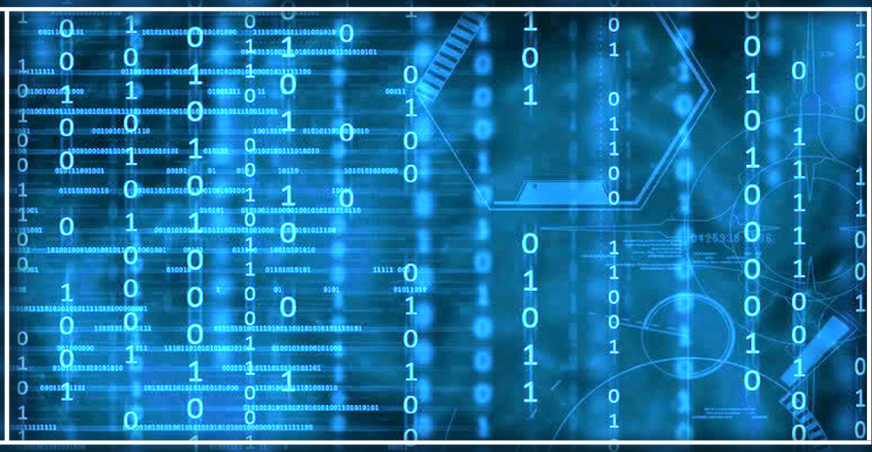
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 12 Issue 12

December 2021



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

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 12 Issue 12 December 2021
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Machine Learning Augmented Breast Tumors Classification using Magnetic Resonance Imaging Histograms

Ahmed M. Sayed
Biomedical Engineering Department, Helwan University, Helwan, Cairo, Egypt
EECS Department, MSOE University, Milwaukee, WI, USA

***Abstract*—At present, breast cancer survival rate significantly varies with the stage at which it was first detected. It is crucial to achieve early detection of malignant tumors to reduce their negative effects. Magnetic resonance imaging (MRI) is currently an important imaging modality in the detection of breast tumors. A need exists to develop computer aided methods to provide early diagnosis of malignancy. In this study, I present machine learning models utilizing new image histogram features using the pixels least significant bit. The models were first trained on an MRI breast dataset that included 227 images captured using the short TI inversion recovery (STIR) sequence and diagnosed as either benign or malignant. Three data classification methods were utilized to differentiate between the tumor's classes. The examined classification methods were the Discriminant Analysis, K-Nearest Neighborhood, and the Random Forest. Algorithms' testing was performed on a completely different dataset that included another 186 MRI STIR images showing breast tumors with verified biopsy diagnostics. A significant tumor classification efficiency was found, as judged by the pathological diagnosis. Classification's accuracy was calculated as 94.1% for the DA, 94.6% for the KNN and 80.6% for the RF algorithm. Receiver operating curves also showed significant classification performances. The proposed tumor classification techniques can be used as non-invasive and fast diagnostic tools for breast tumors, with the capability of significantly reducing false errors associated with common MRI imaging-based diagnosis.**

***Keywords*—Tumor classification; histogram analysis; magnetic resonance imaging; breast cancer; machine learning**

I. Introduction

Breast cancer is the most common cancer type in women worldwide. It is the fifth cause of female deaths due to cancer [1]. Around 300,000 new female cases is estimated to occur each year in the United States alone [2]. The survival rate for breast cancer have generally improved over the past few years, as diagnosis at an early and localized stage is now possible, because of the progressive improvement in treatment strategies [3]. Early diagnosis of malignant tumors is crucial to avoid tumor metastasis and subsequently elevate the survival rate of diseased cases. If the tumor was not diagnosed early, it may spread beyond the original breast organ to other distant organs. Currently the routine method for diagnosing suspected breast tumors is imaging using Mammography, the main imaging modality for the breast organ that is then followed by pathological diagnosis through extraction of a biopsy sample from the tumor invasively. Mammography breast cancer detection sensitivity is generally high [4], however, this sensitivity goes down to near 62% when imaging females with dense breasts [5]. Additionally, the costly biopsy procedure, the gold standard for diagnosis, is routinely performed under ultrasound guidance. However, about 75% of the performed biopsy procedures yield a benign diagnosis [6, 7], which is considered an unnecessary, costly, and time consuming and painful procedure to patients. In order to reduce the wasted biopsy procedures, other imaging modalities were proposed, such as magnetic resonance imaging (MRI) and ultrasound elastography [8-13].

Lately, MRI has become a useful and important modality to visualize and detect breast tumors in today's clinical practice [8, 14, 15]. This imaging modality is becoming increasingly in use to preoperatively evaluate DCIS tumors and define their extent [16, 17]. MRI has the advantages of not producing ionizing radiation, exhibiting high imaging contrast, good sensitivity rate, ability to show auxiliary nodes, and enjoying 3D imaging capabilities [18]. Additionally, Short inversion time Inversion Recovery (STIR) MRI scanning sequence provides a means for suppressing fat and inflammatory tissue from the normal tissue in the resultant images [19, 20]. If MRI is used in daily routine examinations, specific types of cancer would have been significantly diagnosed with higher sensitivity rates at an earlier stage [8], but cost remains a major impediment. Nevertheless, breast imaging using MRI exhibit relatively moderate specificity rates (down to 79%) that increase the erroneous false positive diagnostic percentages [15, 21-23].

One of the MRI imaging characterization methods is histogram analysis that is usually used to distinguish different anatomical and morphological regions, in addition to its more fundamental usage as an image enhancement tool [24, 25]. Some previous studies used histogram methods to illustrate the relation between the tumors physiological changes and their associated histogram parameters to achieve improved utilization of these histogram parameters as substitutive and representative markers describing heterogeneity of the tumor compositions [15, 26, 27]. In the past years, several studies have exploited histogram approaches in various imaging modalities [24, 27-32] with a growing emphasis on different MRI techniques and imaging sequences. Histogram processing methods showed its value for investigating various tumor parameter distributions, for example, in dynamic contrast-enhanced MRI (DCE MRI) it was possible to differentiate

between responder and non-responder groups in brain tumors radiotherapy [33], and in apparent diffusion coefficient (ADC) using diffusion MRI, it was also possible to detect specific types of cervical cancer [34] and endometrial cancer [35]. Despite that, information related to tumor's heterogeneities remain not fully scrutinized [15]. Along with the advances in high resolution MRI and its associated signal processing methods, histogram analysis of cancer tumors scanned using MRI will be used to a greater extent.

In this research, breast tumor's heterogeneity was investigated and described by the least significant byte histogram parameters calculated from STIR MRI imaging sequences for a number of clinically and pathologically verified patients diagnosis. The aim of this study is to differentiate between the two main breast tumors' classes; benign and malignant, with higher accuracy rates. Computer aided diagnosis was achieved using three classification algorithms to categorize the acquired data. Following that, the classification efficiency was calculated and compared with the outcomes of pathological tumor's diagnosis; consequently, the classification errors and receiver operating curves were calculated using two MRI data sets; one dataset for training the classifiers and the other dataset was utilized for testing. Throughout this study, it will be demonstrated that the proposed breast tumor classification technique has the potential as a noninvasive early diagnosis tool. This may lead to earlier and faster tumors characterization, and also may reduce the number of unnecessary biopsies performed pathologically to determine benignancy or malignancy; the applicable criteria that follow.

II. Materials and Methods

In this research, the used training dataset was an online imaging dataset made available for scientific studies. It was published by the Cancer Imaging Archive (TCIA) [36] under the Breast-Diagnosis collection [37]. Table I lists the mass types included in this training dataset and their pathologic diagnosis, as published in the clinical, pathology, and radiologist reports [37]. A different MRI dataset was used for classification testing purposes. This testing dataset was previously acquired from different health care faculties, where tumors diagnosis was also verified with histopathology, as it was used in a previously published study of the research team [38]. The dataset included 186 tumor images, as listed in Table I, and their biopsy results were also available.

TABLE I. Pathologic Diagnosis of the Examined Cases

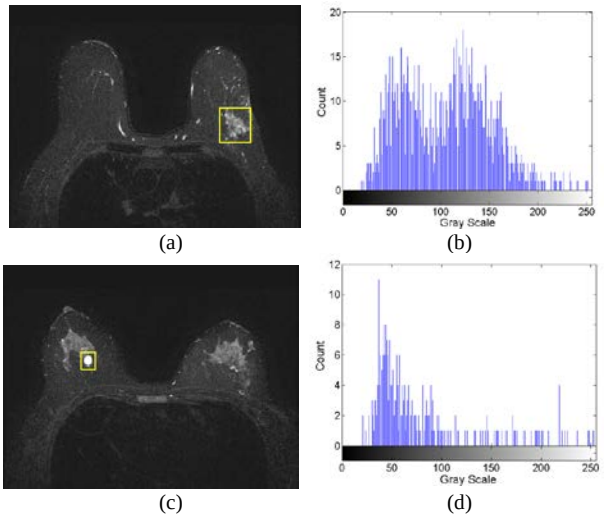
Case Diagnosis	**Pathological Diagnosis (count)**	
	Training Dataset	**Testing Dataset**
Benign	Fibroadenoma (27) Fibrocystic Change (22) Fibrosis (25) Stromal Hyperplasia (8)	Fibroadenoma (43) Fibrocystic Change (15) Cystic Lesion (14)
Malignant	Invasive Ductal Carcinoma (91) Ductal Carcinoma In Situ (19) Invasive lobular Carcinoma (35)	Invasive Ductal Carcinoma (114)
Total Numbers	**Training Dataset**	**Testing Dataset**
Benign	82	72
Malignant	145	114

A. Generating Histograms

The presented analyses in this study focus on obtaining tumor's histograms and identify the important classification features. Apparently, obtaining the whole image's histogram will degrade the classification overall accuracy by including imaging features that represent the surrounding non-tumorous tissue and healthy organs. Selection of the region of interest (ROI); i.e., delineating only the tumor, is a common practice in the routine radiology analysis. This is a particularly necessary step in this study to exclude non-tumorous features from the classification process.

Tumors' locations and pathological diagnosis were already determined in the training dataset's pathological reports. This information was used to manually select ROIs for all independent breast tumors in the dataset. Fig. 1(a) and Fig. 1(c) show selected ROIs for malignant and benign breast tumors, respectively. Following that the histogram for the selected image ROIs was generated for the least significant byte (LSB) only; the reason for that will be explained shortly. The corresponding histograms are presented in Fig. 1(b) and Fig. 1(d), in which the horizontal axes represent gray scale level variations, and the vertical axes represent the number of pixels for a specific gray scale level.

After generating all tumors' ROI imaging histograms, their classifying features were then computed. The classifying features were chosen to be ten histogram parameters. Those parameters were used to describe the shape and profile of a histogram. The ten histogram parameters were generally used in similar studies found in the literature [15, 26, 34, 38, 39] that aimed to differentiate tumors or identify various morphological regions based on imaging data. The used histogram features were: maximum, median, mean, mode, entropy, standard deviation, kurtosis, skewness, 75 percentile and the 25 percentile values. Entropy measures the degree of uniformity of a histogram. Kurtosis represents a measure of the histogram general shape. Skewness represents a histogram's data asymmetry about the mean value. Percentile values represent a value below a specified limit of the calculated histogram data.



(a) (b) (c) (d)

Fig. 1. MRI Imaging Examples of Breast Tumors' ROI and their Least Significant Byte (LSB) Associated Histograms: (a, b) Malignant Tumor; (c, d) Benign Tumor.

Statistically speaking, this study is an observational study, with no control over the classification features. Statistical testing was used to examine the classification significance of the chosen histogram parameters/features. First normality test had to be applied to select a suitable significance test. The Jarque–Bera normality test was utilized to examine the features normality. All features failed the normality test; therefore, the parametric t-test could not be used and the non-parametric Wilcoxon rank sum test was used instead. This test showed that some features had a significant classification power, provided that the alternative testing hypothesis of different features distribution medians for the two classes (benign and malignant) were different. Statistical significance level α of 0.05 was considered throughout this study.

For both training and testing data sets, only Short TI Inversion Recovery (STIR) MRI imaging were examined, which is an imaging procedure that aims at highlighting the breast mass's morphology and facilitates visualization of tumor's heterogeneity. Each pixel of the studied MRI images consists of two bytes; least significant bytes (LSB) and most significant bytes (MSB). Two types of histograms were generated: histograms based on the whole pixel size; LSB and MSB, and histograms based using LSB only. As shown in Table II, the number of significant histogram features using the full pixel size were only 6, compared to 8 significant features when using the LSB. Kurtosis and skewness were the additional significant features in the second case. For the full-length histograms, the mode was the most significant feature with P value of 0.0026, while for the LSB case, skewness was the most significant feature with 8.35E-05 P value.

Evidently, features based on the LSB histograms would provide more classification power between the two tumor types. The LSB's histogram information may have magnified the tumor's image heterogeneity and adherence pattern with the surrounding normal tissue. Fig. 1 shows benign and malignant examples along with their LSB histograms. The malignant LSB histogram has a greater content of low pixel values, in contrast to the benign LSB histogram that has larger content of high pixel values that has been truncated in the calculation of the LSB histograms. Therefore, this interesting and useful effect was encouraging to proceed the classification process using LSB histograms rather than full pixel length histograms, which according to my knowledge, has not been reported before.

B. Data Classifiers

Three classifiers were exploited to automatically categorize the examined images as either malignant or benign, according to the corresponding histogram features. The used classifiers were the discriminant analysis (DA), K-Nearest Neighbor (KNN), and Random Forest (RF) classifiers. The 227 data points were utilized to train and validate the three classifiers, judged by their consequent resubstitution error, and the leave-one-out analysis. The three classifiers were chosen for their popularity, implementation simplicity, and prior use in similar applications [40-42].

The DA classifier tries to find a combination of the classifying features that divides the two disease main classes; benign and malignant. The discriminant analysis as a parametric method, attempts to estimate a categorical or grouping dependent variable based on a number of continuous independent variables; i.e. predictor variables using a preselected discriminant function. The dependent variable in our application was the tumor diagnosis outcome, while the independent variables were the MRI imaging features. Previous studies show that this classification method has shown an acceptable classification performance, even with inappropriate features selections [41, 43].

KNN is a nonparametric classification method. A data point is classified according to the distance between it and its neighbors in the feature space, with the point being assigned to a class that is closest to its K nearest neighbors. The main parameter controlling the performance of such classifier is the number of neighbors, K. A common tradeoff in selecting the right value of the parameter K exists, where larger K values make classification outcomes less vulnerable to the effect of noise or data outliers but results in less distinct classification boundaries. On the other hand, lower values of K produce uneven and irregular classification boundaries [44]. Therefore, the classification analysis was repeated in the training phase for different values of the K parameter, and the resultant classification error was reported accordingly.

The third used classifier in this study was the RF which is also a nonparametric classification method. In this method, a group learning model is constructed using a large number of decision trees. Classification is performed according to the mode of the decision trees. Classification trees have the advantage of making a good fit to the training data [45, 46]. The main parameter in this method is the number of trees T used to build the classification model. Therefore, the classification analysis was repeated in the training phase for different values of the T parameter, and results were compared at each selected value. It is worth mentioning that it has been reported by Lin and Jeon [47], that a relationship exists between RF and KNN methods, where both belong to the weighted neighborhoods schemes.

TABLE II. STATISTICAL TESTING OF HISTOGRAM CLASSIFICATION FEATURES USING TWO IMAGING PIXELS SIZES

P-values based on LSB+MSB										**# Significant features**
Entropy	**Max**	**Median**	**Mean**	**STD**	**Mode**	**Kurtosis**	**skewness**	**prctile75**	**prctile25**	
0.1077	0.0135	0.013	0.0034	0.59	0.0026	0.1077	0.4112	0.0125	0.0063	6
P-values based on LSB										
Entropy	**Max**	**Median**	**Mean**	**STD**	**Mode**	**Kurtosis**	**skewness**	**prctile75**	**prctile25**	
0.847	0.0468	1.34E-04	0.0071	0.327	0.0251	1.47E-04	8.35E-05	1.71E-04	0.0011	8

After training each classifier using the training MRI dataset, the associated resubstitution error was calculated. Additionally, leave one out analysis (LOO) was performed as a validation step. In leave one out analysis each classifier was trained on the whole data set except for one data point, and classification was then predicted for this data point. This process is repeated until all points were diagnosed based on the model generated using the other trained data points. Both resubstitution and LOO analysis data were used to select the classifiers' parameters that generate the lowest false negative error with a high level of accuracy. Achieving low false negative errors is very crucial, as misclassifications of positive malignant tumors are so severe, as they lose the early detection and treatment privileges.

Following training, testing of the classification model follows using another independent dataset that consists of 186 testing images. Classification accuracy and receiver operating characteristic (ROC) curves were plotted for each classification model and analyzed accordingly. A flowchart summarizing the tumor's classification process is shown in Fig. 2.

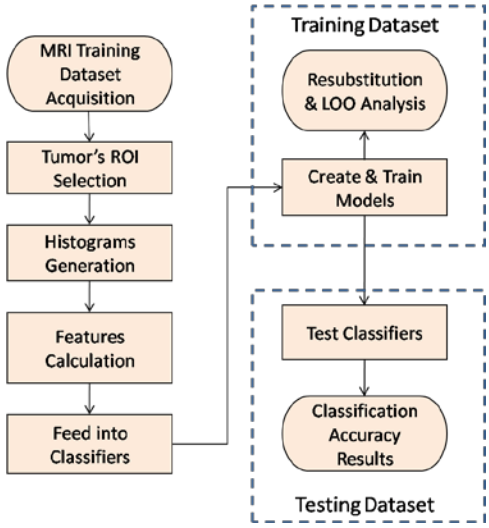


Fig. 2. Flow Chart of the Breast Tumor's Classification Process.

III. Results

Application of the described approach resulted in labeled histogram data points. The points were used for training and evaluation of the three classifiers. The classifiers where then tested, and classification was found to be significant, with different efficiencies according to the used classifier, as will be shown in the following subsections.

A. Classifiers Training Evaluation

1) DA classifier: Five DA discriminant functions were evaluated and the results of resubstitution error and leave one analysis are shown in Table III. The table lists the True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP) values for each discriminant function. Sensitivity, specificity, and accuracy values are used to determine the classifier's performance. The sensitivity value provides a very important indication, as higher sensitivity values point out the classifier's ability to identify malignant tumors as malignant. It is ideal to have a false negative value of zero (sensitivity of 1). The Positive and Negative Likelihood Ratios (PLR and NLR respectively) are also reported in Table III. It is desirable to have a classifier with higher positive likelihood ratio > 2 and NLR < 0.5 for a better discriminatory classification [48].

From the table, it is evident that the Mahalanobis discriminant function provides satisfactory classification results with high sensitivity and low NLR reflecting the low possibility of producing false negative errors. Although the other functions provided a lower FP error, yet the severity of FN errors is much higher than FP errors.

2) KNN classifier: The KNN classifier was trained using the standard Euclidean distance metric. This classifier requires determination of the number of neighboring points; K parameter, to be included while creating the model decision boundary. A good value of K would provide a suitable compromise between the classifier's sensitivity to noise at low values of K form one hand, and the reduced classification accuracy at high values of K on the other hand. Therefore, different K values were examined, as shown in Fig. 3.

In Fig. 3, both the resubstitution and LOO analysis were performed for K values ranging from 1 to 100, to explore any potential useful values of K. Only the classification accuracy is being graphed to show the overall performance of the classifiers, but specificity and sensitivity are also reported and analyzed in Table IV. The resubstitution accuracy profile in Fig. 3(a) shows a rapid accuracy decline as K increases with a small peak that appears at K=15, then the curve declines again until it settles at an accuracy of 79% approximately. The LOO accuracy profile in Fig. 3(b) shows two peaks at K=5 and K=15, then the curve settles at about 79 % accuracy level. From both curves, it is rational to select the value of 15 rather than 5, to make the model more generalized and less sensitive to data outliers and data irregularities. For this value of K = 15, an overall training classification sensitivity of about 85.5% was achieved with a low NLR; which is an indication of a low possibility of producing false negative diagnosis.

3) Random Forest Classifier: In this classifier, the main parameter is the number of trees (T) composing the forest. The resubstitution and LOO analyses were performed for T values from 5 to 300, to explore potentially useful T values, as illustrated in Fig. 4. The resubstitution accuracy profile shown in Fig. 4(a), exhibits a steady accuracy of 100% for T values larger than 40 trees. The LOO profile in Fig. 4(b) shows accuracy fluctuations around the 79% accuracy level for almost all values of T. It can be noticed from the figure that the resubstitution error is infinitesimal and only occurs for small tree numbers. Nevertheless, the LOO analysis reveals the actual performance of the RF algorithm when data points not included in the training are tested, which indicates a model overfitting effect. Based on these results, a classifier with a decision trees number T of 100 was chosen for further testing, and the corresponding training evaluation calculations are listed in the second section of Table IV.

TABLE III. TRAINING AND EVALUATION OF THE DISCRIMINANT ANALYSIS CLASSIFIER USING DIFFERENT DISCRIMINANT FUNCTIONS

A- Resubstitution Analysis										
Discriminant function	**TN**	**TP**	**FN**	**FP**	**Count**	**Sensitivity**	**Specificity**	**Accuracy**	**PLR**	**NLR**
Mahalanobis	52	137	8	30	227	0.945	0.634	0.832	2.583	0.087
Linear	56	134	11	26	227	0.924	0.683	0.837	2.914	0.111
Diagonal Linear	68	106	39	14	227	0.731	0.829	0.767	4.282	0.324
Quadratic	73	107	38	9	227	0.738	0.89	0.793	6.723	0.294
Diagonal Quadratic	72	81	64	10	227	0.559	0.878	0.674	4.58	0.503
B- Leave One Out Analysis										
Discriminant function	**TN**	**TP**	**FN**	**FP**	**Count**	**Sensitivity**	**Specificity**	**Accuracy**	**PLR**	**NLR**
Mahalanobis	52	136	9	30	227	0.938	0.634	0.828	2.564	0.098
Linear	54	133	12	28	227	0.917	0.659	0.824	2.686	0.126
Diagonal Linear	68	104	41	14	227	0.717	0.829	0.758	4.20	0.340
Quadratic	69	107	38	13	227	0.738	0.841	0.775	4.655	0.311
Diagonal Quadratic	72	81	64	10	227	0.559	0.878	0.674	4.580	0.503

TABLE IV. EVALUATION OF THE KNN AND RF CLASSIFIERS' TRAINING PERFORMANCE

	TN	**TP**	**FN**	**FP**	**Count**	**Sensitivity**	**Specificity**	**Accuracy**	**PLR**	**NLR**
KNN Resubstitution Analysis										
K = 15	62	126	19	20	227	0.869	0.756	0.828	3.563	0.173
KNN Leave One Out Analysis										
K = 15	61	122	23	21	227	0.841	0.744	0.806	3.285	0.213
RF Resubstitution Analysis										
T = 100	82	145	0	0	227	1	1	1	Inf	0
RF Leave One Out Analysis										
T = 100	58	127	18	24	227	0.876	0.707	0.815	2.993	0.176

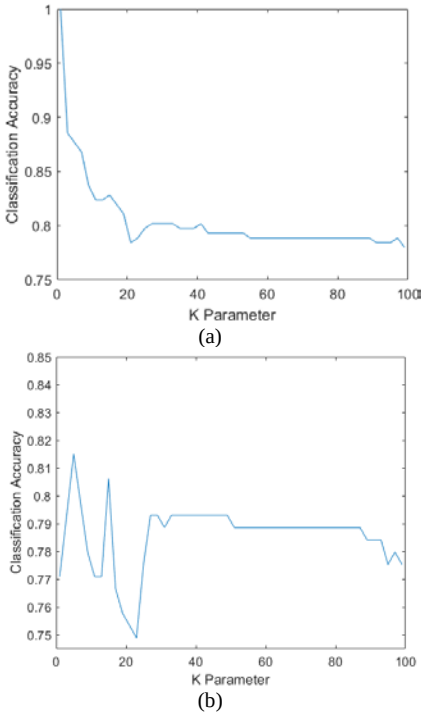


Fig. 3. KNN Classifier: (a) Training Accuracy for different Values of K Parameter using Resubstitution Analysis. (b) Training Accuracy for different Values of K Parameter using LOO Analysis.

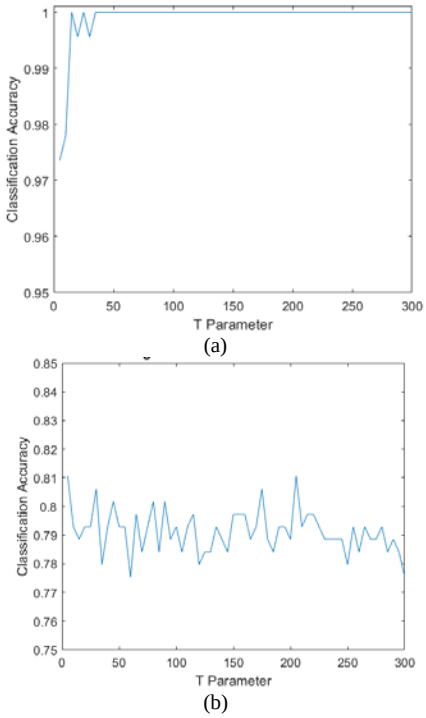


Fig. 4. Random Forest Training Evaluation: (a) Training Accuracy for different Values of T Parameter using Resubstitution Analysis. (b) Training Accuracy for different Values of T Parameter using LOO Analysis.

B. Classifiers Testing Evaluation

One contribution of this study is exploitation of trained classification models to classify an entirely different testing dataset that was not included in the models training. No further tuning or post processing was applied on the models, and testing was performed directly resulting in the following tumor diagnosis outcomes.

1) Testing DA classifie: A summary of DA testing outcomes is presented in Table V. The DA classified the tumors with a sensitivity of 99.0%, specificity of 87.8%, and accuracy of 94.1% with a very low NLR ratio.

2) Testing KNN classifier: The KNN classifier testing results are summarized in Table V for the same K value used in the training process. The testing accuracy of the algorithm was calculated for all values of K; from 1 to 100, as shown in Fig. 5(a). Two accuracy peaks appear at 15 and 23 and giving the same exact classification accuracy.

It is apparent that the KNN has a very close performance to the DA classifier, yet the sensitivity measure was found to be better using the KNN, as the calculated NLR was almost zero with no FN errors.

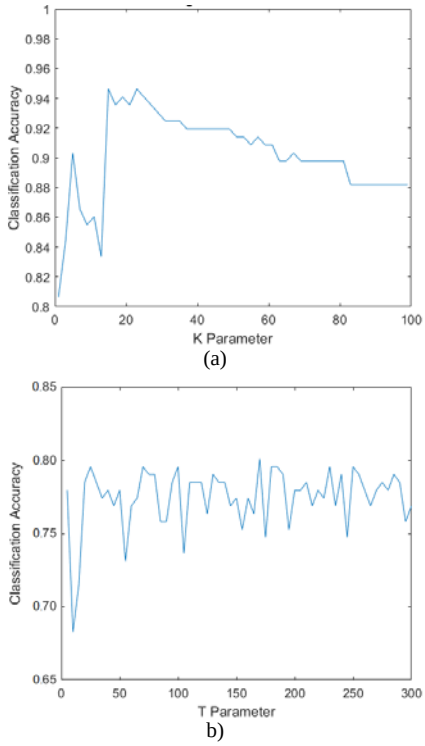


Fig. 5. (a) Accuracy of Classification Testing for all Values of K Parameter for the KNN Classifier. (b) Accuracy of Classification for all Values of T Parameter for the RF Classifier.

3) Testing random forest classifier: The RF algorithm tumor categorization outcomes were summarized in Table V for the same T value used in the training process. Once more, the testing accuracy of the algorithm was calculated for all values of T; from 5 to 300, as shown in Fig. 5(b). The accuracy profile does not show specific peaks or range of T values with higher accuracy levels. A fluctuating, yet steady performance is noticed for T values of more than 25, around the 77.5 % accuracy level. Yet, a sensitivity level of only 76% was noted, as the classifier failed to correctly categorize 25 malignant tumors. The RF testing profile is very similar to the RF LOO training profile. This observation is interesting, as it shows that the RF classifier performance can be accurately predicted based on the LOO training curve profile.

To compare the three classifiers' performance at the selected classification parameters, a combined ROC curve was plotted in Fig. 6. It is clear that the DA and KNN classifiers are superior to the RF classifier, in terms of sensitivity and specificity. The areas under the curves (AUC) were 0.956 and 0.953 for the DA and KNN, respectively, indicating a very good and significant classification performance. The AUC value for the RF model was calculated to be only 0.845 reflecting a moderate classification performance.

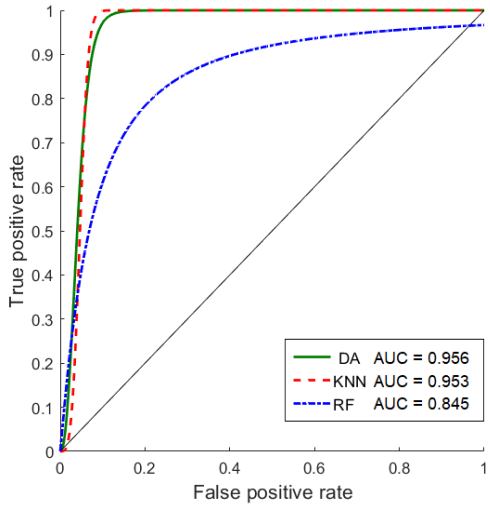


Fig. 6. ROC Curves showing Performance of the Three Breast Tumor Classifiers.

IV. Discussion

In this article, breast tumor classification methods based on MRI LSB histogram parameters were demonstrated. The selected histogram features were used to train three different machine learning methods as tumor classifiers, and their corresponding performances were compared. Previous studies in the literature used histogram parameters to characterize breast cancer and its response to therapy [27, 41, 49-51]. Their main goal was finding the statistical significance of histogram features to categorize the examined breast tumors, yet the usage of machine learning techniques for such a purpose was limited in the literature. The study reported by Vidić and coworkers presented a support victor machine algorithm to evaluate breast tumors classifications [50]. Although the authors showed an overall classification accuracy of up to 0.96, they reported only accuracy values without considering specificity nor sensitivity ratios, therefore, no information were provided about false positive or negative classification errors. Lee and associates reported five machine learning algorithms to predict prognostic biomarkers of breast cancer [51]. The authors reported an AUC value of 0.8 using a random forest

model, which is quite close to the RF AUC value we report herein, although higher AUC values were achieved using other algorithms, as demonstrated in the paper in hand. Additionally, usage of histogram features calculated using only the LSB imaging pixels has never been reported, according to my best knowledge, which have shown more significant differences between benign and malignant tumors. This effect can be explained as follows: important information pertaining to the tumor's heterogeneity, adherence to the surrounding normal tissue, and response to magnetic excitations may have been emphasized in the image's low gray scale values, due to the scanning nature of STIR MRI sequence that suppresses the fatty normal tissues, i.e., it nulls the signal from fat.

Herewith, two different breasts MRI dataset were exploited: one for training the classifiers and the other for independently testing them. One main contribution in this study is the application of a trained machine learning method on a totally different dataset from a different source. This strengthens the hypothesis that the described methods are generalized classification methods that could be efficiently used to classify any given STIR MRI breast tumor images. Furthermore, the described methods can be easily repeated and validated by other research groups on their own datasets.

Statistical analysis of the selected histogram features showed skewness as the most significantly different parameter between benign and malignant tumors, with a P value of 8.35E-05. In general, skewness represents the shape and asymmetry of a given histogram. Based on the training MRI dataset, the average skewness value for the benign images was 5.866, while for the malignant tumors the value was 2.001; approximately 3 folds. This indicates that benign histograms were quite asymmetric around the mean and more right-skewed towards the higher image pixel values as compared to malignant histograms.

As mentioned earlier the selection of the three classifiers was based on their inherent implementation simplicity and prior use in similar applications [40-42]. The aid of machine learning algorithms to improve diagnostic accuracy is of significant interest and utility, as human interpretation of MRI breast data is neither 100% sensitive nor 100% specific. Even though the DA algorithm assumes a multivariate normal distribution between the used features, which is not the case here, yet it was successful in categorizing the testing tumors data. It has been reported that violations of the normality assumption can be permitted in certain cases, and the algorithm outcomes can still be considered reliable, given that the non-normality violations are not caused by data outliers [52]. The other two algorithms; KNN and RF, do not assume normality for the input data points.

The KNN classifier was also very successful and specific in classifying the examined breast tumors. Selection of a certain K parameter was a compromise between good classification performance and robustness against noise and data outliers. Optimization techniques can be used to find the optimum KNN parameters by minimizing the cross-validation loss error. Readily available hyperparameter optimization methods (MATLAB, The MathWorks Inc., Natick, Massachusetts, US) was attempted using the training data and tested as well using the testing dataset. The optimization process recommended using the Spearman distance function (instead of the standard Euclidean method used throughout this study) and a K value of 38. In this case, the classification error occurred only for 1 FP and 1 FN data points out of the 186 testing points (sensitivity of 99% and specificity of 98.8%), which is a remarkable classification performance. Yet, the goal of this paper is to demonstrate the feasibility of using different machine learning techniques to categorize breast tumors and compare between their performances in a pilot study. The task of finding an optimal classifier for that purpose would need more investigation and testing using larger datasets, which is considered future work.

The DA and KNN classification performance metrics showed significantly better outcomes in categorizing testing data over the training data. Training outcomes showed an accuracy of approximately 83%, while testing data showed about 94% accuracy; more than 10% of accuracy increase. This effect can be explained by the fact that the training dataset was larger and more diverse than the testing dataset. As demonstrated in Table I, the training dataset included 227 independent images with 7 different tumor types, while the testing dataset included 186 images showing 4 types of common breast tumors. The training dataset included more tumor types, however, some of them were uncommon and rare tumor types, such as stromal hyperplasia and Ductal Carcinoma in Situ [53, 54]. The trained algorithms used a more generalized data than the testing data, which was the main cause behind the accuracy differences between the two situations. The lack of a more generalized testing dataset is considered another limitation of this study. Despite the encouraging results that have been shown in this study, testing the developed classification methods on a larger and more diverse MRI dataset is an ongoing work.

The RF algorithm had though a moderate classification performance with an utmost accuracy ratio of 80%. It has also the disadvantage of being expensive regarding the computational time. RF LOO analysis was completed in approximately 70 minutes to run using MATALB (The MathWorks Inc., Natick, Massachusetts, US) on a modern computer (Windows 10, 10th Generation Core I5, 2.11 GHz, 16GB RAM), while DA LOO and KNN LOO were completed in 2 and 6 seconds, respectively. Nevertheless, this method's testing performance may be directly predicted from the training LOO data analysis, as the RF algorithm behaved in a very similar and consistent way in both cases, with slight accuracy degradation under the testing mode. This was clear from the demonstrations in Fig. 4(b) and 5(b). Another interesting observation is that the three unoptimized algorithms had very close specificity ratios though; 87.8 for DA, 87.8 for KNN and 85.4 for RF, respectively.

TABLE V. TESTING RESULTS FOR THE BREAST TUMOR'S CLASSIFICATION USING THE THREE ALGORITHMS

Algorithm Parameter	TN	TP	FN	FP	Count	Sensitivity	Specificity	Accuracy	PLR	NLR
DA Algorithm										
Mahalanobis	72	103	1	10	186	0.990	0.878	0.941	8.121	0.011
KNN Algorithm										
K = 15	72	104	0	10	186	1	0.878	0.946	8.2	0
RF Algorithm										
T = 100	70	79	25	12	186	0.76	0.854	0.801	5.191	0.282

V. CONCLUSION AND FUTURE WORK

New breast tumors' classification methods based on MRI imaging were presented. The methods showed the potential to provide more accurate tumor's diagnosis non-invasively and timely efficient. This method may provide an alternative approach to the unnecessary biopsy procedures routinely performed to verify a breast tumor preliminary diagnosis. From the demonstrated results, it has been shown that the discriminant analysis and K nearest neighborhood methods can provide good tumor categorization performance with a significant sensitivity and accuracy levels. The random forest method proved to provide a moderate degree of classification accuracy, however, it showed consistent outcomes in both the training and testing data. The reported least significant byte histogram-based algorithms may be applied on other tumor types, but this requires further investigation to prove being valid. Future research projects include applying the described methods on larger and diverse STIR MRI imaging datasets to find an optimized tumors classification scheme.

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. Eman Zaghloul and her team for providing access to the testing dataset used in this study.

REFERENCES

[1] J. Ferlay et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," Int J Cancer, vol. 136, no. 5, pp. E359-86, Mar 2015.

[2] "American Cancer Society. Breast Cancer Facts & Figures 2015-2016," 2015.

[3] "Diet, Nutrition, Physical Activity and breast cancer," 2017.

[4] N. Houssami, S. J. Lord, and S. Ciatto, "Breast cancer screening: emerging role of new imaging techniques as adjuncts to mammography," Medical Journal of Australia, vol. 190, no. 9, pp. 493-498, May 2009.

[5] P. A. Carney et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," Ann Intern Med, vol. 138, no. 3, pp. 168-175, Feb 2003.

[6] S. P. Poplack, P. A. Carney, J. E. Weiss, L. Titus-Ernstoff, M. E. Goodrich, and A. N. A. Tosteson, "Screening mammography: Costs and use of screening-related services," Radiology, vol. 234, no. 1, pp. 79-85, Jan 2005.

[7] S. Goenezen et al., "Linear and nonlinear elastic modulus imaging: an application to breast cancer diagnosis," IEEE Trans Med Imaging, vol. 31, no. 8, pp. 1628-37, Aug 2012.

[8] C. K. Kuhl et al., "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," J Clin Oncol, vol. 23, no. 33, pp. 8469-76, Nov 20 2005.

[9] A. Sayed, G. Layne, J. Abraham, and O. M. Mukdadi, "Nonlinear Characterization of Breast Cancer using Multi-Compression 3D Ultrasound Elastography In Vivo," Ultrasonics vol. 53, no. 5, pp. 979-991, 2013.

[10] A. M. Sayed, G. Layne, J. Abraham, and O. M. Mukdadi, "3-D visualization and non-linear tissue classification of breast tumors using ultrasound elastography in vivo," Ultrasound Med Biol, vol. 40, no. 7, pp. 1490-502, Jul 2014.

[11] E. Warner et al., "Comparison of breast magnetic resonance imaging, mammography, and ultrasound for surveillance of women at high risk for hereditary breast cancer," J Clin Oncol, vol. 19, no. 15, pp. 3524-31, Aug 2001.

[12] C. D. Lehman et al., "Screening women at high risk for breast cancer with mammography and magnetic resonance imaging," Cancer, vol. 103, no. 9, pp. 1898-905, May 2005.

[13] A. M. Sayed, M. A. Naser, A. A. Wahba, and M. A. A. Eldosoky, "Breast Tumor Diagnosis Using Finite-Element Modeling Based on Clinical in vivo Elastographic Data,", J Ultrasound Med, vol. 39, no. 12, pp. 2351-2363, Dec 2020.

[14] C. H. Lee et al., "Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer," J Am Coll Radiol, vol. 7, no. 1, pp. 18-27, Jan 2010.

[15] N. Just, "Improving tumour heterogeneity MRI assessment with histograms," (in eng), Br J Cancer, vol. 111, no. 12, pp. 2205-13, Dec 2014.

[16] C. J. Allegra et al., "National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ September 22-24, 2009," J Natl Cancer I, vol. 102, no. 3, pp. 161-169, Feb 2010.

[17] M. Pilewskie et al., "Effect of MRI on the Management of Ductal Carcinoma In Situ (DCIS) of the Breast," Ann Surg Oncol, vol. 19, pp. S11-S11, Feb 2012.

[18] J. E. Joy, E. E. Penhoet, D. B. Petitti, National Cancer Policy Board (U.S.). Committee on New Approaches to Early Detection and Diagnosis of Breast Cancer., National Research Council (U.S.). Policy and Global Affairs., and National Research Council (U.S.). Board on Science Technology and Economic Policy., Saving women's lives : strategies for improving breast cancer detection and diagnosis. Washington, D.C.: National Academies Press, 2005.

[19] R. Golfieri, H. Baddeley, J. S. Pringle, and R. Souhami, "The role of the STIR sequence in magnetic resonance imaging examination of bone tumours," Br J Radiol, vol. 63, no. 748, pp. 251-6, Apr 1990.

[20] E. M. Delfaut, J. Beltran, G. Johnson, J. Rousseau, X. Marchandise, and A. Cotten, "Fat suppression in MR imaging: techniques and pitfalls," Radiographics, vol. 19, no. 2, pp. 373-82, Mar-Apr 1999.

[21] E. Othman et al., "Comparison of false positive rates for screening breast magnetic resonance imaging (MRI) in high risk women performed on stacked versus alternating schedules," Springerplus, vol. 4, p. 77, 2015.

[22] C. J. Allegra et al., "National Institutes of Health State-of-the-Science Conference statement: Diagnosis and Management of Ductal Carcinoma In Situ September 22-24, 2009," J Natl Cancer Inst, vol. 102, no. 3, pp. 161-9, Feb 2010.

[23] W. Chen, M. L. Giger, U. Bick, and G. M. Newstead, "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," Med Phys, vol. 33, no. 8, pp. 2878-87, Aug 2006.

[24] J. S. H.D. Cheng, Wen Ju, Yanhui Guo, Ling Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," Pattern Recognition, vol. 43, pp. 299-317, 2010.

[25] H. Satake, S. Ishigaki, R. Ito, and S. Naganawa, "Radiomics in breast MRI: current progress toward clinical application in the era of artificial intelligence, Radiol Med, 2021.

[26] J. S. Carter et al., "Quantitative multiparametric MRI of ovarian cancer," (in eng), J Magn Reson Imaging, vol. 38, no. 6, pp. 1501-9, Dec 2013.

[27] Y. C. Chang, C. S. Huang, Y. J. Liu, J. H. Chen, Y. S. Lu, and W. Y. Tseng, "Angiogenic response of locally advanced breast cancer to neoadjuvant chemotherapy evaluated with parametric histogram from dynamic contrast-enhanced MRI," Phys Med Biol, vol. 49, no. 16, pp. 3593-602, Aug 21 2004.

[28] I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digitized mammograms," Comput Med Imaging Graph, vol. 26, no. 5, pp. 309-19, Sep-Oct 2002.

[29] K. Holli et al., "Characterization of breast cancer types by texture analysis of magnetic resonance images," Acad Radiol, vol. 17, no. 2, pp. 135-41, Feb 2010.

[30] D. R. Chen, R. F. Chang, W. J. Kuo, M. C. Chen, and Y. L. Huang, "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks," Ultrasound Med Biol, vol. 28, no. 10, pp. 1301-10, Oct 2002.

[31] D. Checkley, J. J. Tessier, J. Kendrew, J. C. Waterton, and S. R. Wedge, "Use of dynamic contrast-enhanced MRI to evaluate acute treatment with ZD6474, a VEGF signalling inhibitor, in PC-3 prostate tumours," Br J Cancer, vol. 89, no. 10, pp. 1889-95, Nov 17 2003.

[32] G. Y. Cho et al., "Evaluation of breast cancer using intravoxel incoherent motion (IVIM) histogram analysis: comparison with malignant status, histological subtype, and molecular prognostic factors," Eur Radiol, vol. 26, no. 8, pp. 2547-58, Aug 2016.

[33] S. L. Peng et al., "Analysis of parametric histogram from dynamic contrast-enhanced MRI: application in evaluating brain tumor response to radiotherapy," NMR Biomed, vol. 26, no. 4, pp. 443-50, Apr 2013.

[34] K. Downey et al., "Relationship between imaging biomarkers of stage I cervical cancer and poor-prognosis histologic features: quantitative histogram analysis of diffusion-weighted MR images," AJR Am J Roentgenol, vol. 200, no. 2, pp. 314-20, Feb 2013.

[35] S. Ytre-Hauge et al., "Preoperative tumor texture analysis on MRI predicts high-risk disease and reduced survival in endometrial cancer," J Magn Reson Imaging, vol. 48, no. 6, pp. 1637-1647, 2018.

[36] V. B. Clark K, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," Journal of Digital Imaging, vol. 26, no. 6, pp. 1045-1057, 2013.

[37] B. N. Bloch, A. Jain, and C. C. Jaffe. Data From BREAST-DIAGNOSIS, The Cancer Imaging Archive, 2015a.

[38] Ahmed M Sayed, Eman Zaghloul, and T. M. Nassef, "Automatic Classification of Breast Tumors Using Features Extracted from Magnetic Resonance Images," Procedia Computer Science, vol. 95, pp. 392 – 398, 2016.

[39] H. Chandarana et al., "Histogram analysis of whole-lesion enhancement in differentiating clear cell from papillary subtype of renal cell cancer," Radiology, vol. 265, no. 3, pp. 790-8, Dec 2012.

[40] E. I. Zacharaki et al., "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," Magn Reson Med, vol. 62, no. 6, pp. 1609-18, Dec 2009.

[41] D. J. Tozer, G. R. Davies, D. R. Altmann, D. H. Miller, and P. S. Tofts, "Principal component and linear discriminant analysis of T1 histograms of white and grey matter in multiple sclerosis," Magn Reson Imaging, vol. 24, no. 6, pp. 793-800, Jul 2006.

[42] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput Struct Biotechnol J, vol. 13, pp. 8-17, 2015.

[43] R. Wolz et al., "Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease," PLoS One, vol. 6, no. 10, p. e25446, 2011.

[44] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," IEEE Trans Neural Netw, vol. 12, no. 2, pp. 181-201, 2001.

[45] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings," BMC Genet, vol. 11, p. 49, Jun 2010.

[46] L. Breiman, "Random forests," Mach Learn, vol. 45, pp. 5-32, 2001.

[47] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," Journal of the American Statistical Association, vol. 101, no. 474, 2006.

[48] S. McGee, "Simplifying likelihood ratios," J Gen Intern Med, vol. 17, no. 8, pp. 646-9, Aug 2002.

[49] R. Johansen et al., "Predicting survival and early clinical response to primary chemotherapy for patients with locally advanced breast cancer using DCE-MRI," J Magn Reson Imaging, vol. 29, no. 6, pp. 1300-7, Jun 2009.

[50] I. Vidić et al., "Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study," J Magn Reson Imaging, vol. 47, no. 5, pp. 1205-1216, 2018.

[51] J. Y. Lee et al., "Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI," Eur Radiol, vol. 32, no. 1, pp. 650-660, 2022.

[52] B. G. Tabachnick and L. S. Fidell, Using multivariate statistics, 6th ed. Boston: Pearson Education, 2013.

[53] S. D. Raj et al., "Pseudoangiomatous Stromal Hyperplasia of the Breast: Multimodality Review With Pathologic Correlation," Curr Probl Diagn Radiol, vol. 46, no. 2, pp. 130-135, 2017 Mar - Apr 2017.

[54] S. S. Badve and Y. Gökmen-Polar, "Ductal carcinoma in situ of breast: update 2019," Pathology, vol. 51, no. 6, pp. 563-569, Oct 2019.

New Feature Engineering Framework for Deep Learning in Financial Fraud Detection

Chie Ikeda, Karim Ouazzane, Qicheng Yu, Svetla Hubenova
School of Computing and Digital Media
London Metropolitan University
London, UK

***Abstract*—The total losses through online banking in the United Kingdom have increased because fraudulent techniques have progressed and used advanced technology. Using the history transaction data is the limit for discovering various patterns of fraudsters. Autoencoder has a high possibility to discover fraudulent action without considering the unbalanced fraud class data. Although the autoencoder model uses only the majority class data, in our hypothesis, if the original data itself has various feature vectors related to transactions before inputting the data in autoencoder then the performance of the detection model is improved. A new feature engineering framework is built that can create and select effective features for deep learning in remote banking fraud detection. Based on our proposed framework [19], new features have been created using feature engineering methods that select effective features based on their importance. In the experiment, a real-life transaction dataset has been used which was provided by a private bank in Europe and built autoencoder models with three different types of datasets: With original data, with created features and with selected effective features. We also adjusted the threshold values (1 and 4) in the autoencoder and evaluated them with the different types of datasets. The result demonstrates that using the new framework the deep learning models with the selected features are significantly improved than the ones with original data.**

***Keywords*—Financial fraud; online banking; feature engineering; unbalanced class data; deep learning; autoencoder**

I. Introduction

As the online payment system advances, fraud schemes have shifted from physical fraud actions using ATMs into an advanced technique that uses digital banking accounts. Unauthorized remote banking fraud is formed by three categories: Internet banking, telephone banking and mobile banking. A fraudster accesses a customer's bank account through these remote banking channels and steals money by making an unauthorized money transfer from the account. UK finance announced that total losses through remote banking in the United Kingdom have increased and reached £197.3 million in 2020, 31percent higher than in 2019. The annual number of cases of internet banking fraud and mobile banking fraud has been growing rapidly from 32,721 cases in 2019 to 66,150 cases in 2020. Other financial fraud losses such as payment cards and cheques decreased from £470.2 million to £452.6 million [1].

The fraud Detection System (FDS) used by many financial institutions, has not caught up with the advancement in fraudulent schemes on remote banking. To address constant changes in fraud behavior, some financial industries employ machine learning (ML) methods in FDS [2, 3], but it is still challenging to reveal new fraudulent behaviors by applying ML to raw data only.

Financial transaction data is also very unbalanced because legitimate transactions account for 90% and above of all transaction data and only less than 10% of the rest of the data is fraud. It is difficult to find fraudulent patterns out for ML algorithms specifically for supervised learning.

In the new feature engineering framework published in [19], we created and selected the effect features for fraud detection models built with ML algorithms: We selected Support Vector Machine (SVM) and Isolation Forest (IF) as fraud detection models.

Throughout this research, we apply the feature engineering framework on remote banking data for deep learning with the experimental dataset being provided by a European private bank.

Deep learning has been popularly used for image, audio and video recognition in terms of coping with big data in depth. It learns by dividing input data into a plurality of segmented data patterns through many hidden layers. Recently, it came to be used for classification issues such as fraud detection in the financial area. The original concept of deep learning will be traced to studies of artificial neural networks (ANN). Autoencoder is a type of ANN, an unsupervised deep learning algorithm [4]. It learns how to compress and decompress input data for representation of the original input data and consists of three layers: Encoder, latent (hidden) layer, decoder (Fig. 1). It discovers specific features from the given data during the process of data compression, also known as dimensionality reduction, and how to map the compressed features to the latent layer. The autoencoder finds out how to reconstruct the input data from mapping the features. The most advantage of using autoencoder for financial fraud detection is that autoencoder does not need fraudulent transaction data to learn fraud patterns. As mentioned above, the proportion of fraud transaction data is very little whereas the number of legitimate transaction data is very large. It is difficult to keep track of new fraudulent behavior and state-of-the-art fraud schemes from a few fraud samples because fraudulent actions are not carried out by one person. On the other hand, legitimate transactions are carried out by the same customer who holds his or her own bank account or credit card. Autoencoder can reconstruct customers' behavior patterns by learning from

specific features among large history transaction data. Autoencoder models judge fraudulent data by using loss function with mean squared error (MSE) that measures the error distance of variables in specific features between the learnt data and new input data. There are some related studies of fraud detection using the autoencoder model [5, 6, 7, 8] and they chose autoencoder techniques from the perspective of coping with unbalanced transaction datasets. They commonly use two popular techniques of feature engineering, which are principal component analysis (PCA) and standardization. PCA is a technique of dimensionality reduction and uses orthogonal transformation that computes covariance matrix which represents the correlation between two variables. Unlike machine learning models, deep learning is essential for data processing standardization as it standardizes and weight each attribute to measure how much specific features influence.

Standardization is an essential data processing for using deep learning because deep learning multiplies each attribute and sets the weighting coefficients. Deep learning has not implemented feature engineering on input data from the point of view of adding latent data patterns.

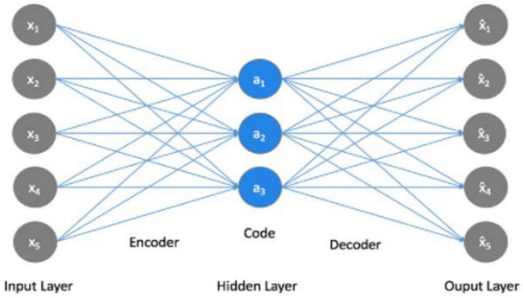


Fig. 1. Autoencoder with Hidden Layer.

In this paper, we propose a new feature engineering framework that newly creates features using feature engineering methods of feature aggregation and feature transformation and selects effective feature candidates for deep learning. In the experiment, the autoencoder is used for building a fraud detection model and verifying the effect of the paper. The rest of the paper is structured as follows: Section 2 reviews the recent development of feature engineering in financial fraud detection for deep learning. Section3 develops a new feature engineering framework that combines feature creation and feature selection processes, and section 4 introduces an autoencoder for fraud detection. Section 5 presents the experimental remote banking dataset, and the final section demonstrates a simulation of the framework for the deep learning model. Finally, section 6 presents results, discussions, and future work.

The main contribution of this study is around improving the accuracy of fraud detection models through using the engineered features produced with the framework, which are the combination of creation and selection processes.

II. Related Work

Until several years ago, the credit card was the great majority of transaction methods in financial services at ATMs, shops, and online shopping. In recent years, remote banking has also become a popular method for transferring money and at the same time, financial fraud losses through remote banking exceed fraud losses of using credit cards according to a report by UK finance in 2021 [1]. Despite the increase of the fraudulent cases of remote banking, there are still very few studies on using feature engineering for deep learning in remote banking.

A. Feature Engineering Framework for Financial Fraud Detection

There exist some similar works of feature engineering framework for financial fraud detection in [9,10,11,12] and they all use feature aggregation methods to create behavior attributes that reveal latent fraudulent patterns. J.S. Kalwihura et al. [11] and Zhang et al. [10] use the HOBA feature engineering methodology which groups into homogenous fraudulent patterns by using feature aggregations based on recency, frequency and monetary (RFM) for insurance fraud detection. The RFM is for behavior analysis which is popularly used in the marketing area. Feature aggregation methods in HOBA feature engineering consist of four aggregation categories related to behavior analysis based on a defined period during a transaction. HOBA also comprises a feature selection method which is a bootstrapped ensemble of bagged trees to select a subset of features from original data. They select random forest as an experimental model which demonstrates a 56.2% increase in the F1-score compared against the original data. Y. Lucas et al. [9] suggest using a feature engineering framework based on multi-perspective Hidden Markov Models (HMMs) for credit card fraud detection. The history of credit card transactions has the card holder's habits of the timing or the place of using a credit card in the last 24h. HMM is a sequence classification model which considers the sequential properties of transaction data. The multi-perspective HMMs categorize a symbol on transactions such as "merchant and amount", "timing", "fraud or customer"," genuine" and observe each symbol as the sequential event on transactions. The HMMs calculate the likelihood of sequences of observed symbols and create features of each event. To measure the effectiveness of the addition of the HMM features, they use perspective, recall and AUC metrics, and random forest as an experimental model. Consequently, the use of the HMM-based features improved the precision-recall AUC of the random forest model significantly compared with the use of the original features only. All the above studies have demonstrated the impact of using feature engineering methods on data with improved performance of machine learning models. In their works, they focused only on the feature aggregations side to reveal latent fraudulent patterns. A. Nagaraja et al. [14] introduced an approach for any network anomaly detection using feature transformation based on mathematical methods. They use feature clustering based on the Gaussian distribution function and a k-Nearest Neighbours (KNN) classifier as a detection model for finding the similarity between observations. The distribution function provides the equivalent deviation and threshold values to carry similarity calculation, and then the distance function of KNN measures the distance of the transformation features and determines if the input is fraud or a legitimate value. Using transformation features improves the detection accuracy in comparison with using the raw data only.

In the new framework [19], we use feature aggregation and feature transformation jointly to create important feature candidates. R. Wedge et al. [15] suggest Deep Feature Synthesis (DFS) that creates new attributes for machine learning models of credit card fraud detection using the relational structure of the dataset. In the processes of DFS, both feature aggregation and transformation methods are used to create new features using attributes of the related transactions. For instance, they applied the Hour in transaction time to determine when a transaction has occurred during the day and use statistical methods i.e., average, mean, sum and standard deviation to express the user behavior on the transaction time base. Timestamps in transactions are significant processes in DFS to compute features of every month and within 24 hours. Eventually, they generated 237 features (over 100 behavioral pattern features) for each transaction and reduced the false positive rate by 54%. However, in their study, they use all 237 generated features which may cause overfitting if the number of the training data is not enough.

In the work described in [19], we used the engineered features created through using the feature engineering framework which has improved the performance of machine learning models.

In this paper, we make use of the new feature engineering framework for deep learning, specifically for autoencoder neural network models.

B. Fraud Detection using Autoencoder Neural Network

Fraud transaction data is always imbalanced and needs to be carefully handled while using machine learning algorithms. Popular methods of coping with imbalanced datasets are oversampling and undersampling which are techniques to balance the class distribution. Oversampling is utilized to synthesize new samples of fraudulent classes but, it will take in noise. Undersampling removes samples from the majority class in the trained dataset but, it may remove useful information or important data. Autoencoder is good for coping with imbalanced datasets without considering the minority class issue because it only uses majority class samples. Some research for credit card fraud detection uses an autoencoder model [16, 17, 18]. P. Jiang et al. [16] designed a six-layer autoencoder for the dataset and selected SoftMax with cross-entropy as the loss function for final classification to detect credit card fraud. The autoencoder model improved the classification accuracy of the fraud class when the threshold was equal to 0.6. A. Pumsirirat et al. [17] used deep learning based on auto-encoder and restricted Boltzmann machine for credit card fraud detection because fraudsters gain new technology that enables them to steal money from customers. Their autoencoder applied backpropagation by setting the input data equal to the output data. Restricted Boltzmann machine can reconstruct legitimate transactions to discover fraudsters from legitimate patterns and holds two layers, input layer and hidden layer. They used the library of TensorFlow to implement autoencoder and restricted Boltzmann machine. The number of studies of financial fraud detection using autoencoder is not a few, but almost all studies use only raw data as the input data for autoencoder. They do not apply feature engineering methods to the raw data.

III. Feature Engineering Framework for Deep Learning Model

The main contribution of the new framework lies in joining two processes of feature creation and feature selection (Fig. 2). Whether machine learning or deep learning algorithms are selected, the processes of feature creation and feature selection in the framework remain the same. In the case of feature creation for deep learning, it is necessary to standardize variables in all features before building a model.

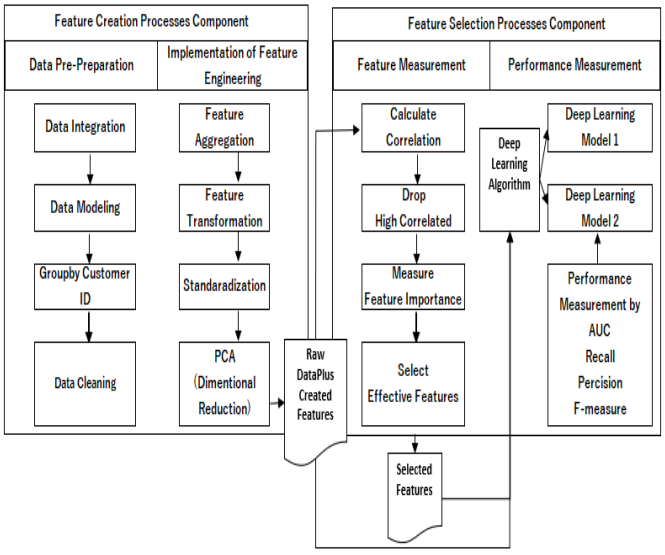


Fig. 2. Feature Engineering Framework.

A. Feature Creation Processes

In the feature creation component, there are two categories: data preparation and implementation of feature engineering. Data preparation contains four processes before the data are ready for implementation using feature engineering methods. The raw data collected from various sources are not clean and need to be maintained by handling missing data and unifying data formats. After the data preparation is made, feature aggregation and feature transformation are sequentially carried out. At this point, a first feature set candidate including all features which include newly created features and original attributes are prepared.

1) Feature aggregation based on customer behavior: Feature aggregation represents a customer's behavior when an online transaction occurs. Based on a unique customer ID, some action attributes e.g., amount, time, access device and network information are aggregated. Aggregation increases the dimensions that can express the data pattern in more detail. The created features by these aggregations represent latent customers' behavior from various angles of data. Table I describes some attribute candidates which can be aggregated with other action attributes for creating individual customer's journeys via online banking.

TABLE I. FEATURE AGGREGATION

Attributes	Combinations
Time	- Days since the last transactions - Hours since the last transactions - Minutes since the last transactions - Days since the last access by same device - Hours since the last access by same device - Minute since the last access by same IP address - Hours since the last access by same IP address - Days since the last event type occurred - Hours since the last event type occurred - Days since the last transaction occurred from specific location/ATM - Hours since the last transactions occurred from specific location/ATM
IP Address	- IP address of access device since last transaction
Amount	- Amount of the last transaction - Amount of the last transaction from specific location/ATM - Amount of the transaction via IP address
Channel	- Channel type when each event is occurred
Event Type	- Event type accessed via IP address - Event type accessed by a specific device

2) Feature transformation based on mathematical functions: There are some available mathematical functions and equations to transform a single attribute into other dimensions by mapping data. The purpose of using transformations is to generate features that discover implications in a given data from mathematical functions i.e., scaling (standardization), log transformation, binning, linear combination, count, on numerical attributes. Some of the functions which are used in the framework are described below:

a) Confidence Interval Formulas

Confidence interval (CI) is a statistic estimation formula that uses the normal distribution for observing a point estimate by calculating maximum, minimum, median, and mean.

b) Standard Deviation

Standard deviation is a method of scaling the values based on z-score which calculates the following equation:

$$Standard\ Deviation = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N-1}} \quad (1)$$

Where:

X_i = Value of each data point

$\bar{X}$ = Mean

N = Number

c) Logarithm Transformation Formula

Log transformation is one of the popular transformation methods used to cope with skewed data because it can remove skewness adapting the formula below.

$$x_i' = \log(x_i) \quad (2)$$

$$x_i' = \log(x_i + 1) \quad \text{in case value can be zero} \quad (3)$$

$$x_i' = \text{sgn}(x_i) log|x_i| = \frac{x^i}{|x_i|} log|x_i|$$

in case value can be negative (4)

$$x_i' = \log(x_i + \sqrt{x_i^2 + \lambda}) \quad (5)$$

generalized log transformation

d) (Linear) Regression Function

This function adapts the concept of linear or multiple regression which classifies the data by fitting two or more attributes to determine the best line. Applying regression helps to discover a mathematical equation for adjusting the data and smoothing out the noise (Fig. 3).

The equation: Let A1, ……, An be n matrices having dimension K x L.

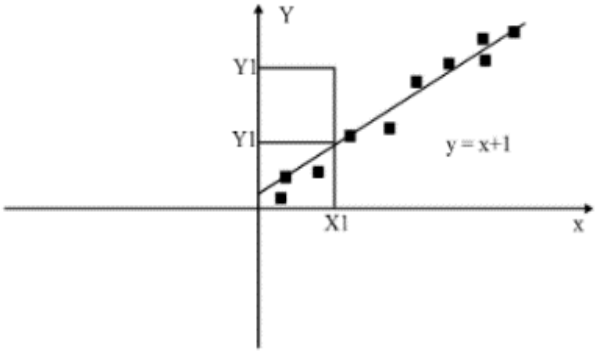


Fig. 3. Linear Regression Function.

$$B = \alpha_1 A_1 + \ldots + \alpha_n A_n \quad (6)$$

e) Clustering (K-Means)

The clustering is to group a set of spots into clusters based on a measured distance. Fraud will be recognised by locating spots out from similar clustering (see Fig. 4). All customers are classified into groups based on similar data patterns by using the K-means clustering method. K-means is an unsupervised learning algorithm that discovers the k number of clusters in a dataset. The K number of clusters are grouped by similarities based on a point at the centre of a cluster. All data are assigned to the closest cluster.

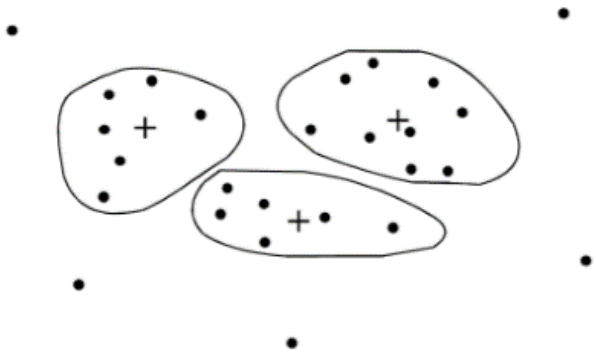


Fig. 4. Clustering.

f) Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised method to reduce feature dimensions from the original feature dimensions but keeps the meaningful variation in the original attributions. PCA explores correlations among the given data and produces new aggregate variables which is a condensed dimensional feature, called principal components (PC).

In Fig. 5, the left side (plot A) shows the original data on the x-axis and y-axis. On the right-side (plot B), the 1st principal axis in the PC1 pivot displays the largest norms. PC2 pivot shows the 2nd principal axis and is orthogonal toward the pivot of PC1. The data in 2-dimension may be diminished to one dimension with extruding each element on the PC1.

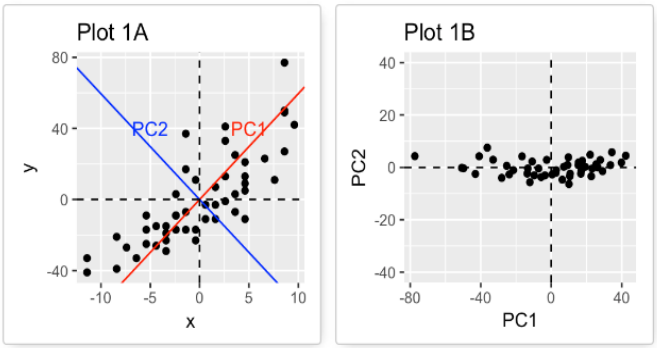


Fig. 5. Principal Component Analysis.

The mathematical approach of PCA is to maximize variances by converting a sequence of values as it is expressed in the following formula. Samples, $X_1, X_2, \ldots, X_N \in R_n$ of the variable X $\in R_n$ that was randomly selected.

$$V_{xx} = \frac{1}{N}\sum_{i=1}^{N}(X^i - \frac{1}{N}\sum_{j=1}^{N}X^j)(X^i - \frac{1}{N}\sum_{j=1}^{N}X^j)^T \quad (7)$$

$$\max_{|a|=1}\frac{1}{N}\sum_{i=1}^{N}(a^T(X^i - \frac{1}{N}\sum_j X^j))^2 = \max_{|a|=1} a^T V_{xx^a} \quad (8)$$

Where a is eigenvector corresponding to the maximum eigenvalue of a variance-covariance matrix of VX_iX_j.

B. Feature Selection Processes

In the feature creation component work introduced above, we created many additional new features. However, redundant features that correlate strongly with other features might be also included. Increasing high dimensional feature space impacts the model performance and causes overfitting, according to Mwadulo [13]. In the feature selection component, there are two main parts for selecting appropriate features from all features having both newly created features and the original data. The first part is feature measurement. In the feature measurement part, we calculate the correlation coefficient and measure feature importance, and then drop redundant features.

- Pearson Correlation Coefficient

When there are high correlations between two or more explanatory variables in the dataset, multicollinearity exists and will cause overfitting in a multiple regression model. The correlation coefficient is a statistical method to measure the degree of intensity of the relationship between feature variables. In the framework, Pearson correlation is selected to calculate the strength between two variables from different types of correlation coefficients. The range of the strength values of the correlation is expressed between -1 and 1. A value of -1 indicates the perfect negative relationship between the two feature values. On the contrary, a value of 1 indicates the perfect positive relationship between the two feature values. Values close to zero means weak or no relationship between the two values (Table II). The equation of the Pearson correlation coefficient is shown below:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \quad (9)$$

Where:

ρ_{xy} = Peason product-moment correlation coefficient

Cov (x, y) = covariance of variables x and y

$\alpha_x = standard\ dviation\ of\ x$

$\alpha_y = standard\ deviation\ of\ y$

TABLE II. BENCHMARK OF CORRELATION COEFFICIENT

Range of Correlation	Interpretation
±0.9 to ± 1.0	Very high positive (negative) corrlation
±0.7 to ± 0.9	High positive (negative) correlation
±0.5 to ± 0.7	Moderate positive (negative) correlation
±0.3 to ± 0.5	Low positive (negative) correlation
0.0 to ± 0.3	No correlation

- Feature Importance Measurement

As an evaluation method of relevant features, we select feature importance to measure the relative importance of each input feature. Scores are calculated by finding a rate of contribution indicating which features influence predictions. In a decision tree model, every node indicates a status of how to split values in an individual feature. The status depends on Gini impurity or information gain in the case of classification. While building a decision tree model, feature importance computes how much a single attribute contributes to reducing the weighted impurity.

IV. DEEP LEARNING ALGORITHM FOR FRAUD DETECTION

- Autoencoder

Autoencoders are unsupervised learning neural networks that learn to encode input data to specific features by reducing dimensions and discovering how the features can be reconstructed and decoded to the original data. In order to measure how well the input data can be reconstructed, a loss function is calculated for updating different weights and reducing the loss between the represented data and the original data. Autoencoder uses unlabeled training data {x (1), x (2), x (3), …}, where x (i) $\in$ Rn and applies backpropagation to learn how to approximate to a function h w, b (x) $\approx$ x displayed in Fig. 6. The output x^ is similar to x.

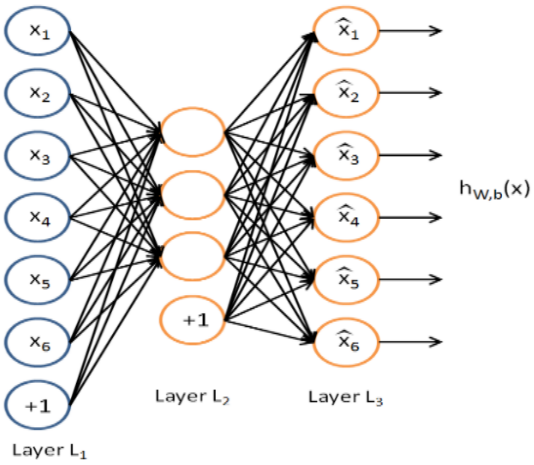


Fig. 6. Autoencoder.

There are three main layers of autoencoder: encoder, hidden and decode.

a) Encoder Layer

An autoencoder model learns how to reduce dimensions of input features and compress the given data into an encoded representation.

b) Hidden Layer

This layer holds the compressed representation of the given data and expresses the most compacted dimensional features.

c) Decoder Layer

The model learns how to reconstruct the compressed data to the original data by using the loss function and calculates the loss between the original data and the reconstructed data. The Mean square error is utilized to measure the error value shown below:

$$l(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} (xi - x^i)\hat{}2 \quad (10)$$

The equation of encoder and decoder are given as follows:

Encoder $h(x) = g((\alpha x))$

$$= \sum (Wx) or \ \tan h(Wx) \quad (11)$$

Decoder $\hat{x} = O(\alpha\hat{}(x))$

$$= \sum (W * h(x)) or \ \tan h(W * h(x)) \quad (12)$$

V. Online Banking Transaction Dataset

The online banking dataset is provided by a European bank for only academic purposes. The dataset contains about 130,000 transactions that occurred via online banking with each customer party ID. The dataset includes fraudulent actions which account for 5% of all transaction records and it is unbalanced labelled data (see Fig. 7).

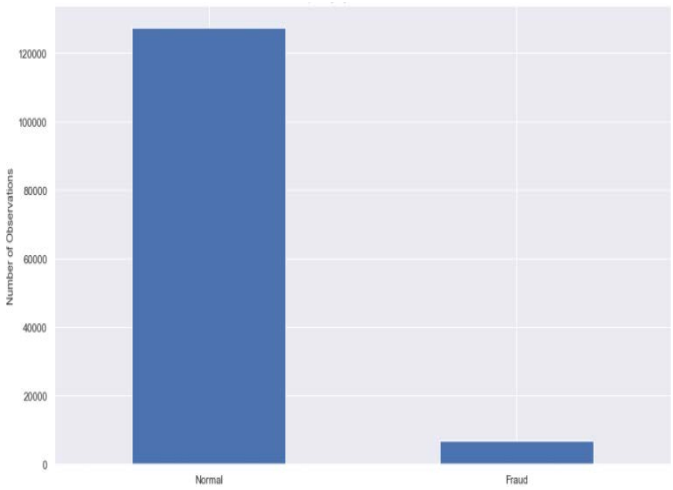


Fig. 7. An Unbalanced Target Data.

There are 39 attributes collected from various data sources such as customer information, bank account, online banking information, device information, network information, timestamp, as described in Table III. Before applying these attributes to the framework, fraudulent tendency from the point of view of transaction amount and timestamp is checked.

In the Fig. 8 shows two distributions of transaction amount frequency. One distribution (in green color) is fraudulent transactions whereas another one (in red) describes normal transactions. Both distributions show little difference between fraud and customer in this context only.

We used a logarithm function on this amount attribute, which is one of the popular mathematical functions. The histogram of log transformation is shown in Fig. 9. The distribution in green shows normal transactions while the one in red represents fraudulent transactions. From the diagnose in Fig. 8, it is shown that the fraudster in red does not steal large money at one time and seems not to be different from normal customers' transactions. It indicates that it is difficult to detect fraud transactions by the rule-based fraud detection system.

The timestamp is considered as an important feature to discover different behavior between a customer and a fraudster as customers will have their usual lifestyle patterns on a time-series basis. Days, Hours and Minutes plot transactions are shown in Fig. 10, 11, 12.

The timestamp in this dataset does not have a remarkable difference between fraud and non-fraud at a glance. Through our framework, this timestamp is segmentalized based on each customer by aggregating customer's information such as amount, network information and access information and creating new features which reveal latent customer behavior or fraudulent pattern.

TABLE III. DESCRIPTION OF ATTRIBUTES IN ORIGINAL DATA

Attribute Name	Description
ED_EVENTTYPETX	Type of event e.g., Customer Login, Make Payment etc
ED_TXNID	Transaction ID
ED_CHANNELIDENTIFIER	A way that customers can interact with a bank. This can be via the telephone, internet banking, branch, mobile.
ED_FINANCIALINSTITUTENM	Financial Institute name
ED_SUBCHANNELNM	Sub-channel name
CUSTD_PARTYID	Customer Party ID
CUSTD_EMAILADDRESSTX	Customer's email address
EVENT	Event of transaction
AUTO_RESPONSE	Auto-response
LATENCY	Latency
IDVD_LOGINTYPE	Login Type
ACTD_BANKACCTNO	Account's bank account number
ACTD_ACCTTYPENM	Account type
ACTD_AVAILABLEBL	Available balance
TRNSD_BENEFICIARYSORTCD	Beneficiary sort code
TRNSD_BENEFICIARYACCTNO	Beneficiary account number
TRNSD_TRNSAM	Transaction amount
TRNSD_PAYMENTDT	Transaction Datetime
TRNSD_TXNREFERENCETX	Transaction reference
TRNSD_PAYMENTDT	Transaction Date Time
IDVD_AUTHENTICD	Authentication code
IDVD_INTESESSIONID	Internet session ID
IDVD_IPADDRESSID	IP address
IDVD_CLIENTSCREENRESOID	Client screen resolution
IDVD_USERAGENTTX	User-agent
IDVD_DEVICEID	Device ID
IDVD_INTESESSIONID	Internet session ID
IDVD_CLIENTSCREENRESOID	Client screen resolution
IDVD_USERAGENTTX	User-agent
IDVD_BROWSERLANGTX	Browser language
IDVD_IPADDRESSID	IP address
IDVD_DEVICEID	Device ID
IDVD_TELSESSIONID	Telephone session ID
IDVDATA_TRNSTS	Transactions timestamps
EVENT	Event of transaction
AUTO_RESPONSE	Auto-response
Last_LATENCY	Latency
IDVD_LOGINTYPE	Login Type
IDVD_AUTHDETAILS1	Authentication details
Is Fraud	Fraud flag whether fraud or not

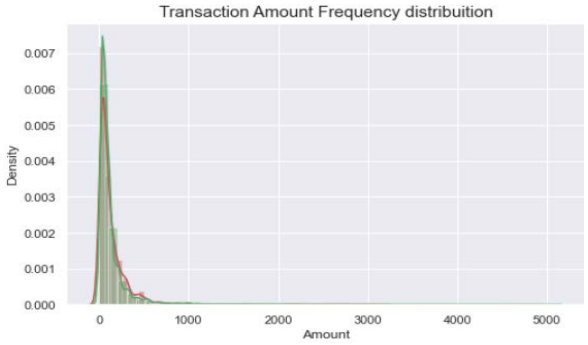


Fig. 8. Fraudulent and Customer's Distributions of Transaction Amount.

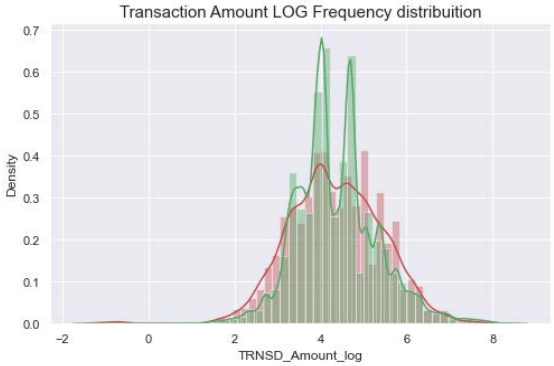


Fig. 9. Distributions of Transaction Amount with Log Transformation.

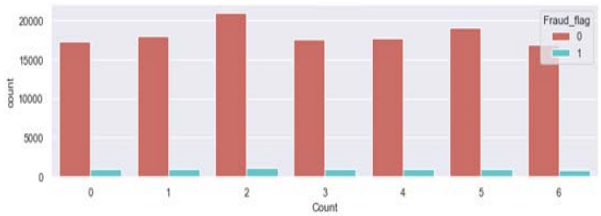


Fig. 10. Transactions base on Weekdays.

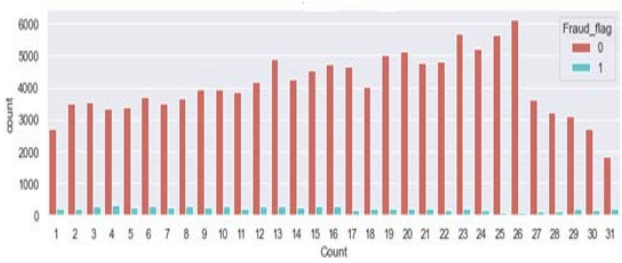


Fig. 11. Transactions base on Days.

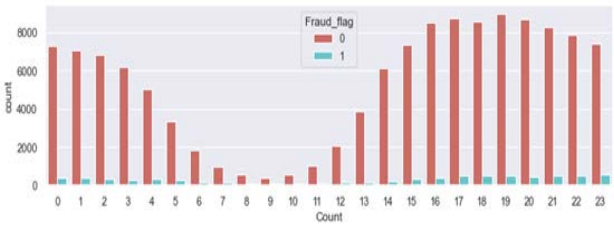


Fig. 12. Transactions base on Hours.

VI. Experiments and Results

A. Experiments

From our previous published work in [19], it has been demonstrated that the performance of fraud detection models with prepared feature sets, performs better than the models with original data only.

The purpose of the experiments in this research is to verify the effectiveness of using a feature set that is created through the processes using our framework for deep learning. The online banking dataset described in Section 3 is used. Target attributes in original data that are used for implementation of feature aggregation and transformation methods are almost fixed specifically for online banking transaction data. This is because the banking system has common attributes in some tables such as customer information, banking information, network information. According to processes in the feature creation component, we apply feature engineering methods on the original data and create 65 new features after aggregating features and transforming features (see Table IV).

TABLE IV. Created New Features

New Attributes Created by Feature Engineering
LATUPDATE_Weekdays
LATUPDATE_Hours
LATUPDATE_Days
LATUPDATE_Minute
BALANCE_log
Balance_min_mean
Balance_min_std
Amount_log
AUTO_RESPONSE_std
ACTD_AVAILABLEBALANCE_log
ACTD_AVAILABLEBALANCE_min_mean
ACTD_AVAILABLEBALANCE_min_std
Trans_min_mean
Trans_min_std
mean_last
mean_last_count
mean_balance
min_last
min_balance
min_last_balance
max_last
max_balance
max_last_baalance
count_last
count_balance
count_last_balane
mean_last
mean_balance
mean_last_balance
max_last
max_balance
max_last_balance
CUSTD_PARTYID_IDVD_CLIENTSCREENRESOID_count_LATUPDATE_Weekdays
CUSTD_PARTYID_IPADDRESSID_count_LATUPDATE_Weekdays
CUSTD_PARTYID_IDVDATE_TRNSTS_count_LATUPDATE_Weekdays
CUSTD_PARTYID_LAST_LATENCY_count_LATUPDATE_Weekdays
CUSTD_PARTYID_TRNSD_TRANSSESSIONCD_count_LATUPDATE_Weekdays
CUSTD_PARTYID_TRNSD_Amonut_log_count_LATUPDATE_Weekdays
CUSTD_PARTYID_ACTD_AVAILABLBALANCE_log_count_LATUPDATE_Weekday
CUSTD_PARTYID_ACCESS_CD_count_LATUPDATE_Weekdays
CUSTD_PARTYID_TRNSD_Amount_log_count_LATUPDATE_weekdays
CUSTD_PARTYID_TRNSD_Amount_count_LATUPDATE_Weekdays
LATENCY1_std
LATENCY2_std
LAST_LATENCY_std
LGIN_LATENCY1
LGIN_LATENCY2
LATENCY1_std
LATENCY2_std
clusters_1
clusters_2
clusters_3
count_cluser
Days_std
Weekday_std
Hours_std
PCA_EVENT0
PCA_EVENT1
PCA_PASS0
PCA_PASS1
PCA_PASS2
PCA_FinancialInfo0
PCA_FinancialInfo1
PCA_FinancialInfo2
PCA_CustomerID_IP_Amount

The result of feature importance measurement is presented in Table V. We measured the feature importance of all features both original and the created features and recognized that the most of features with higher importance rate are the new features created via the feature engineering framework. Based on higher scores, we selected 57 features among all 104 features and the rest of feature's importance rate were nearly equal to zero.

TABLE V. FEATURE IMPORTANCE MEASUREMENT (TOP30)

Attribute	Importance
count_last_balance	0.103799286
CUSTD_PARTYID_ACCESS_CD_count _LATUPDATE_Weekdays	0.095292695
min_last_balance	0.094270386
count_balance	0.091391504
count_last	0.07332497
TRNSD_BENEFICIARYACCTNO	0.064115062
BALANCE_log	0.060077297
Balance_min_mean	0.05803433
IDVDATA_TRNSTS	0.04700098
CUSTD_PARTYID	0.041534992
CUSTD_PARTYID_TRNSD_Amount_log_count _LATUPDATE_Weekdays	0.041354892
mean_last_balance	0.025604612
mean_balance	0.024146388
min_last	0.015255225
ACTD_AVAILABLEBLBALANCE_min_mean	0.014513752
ACTD_AVAILABLEBLBALANCE_log	0.013771143
IDVD_SCREENSIZE	0.013566704
TRNSD_TRANSSESSIONCD	0.011465614
LATENCY1_std	0.010989958
PCAID_D0	0.010124042
CUSTD_PARTYID_ACTD_AVAILABLEBLBALANCE _log_count_LATUPDATE_Weekdays	0.009920134
LGIN_LATENCY1	0.009015999
max_last_balance	0.008283598
ACTD_AVAILABLEBLBALANCE_min_std	0.007470134
CUSTD_PARTYID_TRNSD_TRANSSESSIONCD _count_LATUPDATE_Weekdays	0.004289627
PCAID_D1	0.003624484
PCA_PASS0	0.003490054
LATUPDATE_Days	0.003359021
max_last	0.00317593
Days_std	0.003111183

Now, three different deep learning models are built with three types of feature sets: (1) original dataset only, (2) original dataset plus newly created features, (3) only selected features following feature importance scores (see Tables VI to VIII).

We then use Tensor Flow which provides a simple autoencoder program from Python libraries.

Autoencoder requires the setting of some parameters and we manually determine optimal parameter values. The Autoencoder algorithm is applied in the settings below:

1) The data are divided into 80% training data and 20% testing data. The training data consists of customer transactions only excluding fraudulent data. In the testing, the autoencoder encodes and compresses the input data and tries to represent the original data based on leaned dimensional reduction and reconstruction. Then, it can distinguish a fraudulent transaction if it cannot represent the data again.

2) The number and size of layers are set from left to right 57-18-10-6-6-10-18-57 in the case of the selected feature set. These numbers show how to encode and decode in the neural networks. From the fifth to the eighth layers the data is reconstructed, and the mean squared error as a loss function is calculated. The significant point in the layers is that the number of input data size is the same as the output data size.

TABLE VI. PARAMETERS OF AUTOENCODER

Parameter Name	Value
Optimizer	Adam Optimize
Loss Function	Mean_Squared_Error
# of Epoc	1000
Batch Size	128
Test_size	0.2

TABLE VII. AUTOENCODER MODEL USING TENSORFLOW

```
input_layer = Input (shape= (input_dim,))
encoder = Dense (encoding_dim, activation=" tanh", activity_regularizer =
regularizers. l1(learning_rate)) (input_layer)
encoder = Dense (hidden_dim1, activation =" elu") (encoder)
encoder = Dense (hidden_dim2, activation =" tanh") (encoder)
decoder = Dense (hidden_dim2, activation = "elu") (encoder)
decoder = Dense (hidden_dim1, activation = "tanh") (decoder)
decoder = Dense (input_dim, activation =" elu") (decoder)
autoencoder = Mode (inputs = input_layer, outputs = decoder)
```

TABLE VIII. AUTOENCODER LAYERS (SELECTED FEATURES)

Layer (type)	Output Shape	Param #
input_1 (Input Layer)	[(None, 57)]	0
dense_(Dense)	(None, 18)	1044
dense_1 (Dense)	(None, 10)	190
dense_2 (Dense)	(None, 6)	66
dense_3 (Dense)	(None, 6)	42
dense_4 (Dense)	(None, 10)	70
dense_5 (Dense)	(None, 57)	627
Total params: 2,039 Trainable params: 2,039		627

B. Performance Metrics for Fraud Detection Models

Classification problems using unbalanced labelled data cannot be evaluated by the accuracy only. Especially in the case of financial fraud detection, we should evaluate and compare the model performance with plural metrics because the classification problem is necessary to be considered as a balance between the true positives ratio (TP) and the false-positive ratio (FP). TP is the number of predictions as fraud where the actual result is also fraud. FP is the number of predictions as a legitimate transaction where the actual result is the customer. The true negatives (TN) and the false negatives (FN) are also significant metrics when measuring the performance of recall and precision. A recall is the ratio of frauds that are perfectly classified whereas precision is the ratio of the accuracy of fraud predictions. When the score of recall is high, it indicates a poor rate of FN which is the number of predictions as a legitimate transaction where the actual result is fraud. When the score of precision is high, it indicates a poor rate of FP which is the number of predictions as fraud where the actual result is the customer.

F1-Measure is the harmonic mean of precision and recall. The best score is 1 whereas the worst score is 0. This metric seeks the balance between precision and recall (see Table IX).

TABLE IX. PERFORMANCE METRICS DEFINITION

Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
F1-measure	2*Precision*Recall/ (Precision + Recall)

The confusion matrix shows a matrix describing the performance of the model using True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) (see Table X).

TABLE X. CONFUSION MATRIX

# of Observations	**Predicted Normal**	**Predicted Fraud**
Actual Normal	TN	FP
Actual Fraud	FN	TP

C. Results and Evaluations

The effectiveness of the feature engineering framework is measured by comparison with the performance of the autoencoder model with the original data only. As stated in the previous section, in order to evaluate the efficiency of the created and selected features, the model performance is assessed by AUC, recall, precision, and F-measure. The following Table XI shows the comparison of autoencoder models with two threshold values (Threshold=4 and 1) with three different types of datasets.

TABLE XI. COMPARISON OF AUTOENCODER MODELS WITH THRESHOLD VALUE =4 IN THE DIFFERENT DATASETS

Threshold=4	**AUC**	**Recall**	**Precision**	**F-measure**
Model1 with only original data	0.65	0.058	0.188	0.0896
Model 2 with original data plus new features	0.83	0.064	0.215	0.0986
Model 3 with the selected features	0.92	0.064	0.215	0.0986

The results in Table XI shows that model 3 with the selected features and model 2 with original data plus newly created features are higher in all performance metrics than model 1 (with original data only). Model 3 has a higher AUC than Model 2.

Different threshold values are chosen based on the situation shown in Fig. 13. We can adjust the threshold value to a better classification part.

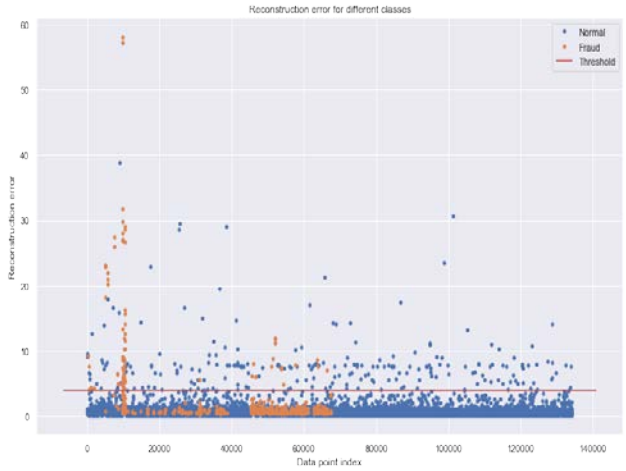


Fig. 13. Data Distribution in Threshold 4.

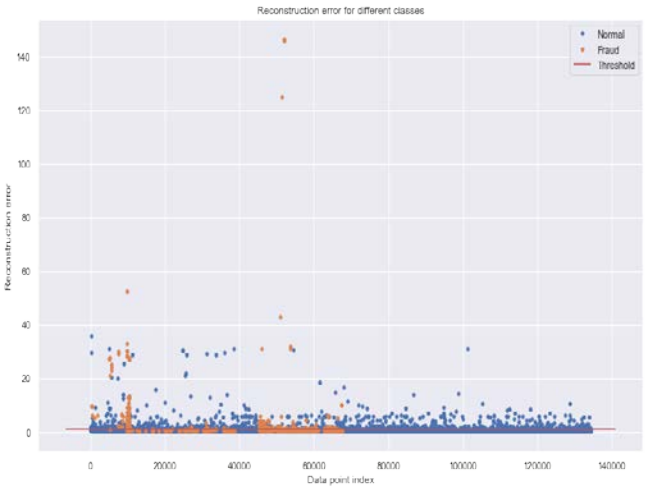


Fig. 14. Data Distribution in Threshold 1.

The data distribution presents the thresholds between fraudulent data and legitimate data. In the case of threshold 4, most fraud transactions are not classified well with the confusion matrix confirming this. Now let's change the threshold value from 4 to 1. Fig. 14 shows the data distribution when the threshold value is equal to 1.

It appears that fraud transactions with the threshold value of 1 are better classified than the ones with a threshold equal to 4. The performance of the models with a threshold value of 1 becomes as described in the table below:

TABLE XII. COMPARISON OF AUTOENCODER MODELS WITH THRESHOLD VALUE=1 IN THE DIFFERENT DATASETS

Threshold=1	AUC	Recall	Precision	F-measure
Model1 with only original data	0.73	0.161	0.104	0.1263
Model 2 with original data plus new features	0.91	0.358	0.451	0.3994
Model 3 with the selected features	0.96	0.648	0.430	0.5167

Table XII demonstrates the superiority of Model 3 with selected features.

From the above results two suggestions are drawn. First, the performance of the autoencoder models significantly improves when the new features are created based on the proposed feature engineering framework. The experiments indicate that it is efficient for a deep learning model (for classification) to implement feature engineering on original data before inputting the data. Second, adjustment of threshold in autoencoder also made an impact on the model accuracy. A combination of an appropriate setting of threshold and optimal feature set can improve a deep learning model performance for fraud classification.

Another point of view from the experiments is about scores of Recall and Precision of each model. Recall has an impact on huge money loss whereas Precision influences customer satisfaction and confidence. All measurements of the models using the selected feature set are the highest scores than model 1. Precision in model 2 is higher than the precision in model 3 which indicates that a balance between precision and recall is a trade-off. As mentioned above, a high score of recall indicates the model can identify fraudulent activities without mislabeling them as actual customers. On the other hand, a high score of precision means the model can identify actual customers' transactions without mislabeling them as fraud. In either case, using the created features in the newly built feature engineering framework could improve the model performance and detect fraudulent transactions based on the reconstruction of the input data.

VII. CONCLUSION

A fraud detection system in recent years adopts machine learning models which learn anomaly data patterns from past transaction records. However, the total losses through online banking in the United Kingdom have been increasing because fraud schemes continue to further evolve as the online payment system advances. Feature engineering is a key to improving the accuracy of fraud detection models and can reveal latent data patterns by transforming raw data into another dimension. In the paper, we used the feature engineering framework which creates new features and selects effective features through feature engineering techniques for autoencoder, a deep neural network, this time. As a result, the performance of the autoencoder models built with selected features from the framework was better in comparison to the performance of the autoencoder models built with raw data only.

Deep learning methods have a function of feature extraction to reduce the number of features in an input data and automatically learns features at multiple levels by combining the input features. Although they already have a part of feature engineering function in the algorithms, using the prepared dataset including new features created through the feature engineering framework was more effective for improving the deep learning model performance.

In this paper, we used an autoencoder as a deep learning model and presented the effectiveness of using the feature engineering framework. In further work, we will use other deep learning algorithms such as recurrent neural network (RNN) and convolutional neural network (CNN) which are often used for financial fraud detection. Moreover, the feature selection component in the framework will be studied and improved.

REFERENCES

[1] FRAUD – THE FACTS 2021:The definitive overview of payment industry fraud:UK Finance, 2021.

[2] Casilda Aresti, "Technology and operations management: PayPal's Use of Machine Learning to Enhance Fraud Detection" (PayPal's Use of Machine Learning to Enhance Fraud Detection (and more) - Technology and Operations Management (hbs.edu)) , 2018.

[3] Niccolo Mejia, "AI-Based Fraud Detection in Banking – Current Applications and Trends" (AI-Based Fraud Detection in Banking – Current Applications and Trends | Emerj), 2020.

[4] J.Jordan. "Introduction to autoencoders" [Cited:14/04/2021] https://www.jeremyjordan.me/autoencoders/. 2018.

[5] P. Jiang, J. Zhang and J Zou. "Credit Card Fraud Detection Using Autoencoder Neural Network", Cornell University, arXiv.org. vol. 1, site:1908.11553, 2021.

[6] M.A. AI-Shabi. "Credit Card Fraud Detection Using Autoencoder Model in Unbalanced Dataset", ELSEVIER, Expert Systems With Applications vol.167, pp.254–262, doi: 10.1016/j.procs.2020.03.219, 2019.

[7] S.Misra, S.Thakur, M.Ghosh, K.Saha. An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction 2020.

[8] A. Pumsirirat and L.Yan. "Credit CardFraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine", International Journal of Advanced Computer Science and Applications(IJACSA), Volume 9 Issue 1, 2018.

[9] Y. Lucas, P.Portier and S.Calabretto. "Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs", ELSEVIER, Expert Systems With Applications vol.102, pp.393–402, doi: 10.1016/j.future.2019.08.029, 2020.

[10] X.Zhang, Y.Han, W.Xu and Q.Wang. "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture", ELSEVIER, Expert Systems With Applications vol.557, pp.302–316,doi: 10.1016/j.ins.2019.05.023, 2021.

[11] J.S. Kalwihura and R. Logeswaran. "Auto-Insurance fraud detection: a behavioural feature engineering approach",Cornell University, arXiv.org. site: 1805/1805.09741.pdf, 2020.

[12] A.C. Bahnsen, D.Aouada and S.B. Ottersterm. “Feature engineering for credit card fraud detection”, ELSEVIER, Expert Systems With Applications vol.51, pp.134–142, 2016,doi: 10.1016/j.eswa.2015.12.030.

[13] M.Mwadulo. “A review on feature selection methods for classification tasks”, 2016, International Journal of Computer Applications Technology and Research, Vol. 5, Issue 6, pp. 395-402, 2016, ISSN:-2319–8656.

[14] A. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty and V. Sravan Kiran, "Similarity Based Feature Transformation for Network Anomaly Detection," in IEEE Access, vol. 8, pp. 39184-39196, 2020, doi: 10.1109/ACCESS.2020.2975716.

[15] R.Wedgy, J.Max.Kanter, K.Veeramachaneni, S.M.Rubio and S.l.Perez. “Solving the false positive problem in fraud prediction using animated feature engineering.”, Cornell University, arXiv.org, Computer Science, doi: 10.1007/978-3-030-10997-4_23.

[16] P.Jiang, J.Zhang and J.Zou. ”Credit card fraud detection for autoencoder neural network.”, Cornell University, arXiv.org. site:1908.11553v1,2018.

[17] A.Pumsiriat and L.Yan. “Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine.”, 2018, the International Journal of Advanced Computer Science and Applications(IJACSA),vol. 9, doi:10.14569/IJACSA.2018.090103.

[18] M.A.Al-Shabi. “Credit card fraud detection using auto encoder model in unbalanced datasets.”, 2019, Journal of Advances in Mathematics and Computer Science, pp. 1-16, doi: 10.9734/jamcs/2019/v33i530192.

[19] C.Ikeda K.Ouazzane and Q.Yu, “A new framework of feature engineering for machine learning in financial fraud detection.”, 2020 10th International Conference on Artificial Intelligence, Soft Computing and Applications, London, 2021 vol 10, doi:10.5121/csit.2020.101517.

Multifractal Analysis of Heart Rate Variability by Applying Wavelet Transform Modulas Maxima Method

Evgeniya Gospodinova, Galya Georgieva-Tsaneva, Penio Lebamovski
Institute of Robotics, Bulgarian Academy of Sciences, Sofia, Bulgaria

***Abstract*—The analysis of heart rate variability is based on the intervals between the successive heartbeats and thanks to it information about the functional state of the person can be obtained and the dynamics of its change can be traced. The nonlinear dynamics methods provide additional, prognostic information about the patient's health, complementing traditional analyses and are considered potentially promising tools for assessing heart rate variability. In this article, studies have been carried out to identify the mono- and multifractal properties of two groups of people: healthy controls and patients with arrhythmia using Wavelet Transform Modulas Maxima Method. The obtained results from the studies show that for healthy subjects the multifractal spectrum is broader than the spectrum of patients with arrhythmia. The value of the Hurst exponent is lower in healthy controls, and in patients with arrhythmia this parameter tends to one. For the healthy subjects, the scaling exponent showed nonlinear behaviour, while for patients with arrhythmia it was linear. This indicates that heart rate variability in healthy controls has multifractal behaviour while patients with arrhythmia have monofractal behaviour. The finding may be useful in diagnosing subjects with cardiovascular disease, as well as in predicting future diseases, as the heart rate variability changes at the slightest deviation in the health status of subjects before the onset of relevant signs of the disease.**

***Keywords*—RR time series; heart rate variability; wavelet transform modulas maxima method; monofractal; multifractal**

I. Introduction

The effectiveness of the modern medical technologies is closely linked to the improvement of the methods and the instruments for monitoring and analysing the condition of patients during their treatment. In medicine the problem with patients' clinical surveillance occupies a special place, as the monitoring of their current state can be of vital importance.

The use of electrocardiographic data for analysis of the cardiac activity of the patient is a generally accepted method. The presentation of heart rhythm as a dynamic row of RR time intervals (the distances between the R-tops of the electrocardiogram) and the mathematical analysis of this data [1] is widely used in the research of the cardiac activity. Based on the RR interval series, heart rate variability (HRV) is determined, which is one of the most accessible physiological parameters [2] reflecting the processes of autonomic regulation in the cardiovascular system. The dynamic characteristics of the heart rate make it possible to assess the severity of changes in the sympathetic and parasympathetic activity of the autonomic nervous system in changing the patient's health. The sympathetic branch reduces the intervals between heartbeats [3], while the parasympathetic branch increases them.

In modern cardiology, more and more attention is paid to the analysis of heart rate variability and more specifically to the changes in heart rate intervals [4]. Heart rate is the most objective characteristic of the functional state of the human body and depends on several factors: age, gender, environmental conditions, stress, body temperature, etc. [5, 6]. HRV analysis is a unique diagnostic technique that allows not only to assess the functional state of the human body, but also to monitor its dynamics and to identify the occurrence of pathological conditions when they are at a very early stage. The heavy physical work, the psychological stress, as well as the disease states of the human body lead to an increase in heart rate and to a decrease in HRV. Conversely, when the body is at rest, the heart rate is usually lower and the HRV is higher.

The mathematical methods for assessing the functional state of the human body through the study of HRV are combined into the following two groups: linear and nonlinear methods.

The linear methods include time-domain analysis and frequency-domain analysis. These methods are standardized, knowing the reference values of the studied parameters, but this is often not enough to characterize the complex dynamics of the RR time series of heart rate.

The nonlinear methods such as: Poincare plot, Detrended Fluctuation Analysis (DFA), Multifractal Detrended Fluctuation Analysis (MFDFA), Wavelet Transform Modulas Maxima Method (WTMM), AppEn, SampEn [7, 8, 9, 10] and others are not standardized, which is the reason for their limited use in clinical practice. The nonlinear methods for HRV analysis are based on the theory of chaos and fractals. These methods are in the process of active research, and it is expected that in the near future they will be able to give a new idea of the dynamics of heart rate in the context of physiological changes in patients with cardiovascular disease. Practically, each cardio interval contains elements of nonstationarity (fractal components) and for their evaluation in recent years methods of nonlinear dynamics are actively applied. These methods provide additional prognostic information of the studied signals, which complements the traditional analyses in the time and frequency domains.

This article is funded by the National Science Fund of Bulgaria (BNSF).

To be able to deal with the problem of the accurate HRV assessment of the studied cardiac signals, it is necessary to choose an appropriate method of analysis that represents the dynamics of the heart rate.

The aim of this article is to analyse and evaluate HRV in two groups of people: healthy controls and patients with arrhythmia, using the Wavelet Transform Modulas Maxima method. The effectiveness of the method used was evaluated by statistical t-analysis.

The rest of the paper is organized as follows: Section II provides an overview of related research in the scientific literature. Section III focuses on the Wavelet Transform Modulas Maxima Method used in this paper for HRV analysis. Section IV describes the data used for the analyses performed. Section V presents the results and discussions. The final section (Section VI) of the article contains the conclusion that can be made from the obtained results.

II. Related Work

In recent years, there has been an active introduction of the mathematical methods of analysis in the medical practice. Many scientists have studied the complex nature of the changes in the parameters of electrocardiographic (ECG) signals using nonlinear dynamic methods. The fractal and multifractal approach in the analysis of the cardiological data allows to obtain new knowledge and assessments that give an idea of the nonlinear dynamic processes taking place in them. As a result of the work of [11] on the heartbeat dynamics, the multifractal analysis has become a widely used tool for applied research, in cases where the non-stationary processes are limited by the application of the classical methods of analysis.

The multifractal analysis expands the possibilities for cardio-diagnostics based on the wavelet theory. The proposed approach by [12] for multiresolution wavelet analysis of heartbeat intervals allows distinguishing healthy patients from those with cardiac pathology. This universal approach is applicable in the analysis of non-stationary processes in the physical and biological sciences, including the analysis of ECG signals.

The authors in [13] proposed a technique for multifractal analysis to determine the degree of multifractality of the heart rate of patients suffering from partial seizures. The results show that the degree of multifractality varies depending on the severity of the disease.

The authors of [14] advise physicians to interpret the results obtained from fractal and traditional methods with caution, as they are still in the process of research and the measurements obtained are not fully described as biomarkers for clinical use.

In [15], the authors show how the entropy and the multifractal analysis can depict the dynamics of heart rate when students performed selective inhibition tasks. The results show that the entropy and the fractal markers outperform markers in the time and frequency range of the heart rate variability in distinguishing the cognitive tasks.

The WTMM method presented by Arneodo et al. [16] can be used to study the structures of inhomogeneous processes of various natures, based on wavelet and multifractal analysis.

Recent studies [17, 18] have shown that HRV changes in individuals infected with Covid-19 even before the onset of symptoms of the disease. This indicator can be useful for early detection of this disease.

The need to study HRV through the application of mathematical methods is determined by the fact that it accurately reflects the state of regulatory processes in the human body and provides information that is important in the diagnosis, prognosis, treatment, and prevention of diseases of various kinds.

III. Methodology

The choice of an appropriate mathematical method for time series analysis is determined by its flexibility and ability to be effectively applied to real processes. Among such universal methods for time series analysis, the WTMM method, which is discussed in this article, can be applied.

The WTMM method [19, 20, 21] is based on the mathematical tools of wavelet, fractal and multifractal analysis, which can reveal the hidden dynamics of the studied time series of heart rate in the context of mono- and multifractality.

The wavelet theory allows the studied signal to be analysed in more detail than Fourier analysis [22]. The wavelets are localized at both frequency and time, while the standard Fourier transform has only frequency localization.

The continuous wavelet transform (CWT) is used to decompose the continuous wavelet function. Unlike the Fourier transform, CWT provides the ability to build a time-frequency representation of the signal [23, 24], which achieves very good localization in time and frequency.

The dynamic characteristics of the RR interval series have fractal and in some cases multifractal properties [25, 26]. The fractal concept is related to processes (objects) that meet the following two criteria:

- Self-similarity: The process consists of many segments that are similar to each other and to the whole object.
- Fractional dimension: According to this criterion the fractal objects are distinguished from Euclidean ones, which are characterized by a dimension that is an integer.

The fractal processes are of two types: monofractal and multifractal [25]. The monofractal process is homogeneous in the sense that it has the same scaling properties, which both locally and globally can be characterized by a single scale indicator, such as: fractal dimension and Hurst exponent. Unlike the monofractal processes, the multifractal processes decompose into a large number of homogeneous fractal subsets, whose properties can be characterized by a spectrum of local fractal dimensions or Hearst exponents.

The behaviour of the studied signal by applying the WTMM method is performed in two stages [27, 28, 29, 30]. In the first stage, a continuous wavelet transform is performed according to the following formula:

$$W(a,b) = \frac{1}{a}\int_0^T x(t)\psi^*\left(\frac{t-b}{a}\right)d \quad (1)$$

Where:

- W are the wavelet coefficients;
- a is a scale parameter, $a \in R^+$;
- b is the translation coefficient, $b \in R$;
- x(t) is the input signal;
- ψ^* is a continuous function in the field of time and frequency, called the mother wavelet, which is complexly conjugated. The main purpose of the mother wavelet is to provide a function to generate daughter wavelets;
- T is the maximum value of time.

The scale parameter **a** can stretch or contract the signal under study. When the value of this parameter is small, the signal is compressed, which in turn leads to a more detailed graph. On the other hand, when the scale parameter is higher, the signal is stretched, which means that the resulting graph will be presented in lower detail.

To determine the singularity of the function, it is sufficient to use only the information about the maximum of the obtained wavelet coefficients, constructing the skeleton of the wavelet transform [31,32]. Following the lines of the skeleton, the behaviour of the singularities of the function x (t) can be traced [33].

The second stage of the WTMM method consists in the creation of partition functions Z(q, a), which allow to obtain reliable estimates of the characteristics of the studied process:

$$Z(q,a) = \sum_{l\in L(a)}(\sup_{a'\le a}|W(a',x_l(a'))|)^q \quad (2)$$

Where:

- L(a) is a set of all lines;
- l are the local maxima of the modules of the wavelet coefficients that exist for the scale a.

Equation (2) shows that the maximum value of the modulus is selected for each line at scales smaller than a set value of a. As a rule, it is expected that at small values of a, the partition function will have a power dependence, which will quantitatively characterize the scaling exponents τ(q):

$$Z(q,a) \sim a^{\tau(q)} \quad (3)$$

The scaling exponent τ(q) is defined by the following expression:

$$\tau(q) \sim log_{10}Z(q,a)/log_{10}a \quad (4)$$

By selecting different values of the parameter q, a linear or nonlinear dependence of τ(q) can be obtained, depending on the type of the studied process [34]:

- if the process is monofractal, then the function τ(q) is linear and the exponent h(q)=dτ(q)/dq=const;
- if the process is multifractal, then the function τ(q) is nonlinear and the exponent h(q)=dτ(q)/dq≠const.

By analogy with thermodynamic formalism, the spectrum of singularities is calculated on the basis of the Legender transformation:

$$D(h) = qh(q) - \tau(q) \quad (5)$$

The spectrum D (h) is determined by the set of values of the fractal dimensions of the original time series. The maximum of the spectral curve corresponds to the Hurst parameter. The following conclusions about the signal behaviour can be made from the value of the Hurst parameter:

- if $0<h<0.5$, then the signal has anticorrelation dynamics;
- if $0.5<h<1.0$, then the signal has a correlation behavior;
- if h=0.5, it lacks correlation in the signal.

IV. Data

The application of wavelet and fractal analysis of the studied RR interval series is designed to determine and visually assess the degree of harmonization of the studied time series that have fractal-like structures. The purpose of this analysis is to identify functional and pathological changes, as well as to predict changes in the health status of patients.

To test whether HRV can provide information outside of linear indices, the following two groups of people were examined in this study:

- healthy controls (10 men and 10 women aged 56 ± 4).
- patients with arrhythmia (10 men and 10 women aged 58 ± 3 years).

V. Results and Discussion

The software for the analysis of the fractal and multifractal properties of the test signals was created with MATLAB.

On Fig. 1A and Fig. 1B are shown RR interval series corresponding to the heart rate variations in a healthy individual and a sick patient with arrhythmia. Variations in a healthy individual are greater than those of a sick patient. This property can be used as a criterion in cardiovascular diagnosis. In practice, heart rate dynamics can be investigated using the linear methods by applying methods in the time- and frequency domains. The obvious shortcomings of these methods are that they can only be applied only to stationary time series and the heart rate shows heterogeneity and non-stationary of its fluctuation.

On Fig. 1C and Fig. 1D are shown the results of the wavelet transform and the graphics obtained have tree structures. The tree structure is more pronounced in the healthy patient. This property allows revealing the structure of the RR interval series and can also be used as a diagnostic criterion. The colour code of the graphics presents the values of wavelet

coefficients. The light colours correspond to the higher absolute values of the coefficients and darker colours correspond to the lower values.

Important information about the behaviour of the studied RR interval series is also contained in the wavelet skeleton of the local maxima lines on each scale of wavelet coefficients matrix (Fig. 1E and Fig. 1F). The local maximum modules of wavelet transformation |W(a, в)| there are the greatest values in those points of the analysed function in which it undergoes the most significant changes (jumps).

On Fig. 2A and Fig. 2B are shown the partition functions Z(q,a) for a healthy subject and for a patient with arrhythmia. The calculation of the Z(q,a) allow the signal fluctuations to be monitored. Positive values of the parameter q accentuate the large fluctuations of the signal (strong inhomogeneity), while the negative values of q accentuate on small fluctuations.

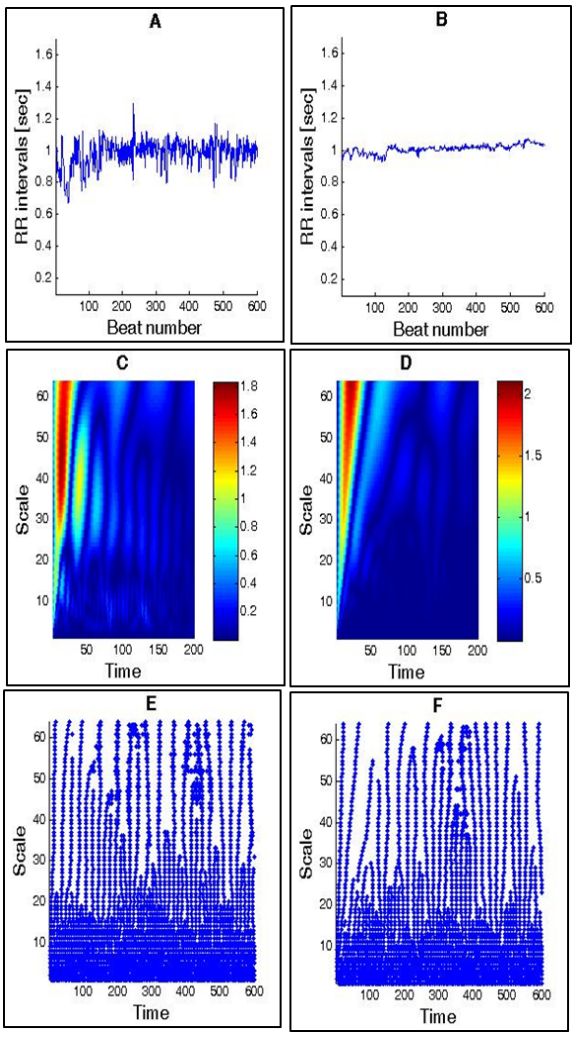


Fig. 1. RR Time Series for Healthy Subject (A) and for Subject with Arrhythmia (B); CWT Coefficients Plot in the Case of RR Intervals for Healthy (C) and Subject with Arrhythmia (D); WTMM Skeleton Plots for Healthy Subject (E) and Subject with Arrhythmia (F).

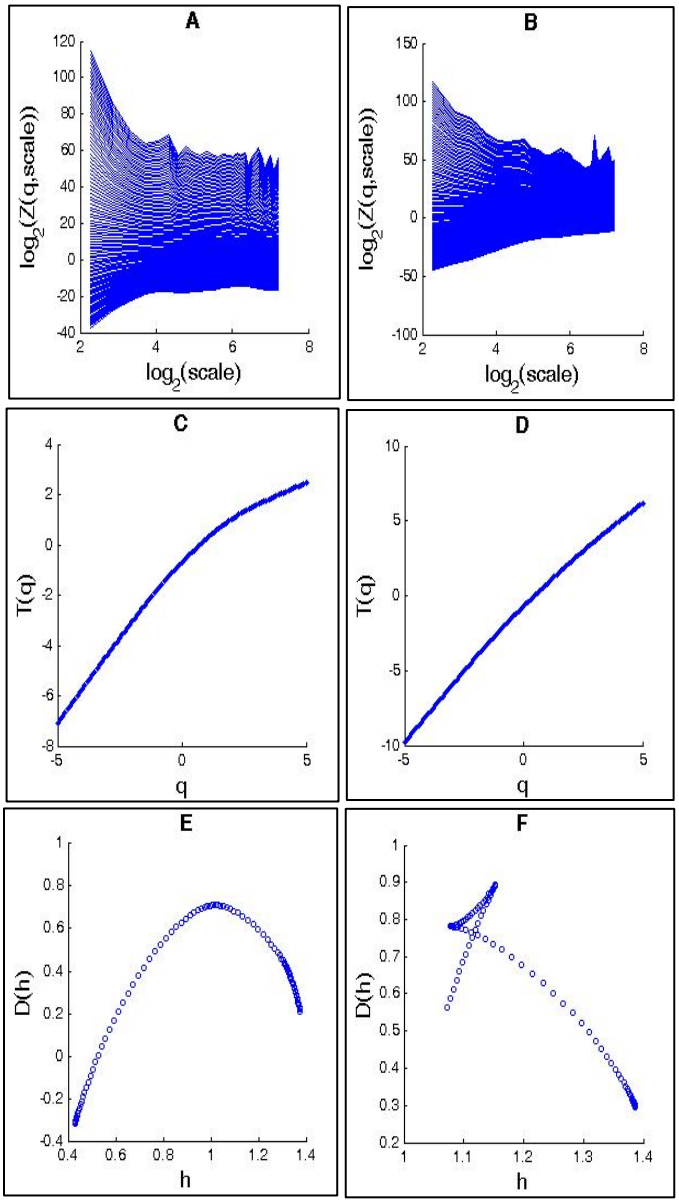


Fig. 2. Partition Functions for Healthy Subject (A) and for Subject with Arrhythmia (B); Scaling Exponent τ(q) for Healthy Subject (C) and for Subject with Arrhythmia (D); Multifractal Spectrum for for Healthy Subject (E) and for Subject with Arrhythmia (F).

The function τ(q) characterizes the fractal properties of the RR time series investigated. The graphics of the scaling exponent τ(q) have been shown on Fig. 2C and Fig. 2D for a healthy subject and for a patient with arrhythmia. For the healthy subject, the τ(q) spectrum has nonlinear behaviour and for the patient with arrhythmia this function is linear.

The statistical differences for multifractal spectrum between the RR time series of the two investigated subjects have been shown on Fig. 2E and Fig. 2F. The key characteristics are: the width of the spectrum of the singularity and the value of the exponent corresponding to the maximum value of the spectrum (Hurst exponent). For healthy people, the singularity spectrum is wide with non-zero singularities. On the other hand, for the patient with arrhythmia the singularity

spectrum is a very narrow range. In Table I is shown the values of the multifractal spectrum parameters: h_{max}, h_{min}, $\Delta h=h_{max}-h_{min}$ and the value of the Hurst exponent observed for healthy subjects and patients with arrhythmia. The results show that the width of the multifractal spectrum of the group: healthy controls are greater than the group: patients with arrhythmia. The value of the Hurst exponent is lower in healthy subjects. The studied parameters determined by t-test have statistical significance (p-value <0.05), therefore with these parameters the two studied groups can be distinguished.

The obtained results show that the heart rate of the healthy people is characterized by uneven dynamics in the form of multifractal behaviour, which cannot be detected by traditional methods, but can be observed using the WTMM method. In the pathological cases, the uneven dynamics of HRV is destroyed, which reflects the state of the cardiovascular system.

The results demonstrate the effectiveness of applying the WTMM method for HRV analysis as an additional measure that can expand and improve the information obtained from the RR time series in the diagnosis and prognosis of cardiovascular disease.

TABLE I. MULTIFRACTAL SPECTRUM PARAMETERS AND HURST EXPONENT

Parameter	**healthy controls N=20**	**patients -arrhythmia N=20**	**p-value**
	mean±std	*mean±std*	
h_{max}	1.42±0.12	1.31±0.15	0.01
h_{min}	0.30±0.20	0.8±0.14	0.0001
$\Delta h=h_{max}h_{min}$	1.0±0.22	0.6±0.20	0.0001
Hurst	0.72±0.10	0.93±0.3	0.01

VI. CONCLUSION

The results obtained in this article confirm the hypothesis that monofractality is a marker of pathological dynamics of heart rhythm in the case of cardiovascular disease such as arrhythmia. Conversely, it has been shown that multifractality is an indicator of a healthy individual. Therefore, HRV analysis using the WTMM method can be a useful approach to distinguish healthy controls from patients with arrhythmia, as the parameters studied have statistical significance (p-value<0.05). Interpretation of the results of this type of analysis may be useful before the possibility of using this method for physiological or clinical studies.

ACKNOWLEDGMENT

This research work was carried out as part of the scientific project “Investigation of the application of new mathematical methods for the analysis of cardiac data” No KP-06-N22/5, date 07.12.2018, funded by the National Science Fund of Bulgaria (BNSF).

REFERENCES

[1] F. Shaffer, J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” Front Public Health, 5:258. doi: 10.3389/fpubh.2017.00258. PMID: 29034226; PMCID: PMC5624990, September 2017.

[2] F. Sessa, V. Anna, G. Messina, G. Cibelli, V. Monda, G. Marsala, M. Ruberto, A. Biondi, O. Cascio, G. Bertozzi, D. Pisanelli, F. Maglietta, A. Messina, M. P. Mollica, M. Salerno, “Heart rate variability as predictive factor for sudden cardiac death,” Aging, 10(2), pp. 166–177, 2018. https://doi.org/10.18632/aging.101386.

[3] B. Grässler, B. Thielmann, I. Böckelmann, A. Hökelmann, “Effects of Different Training Interventions on Heart Rate Variability and Cardiovascular Health and Risk Factors in Young and Middle-Aged Adults: A Systematic Review”, Frontiers in Physiology, vol.12, pages=532, 2021 https://www.frontiersin.org/article/10.3389/fphys.2021.657274.

[4] S. U. Marasingha-Arachchige, J. Á. Rubio-Arias, P. E. Alcaraz, L. H. Chung, “Factors that affect heart rate variability following acute resistance exercise: A systematic review and meta-analysis,” Journal of Sport and Health Science, 2020, ISSN 2095-2546. https://doi.org/10.1016/j.jshs.2020.11.008.

[5] F. Shaffer, J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” Frontiers in public health, vol. 5, pages=258, 2017. https://doi.org/10.3389/fpubh.2017.00258.

[6] R. McCraty, F. Shaffer, “Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk,” Glob Adv Health Med, vol.4, issue 1, pp. 46–61, 2015. https://doi.org/10.7453/gahmj.2014.073.

[7] Z. Germán-Salló, M. Germán-Salló, “Non-linear Methods in HRV Analysis”, Procedia Technology, Vol. 22, pp. 645-651, 2016, ISSN 2212-0173, https://doi.org/10.1016/j.protcy.2016.01.134.

[8] C. Fiskum, T. G. Andersen, X. Bornas, P. M. Aslaksen, M. A. Flaten, K. Jacobsen, “Non-linear Heart Rate Variability as a Discriminator of Internalizing Psychopathology and Negative Affect in Children With Internalizing Problems and Healthy Controls,” Frontiers in Physiology, vol. 9, pp.561, 2018. https://www.frontiersin.org/article/10.3389/fphys.2018.00561.

[9] P.Melillo, M. Bracale, L. Pecchia, “Nonlinear Heart Rate Variability features for real-life stress detection. Case study: students under stress due to university examination,” BioMed Eng OnLine, vol 10(96), 2011. https://doi.org/10.1186/1475-925X-10-96.

[10] T. Henriques, M. Ribeiro, A. Teixeira, L. Castro, L. Antunes, C. Costa-Santos, “Nonlinear Methods Most Applied to Heart-Rate Time Series: A Review,” Entropy (Basel, Switzerland), *22*(3), 309, 2020. https://doi.org/10.3390/e22030309.

[11] P. Ivanov, L. Amaral, A. Goldberger, S. Havlin, M. G. Rosenblum, Z. R. Struzik, H. E. Stanley, “Multifractality in human heartbeat dynamics”. Nature 399, 461–465, 1999. https://doi.org/10.1038/20924.

[12] S. Thurner, M. C. Feurstein, M. C. Teich, „Multiresolution Wavelet Analysis of Heartbeat Intervals Discriminates Healthy Patients from Those with Cardiac Pathology“. Phys. Rev. Lett. 80, 1544, 1998.

[13] D. Ghosh, S. Dutta, S. Chakraborty, S. Samanta, “Epileptic Seizure: A New Approach for Quantification of Autonomic Deregulation with Chaos Based Technique”. Transl Biomed. 2017, 8:1. https://doi.org/10.21767/2172-0479.100106.

[14] J. Sen, D. McGill. “Fractal analysis of heart rate variability as a predictor of mortality: A systematic review and meta-analysis”. An Interdisciplinary Journal of Nonlinear Science, Volume 28, Issue 7, 2018. https://doi.org/10.1063/1.5038818.

[15] P. Bouny, L. M. Arsac, C. R. Touré, V. Deschodt-Arsac, „Entropy and Multifractal-Multiscale Indices of Heart Rate Time Series to Evaluate Intricate Cognitive-Autonomic Interactions“. Entropy. 2021; 23(6):663. https://doi.org/10.3390/e23060663.

[16] A. Arneodo, B. Audit, E. Bacry, S. Manneville, J.F. Muzy, S.G. Roux, “Thermodynamics of fractal signals based on wavelet analysis: Application to fully developed turbulence data and DNA sequences”. Physica A-statistical Mechanics and Its Applications - PHYSICA A. 254. 24-45, 1998. https://doi.org/10.1016/S0378-4371(98)00002-8.

[17] MBA Mol, MTA Strous, FHM van Osch, FJ Vogelaar, DG Barten, M. Farchi, NA Foudraine, Y. Gidron, “Heart-rate-variability (HRV), predicts outcomes in COVID-19,” PLoS One., vol 16 (10), October 2021, doi: 10.1371/journal.pone.0258841. PMID: 34710127; PMCID: PMC8553073.

[18] S.-A. Ouadfeul, “Multifractal behavior of SARS-CoV2 COVID-19 pandemic spread, case of: Algeria, Russia, USA and Italy”, medRxiv, 2020. https://doi.org/10.1101/2020.09.16.20196188.

[19] H. Salat, R. Murcio, E. Arcaute, “Multifractal methodology”, Physica A: Statistical Mechanics and its Applications, vol. 473, pp. 467-487, 2017, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2017.01.041.

[20] A. Arneodo, B. Audit, P. Kestener, S. Roux, “Wavelet-based multifractal analysis”. Scholarpedia. 3. 4103, 2008. https://doi.org/10.4249/scholarpedia.4103.

[21] P. Ivanov, A. Luis, A. L. Goldberger, H. Shalomo, H. E. Stanley, Z. Struzik, "From 1/f noise to multifractal cascades in heartbeat dynamics," Chaos, vol 11(3), pp. 641-645q 2001. https://doi.org/10.1063/1. 1395631.

[22] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, L. Marcourt, “Fourier and wavelet transform analysis, a tool for visualising regular patterns in dna,” Journal of Theoretical Biology, 206(ARTICLE), 2000, pp.323-326.

[23] V. I. Kovalchuk, O. S. Svechnikova, L. A. Bulavin, “Multifractal Analysis of Cardiac Series and Predictors of Sudden Cardiac Death,” Ukrainian Journal of Physics, vol. 66, No.10, 879, 2021. https://doi.org/10.15407/ujpe66.10.879.

[24] D. Makowiec, R. Galaska, A. Rynkiewicz, J. Wdowczyk-Szulc, “Multifractal estimators of short-time autonomiccontrol of the heart rate”, Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 405–411, 2009.

[25] H. E. Stanley, L. A. Amaral, A. L. Goldberger, S. Havlin, P. Ivanov, and C. K. Peng, "Statistical physics and physiology: monofractal and multifractal approaches," Physica A, vol. 270, , pp. 309-324, 1999.

[26] E. Gerasimova, B. Audit, S. G. Roux, A. Khalil, O. Gileva, F. Argoul, O. Naimark, A. Arneodo, “Wavelet-based multifractal analysis of dynamic infrared thermograms to assist in early breast cancer diagnosis,” Frontiers in physiology, vol. *5*, 176, 2014. https://doi.org/10.3389/fphys.2014.00176.

[27] A. Puckovs, A. Matvejevs, “Wavelet Transform Modulus Maxima Approach for World Stock Index Multifractal Analysis,” Information Technology and Management Science, December 2012, https://doi.org/10.2478/v10313-012-0016-5.

[28] A. Arneodo, B. Audit, N. Decoster, J. F. Muzy, C. Vaillant, “A wavelet based multifractal formalism: application to DNA sequences, satellite images of the cloud structure and stock market data,” in The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes, eds A. Bunde, J. Kropp, and H. J. Schellnhuber (Berlin: Springer Verlag), pp. 26–102, 2002.

[29] J. F. Muzy, E. Bacry, A. Arneodo, “Wavelets and multifractal formalism for singular signals: application to turbulence data,” Phys. Rev. Lett. Vol. 67, Issue 25, pp. 3515–3518, 1991. https://doi.org/10.1103/PhysRevLett.67.3515.

[30] J.-F. Muzy, E. Bacry, A. Arneodo, “Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method,” Phys. Rev. E , vol. 47, issue 2, pp. 875–884, 1993. https://doi.org/10.1103/PhysRevE.47.875.

[31] J.-F. Muzy, E. Bacry, A. Arneodo, “The multifractal formalism revisited with wavelets,” Int. J. Bifurc. Chaos, vol. 4, pp. 245–302, 1994.

[32] R. Galaska, D. Makowiec, A. Dudkowska, A. Koprowski, K. Chlebus, J. Wdowczyk-Szulc, A. Rynkiewicz, “Comparison of wavelet transform modulus maxima and multifractal detrended fluctuation analysis of heart rate in patients with systolic dysfunction of left ventricle,” Ann Noninvasive Electrocardiol, vol. 13(2), pp. 155-64, 2008. doi: 10.1111/j.1542-474X.2008.00215.x. PMID: 18426441; PMCID: PMC6932668.

[33] P. S. Addison, “The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance”. CRC press; January 2017.

[34] P. Lio, “Wavelets in bioinformatics and computational biology: state of art and perspectives,” Bioinformatics, vol 19(1), pp.2-9, 2003.

Neural Network Model for Artifacts Marking in EEG Signals

Olga Komisaruk, Evgeny Nikulchev
MIREA — Russian Technological University
Moscow 119454, Russia

***Abstract*—One of the main methods for research of the holistic activity system of human brain is the method of electroencephalography (EEG). For example, eye movements, blink, hearth activity, muscle activity that affects EEG signal interfere with cerebral activity. The paper describes the development of an intelligent neural network model aimed at detecting the artifacts in EEG signals. The series of experiments were conducted to investigate the performance of different neural networks architectures for the task of artifact detection. As a result, the performance rates for different ML methods were obtained. The neural network model based on U-net architecture with recurrent networks elements was developed. The system detects the artifacts in EEG signals using the model with 128 channels and 70% accuracy. The system can be used as an auxiliary instrument for EEG signal analysis.**

Keywords*—Artifacts in EEG signal; neural network model; recurrent neural network; U-net architecture

I. Introduction

Electroencephalography provides quantitative and qualitative analysis of human brain functionality and its reactions to stimulants. Electroencephalogram (EEG) is important for brain activity and behavior recognition, but there are always artifacts in electrical activity records that have influence on EEG signal analysis.

Measuring instruments, including defective electrodes, disturbances and high electrode resistance can be the reason of artifact occurrence. These artifacts can be recognized by more accurate recording system, but physiological artifacts are more complex. The eye movements, blink, hearth activity, muscle activity that affects EEG signal interfere with neural activity and can be used as normal phenomenon [1].

Artifact is a signal, caused by an extracerebral source, observed during EEG recording. They identify physical and physiological causes of artifacts [2]. Artifacts obtained during an electroencephalographic investigation represent a recording defect [3]. Modern electroencephalographic equipment records extremely small values of changes in bioelectric potentials, and therefore the true EEG recording can be distorted due to the influence of a variety of physical (technical) or physiological artifacts [4]. In some cases, such artifacts can be removed using analog-digital converters and various filters, but if the artifact effect coincides in characteristics of wave frequency with a real EEG recording, then these methods become ineffective.

The most common physical artifacts are mains frequency, phone artifact, wire breakage, poor electrode contact, high resistance artifact.

The following physiological artifacts are often recorded: ECG artifact, vascular artifact, galvanic skin artifact, oculomotor artifact, electrooculogram, myographic artifact - electromyogam [2]. The appearance of such artifacts is due to various biological processes occurring in the patient's body.

An ECG artifact most often occurs in the examined patients suffering from increase in arterial pressure, mainly in monopolar and transverse biopolar leads [5]. Usually, its occurrence is associated with an increase in the activity of the sympathetic nervous system, which facilitates the conduction of an ECG signal to peripheral tissues. Galvanic skin artifact occurs due to the activation of the patient's parasympathetic nervous system and increased sweating. As a result, there is a general cyclical change in the resistance of the skin and the skin-electrode system [1]. An oculomotor artifact, an electrooculogram (EOG), appears as slow-wave oscillations in the frontopolar leads with a frequency of 0.3–2 Hz. The appearance of EOG is associated with a change in the position of the eyeball (retina). Myographic artifact occurs when the frontal, chewing and occipital muscles are strained. The appearance of such an artifact can be both a spontaneous stress of the patient and involuntary reaction to an overly tightly put on fixing electrodes system [6].

The use of machine learning methods and neural networks determines promising research in the field of automatic artifact detection. In neurocomputer technologies, there is a general training scheme [7], which is divided into a training set, in which optimization of parameters is carried out, and a test set, according to which the quality of the resulting model is assessed. At the stage of training, it is necessary to understand the signs by which the classifier will be trained [8].

The paper contains six sections. The second section presents the overview of the approaches used. The third section describes the source data for current study. The fourth section presents methods used in the study including the description of software and the types of the neural network architectures. In the fifth section, conclusions of the study are given. This section presents the result of the searching for effective architecture for a qualitative solution to the problem of searching for artifacts. The sixth section contains general conclusion.

II. Related Work

EEG is a tool for psychophysiologic researches. However, the record filtering is often accomplished by high qualified professionals and takes a lot of resources and special filtering techniques [9]. Under these conditions development of effective EEG data filtering methods is an urgent task.

Fast development of cheap high parallel computation infrastructure, powerful machine learning algorithms and big data caused a huge progress in deep learning. The modern approaches of automatic interpretation of EEG use modern techniques such as neural networks and support vector machine.

Machine learning and neural network techniques in particular [2] determine perspective in researches in the automatic artifact recognition domain.

There are five basic algorithms [10] that are widely used in classifiers:

- Linear classifier [11]. It is more popular in online applications including real-time applications. One of the most effective method is support vector machine that usually better than other classifiers.
- Neural networks [12]. The most frequent methods for time series analysis are such architectures as convolutional neural network and recurrent neural network.
- Non-linear classifier [13]. Common methods are hidden Markov models and Bayesian classifiers.
- K-means [14]. These classifiers are based on neighbor distance values.
- Classifier combinations [15]. This method combines different classifiers and demonstrates good efficiency for autonomous applications.

Due to real-time classification, described classifier methods are more optimal for EEG signal analysis.

The task of the machine in unsupervised learning is to find relationship between individual data, to identify patterns, to select patterns, to organize data or describe their structure, and to classify data.

One of the most known drawbacks of machine learning methods is that the source data for training and data for test belong to the same feature space and follow the same probability distribution.

The aim of the research is development of intelligent tools based on neural network technologies that can recognize artifacts in EEG obtained via 64-channel electroencephalograph.

III. Data

EEG data is recorded using electroencephalograph Brain Products, containing 128 channels, 64 sensors placed on the international system “10-10%”.

The aim of the experiment was analyzing brain activity zones in resting state and nonverbal intelligence dependencies.

The study was conducted in a sound-attenuated and electrically shielded dimly lit room. Impedance was kept under 25 kOhm with high conductive chloride gel. The time of EEG settling was approximately 15 minutes.

The BrainProducts PyCorder system was used as a data collection system. This system allows continuous recording without any filtering and continuous sampling at 500 Hz. The reference electrode was located at Cz. The data was re-referenced to the common reference after the recording and downsampled to 256 Hz. The data were filtered from 0.1 Hz to 30 Hz and then re-referenced to an averaged reference and manually cleaned from artifacts, with noisy channels excluded.

To remove blink and vertical eye-movement artifacts, independent component analysis (ICA) was performed on the following electrodes: VEOG — Fp1, HEOG — FT9 and FT10. After ICA, the excluded channels were topographically interpolated, and semiautomatic artifact rejection was conducted.

Dataset contains two types of files:

1) Edf files are source data of EEG recording process (see example in Fig. 1)

2) “Markers” files contain description of artifacts (see example in Fig. 2):

- type - type of interval;
- description - artifact description (for example, “Blink”);
- position - time of artifact appearance (unit of time represented in “SamplingInterval” field, that equals to 3.90625 ms);
- length - artifact duration;
- channel - channel name, representing the location of artifact (Fp1, Fp2 – “Blink”, All - artifact that appeared in all channels).

There are only two types of artifacts. Thereby, neural network will classify three classes: “Blink”, “Global artifact” and “Resting state” (when there are no artifacts).

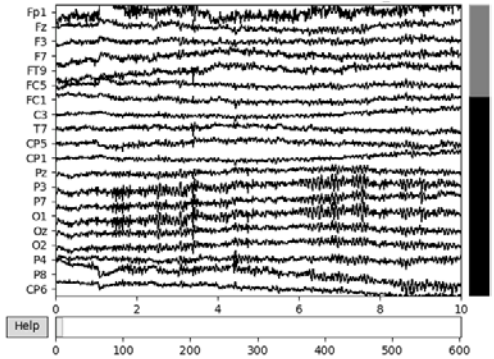


Fig. 1. Example of Edf File Format Content.

```
Sampling rate: 256Hz, SamplingInterval: 3.90625ms
Type, Description, Position, Length, Channel
New Segment, , 1, 1, All
Bad Interval, Userdefined, 32420, 1117, All
UserDefined, Blink, 33374, 63, Fp1
Bad Interval, Userdefined, 33470, 601, All
UserDefined, Blink, 33708, 84, Fp1
UserDefined, Blink, 34128, 55, Fp1
UserDefined, Blink, 34410, 59, Fp1
UserDefined, Blink, 34554, 53, Fp1
UserDefined, Blink, 35239, 83, Fp1
UserDefined, Blink, 35466, 55, Fp1
UserDefined, Blink, 35539, 57, Fp1
UserDefined, Blink, 35818, 52, Fp1
UserDefined, Blink, 35879, 57, Fp1
UserDefined, Blink, 36211, 80, Fp1
UserDefined, Blink, 36362, 98, Fp1
UserDefined, Blink, 37139, 52, Fp1
UserDefined, Blink, 37517, 57, Fp1
UserDefined, Blink, 38510, 58, Fp1
```

Fig. 2. Example of «Markers» File Format Content.

IV. Methods

To select a neural network model, it is necessary to conduct experimental studies of various architectures. An intelligent EEG signal analysis circuit has been developed (Fig. 3). Intelligent analysis of EEG signals consists of the process of recording and forming a database, processing signals and training a neural network model.

Recording process consists of taking readings using an electroencephalograph, the data of the electrodes located on the surface of the head are sent to the BrainProductsPyCorder data acquisition system. Next, expert analysis and processing of the generated database is carried out, in which different types of artifacts are marked, and then a new database is formed containing information about artifacts in each .edf file.

Based on the Database analysis, the size of input and output of neural network was determined. Pre-processing block reads Markers Database. Then, Data analyze block analyzes it. After that, train and test samples formed.

It is necessary to determine input and output. To find the solution, data was analyzed where distance between artifacts and maximum duration of every type of artifact were found. Also, quantity for every type of artifact was analyzed for data balance. For that, Data_analyzer.py library was created. The library consists of the following methods:

- max_artifact_length - returns maximum length of the artifacts;
- max_type_length - returns maximum length of the artifact of the specific type;
- channel_stats - based on markers data, it returns quantity of artifacts for every channel;
- normal-state-lengths - returns distances between artifacts (lengths of «resting state»);
- getMaxMin_by_edf - returns maximum and minimum values of frequency in edf file;
- getMaxMin_by_train - returns maximum and minimum values of frequency in input samples.

Using the described methods, the most optimal time window was selected for determining artifacts, based on the maximum length of the artifact Blinking (1.8 seconds) [16].

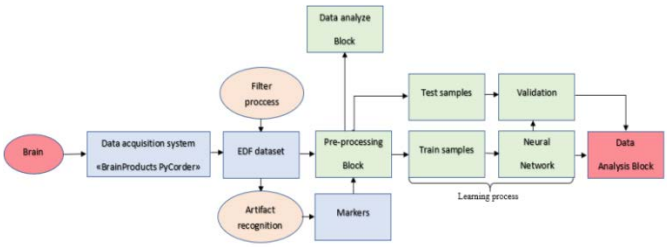


Fig. 3. EEG Signal Intelligent Analysis Scheme.

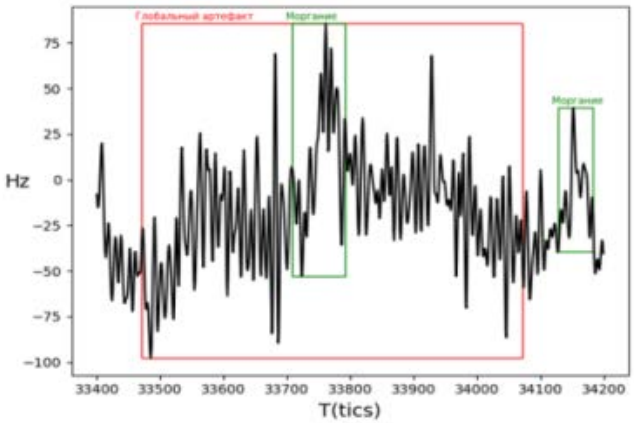


Fig. 4. Signal Graph with Artifact Marking.

It is necessary to determine number of channels that will be included in classification. Based on fact that most «Blinks» appear in Fp1 sensor, neural network can be trained only on one sensor. There are two artifacts: Blink and global artifact. Output of neural network consists of three classes: «Blink», «Global artifact» and «Resting state».

Samples were formed based on Markers database. Blink artifacts are put randomly in samples (Fig. 4).

Raw data in dataset still has noise. To filter the signal Fast Fourier Transform was implemented. The result is showed in Fig. 5.

Based on the developed mining analysis scheme presented in Fig. 4, it is necessary to develop an environment for conducting experiments. The interaction of software is shown in Fig. 6. Number of neural network models were trained.

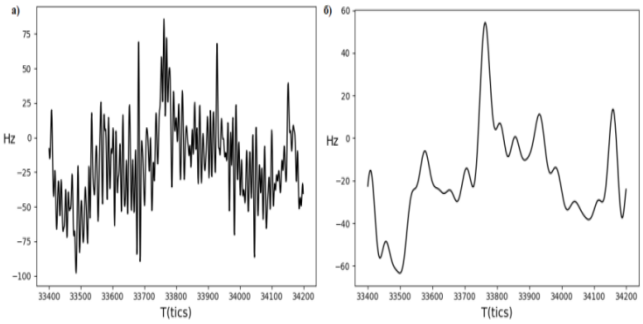


Fig. 5. Result of FFT: a) EEG Signal; b) EEG Signal with FFT.

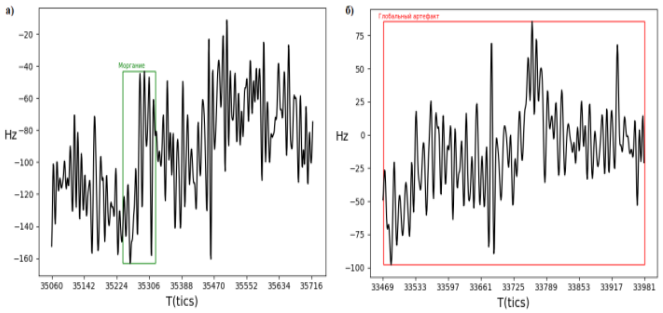


Fig. 6. Train Sample.

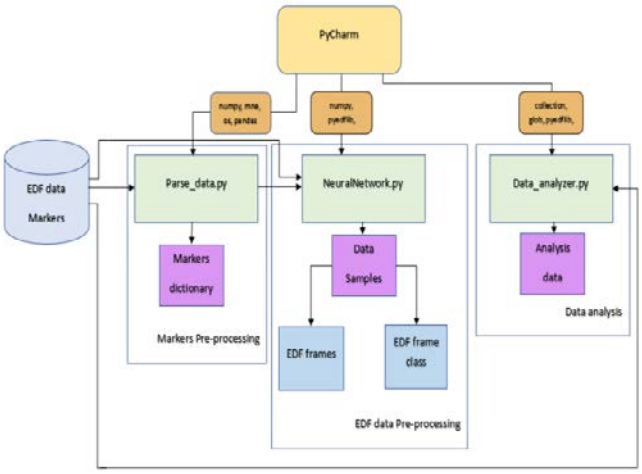


Fig. 7. Software Structure for the Analysis and Formation of Training Samples for a Neural Network Model.

In the PyCharm development environment, the main source code was developed to analyze and process the input values of the neural network model. The software tools interaction is shown in Fig. 7. The Parse_data.py library is used to convert Markers files into an associative array containing all the artifact information for each record in .edf files. The library contains the artifacts_supression method, which is used to translate the Position and Length format in seconds. In the read_markers_from_dir method, an associative array is formed from the specified directory using the pandas data analysis library containing the file name and its information about artifacts: the position of the artifacts, their description and length. This approach is used to obtain data in the function of generating training samples for a neural network.

The NeuralNetwork.py library allows creating samples for training a neural network based on arrays that are generated using the Parse_data.py library. The main method is prepare_data, which is based on information about artifacts, a database of EEG signals, used channels, and the size of the input window (in seconds) and a given ratio of samples with a normal state to samples with artifacts forms training samples for a neural network. Since the window size is larger than the maximum length of the Blinking artifact, this class is added to the selection completely. This takes into account the random shift of the artifact relative to the start of the sample. The Global artifact class is divided into several samples, from the beginning of the artifact to the sample that captures the end of the artifact and part of the signal without artifacts. In the process of recording samples with artifacts, the distance between them is calculated, and samples with the "normal state" class are taken, located between the artifacts. The Data_analyzer.py library contains the methods for analyzing the database described previously. An executable file "main.ipynb" was created in the Colaboratory environment, which contains the interactions of the libraries shown in Fig. 7, and also contains the architecture and process of training a neural network. The implementation scheme of an intelligent system for determining artifacts in an EEG signal is described in Fig. 8.

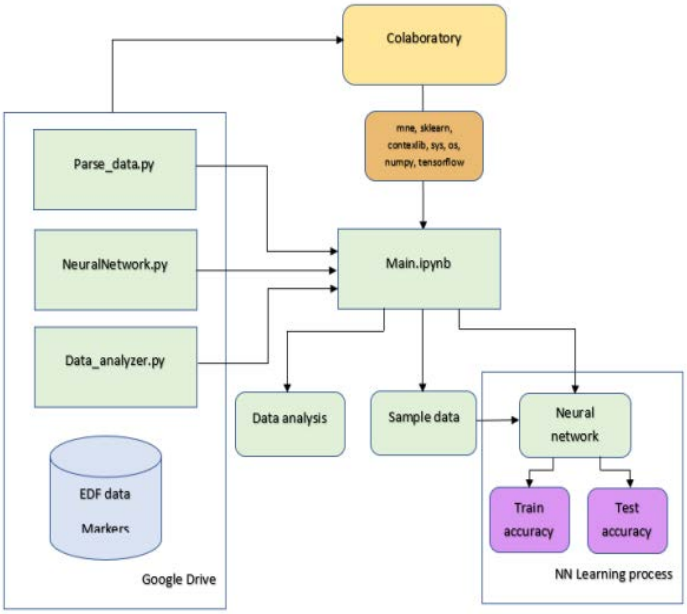


Fig. 8. Neural Network Training Scheme.

The neural network training scheme is an interaction of the libraries described earlier in the executing part of the Main.ipynb program. The executable file contains methods from the libraries for processing the database, conclusions of analytical data, sampling, the architecture of the neural network and its learning process located in the GoogleDrive cloud storage. The training process was conducted on the Colaboratory platform.

V. Experimental Selection of Neural Network Architecture

After analyzing the results of studies related to signal processing using neural networks, the architectures were selected based on convolutional and recurrent neural networks. Thus, 4 architectures were obtained:

- Batch_normalization + CNN + Dense using spectrograms;
- RNN (LSTM) + CNN + RNN (LSTM) + Dense;
- Batch_normalization + CNN + Dense;
- LSTM + NN based on "U-net"

1) Batch_normalization + CNN + Dense using spectrograms: The signal was converted to a spectrogram, a corresponding function was created using the fast Fourier transform (performed using the spectrogram method of the Scipy library) (Fig. 9).

A neural network model was applied to this type of data (Fig. 10), which is based on the convolutional neural network (CNN) [17]. The architecture was selected experimentally. It is the input data that comes to the normalization layer (BatchNormalization) with the aim of uniform learning.

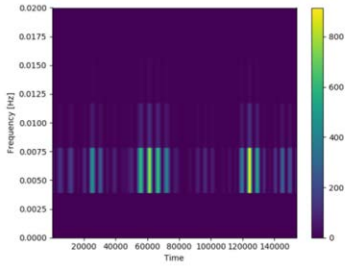


Fig. 9. Spectrogram of the EEG Sensor Signal.

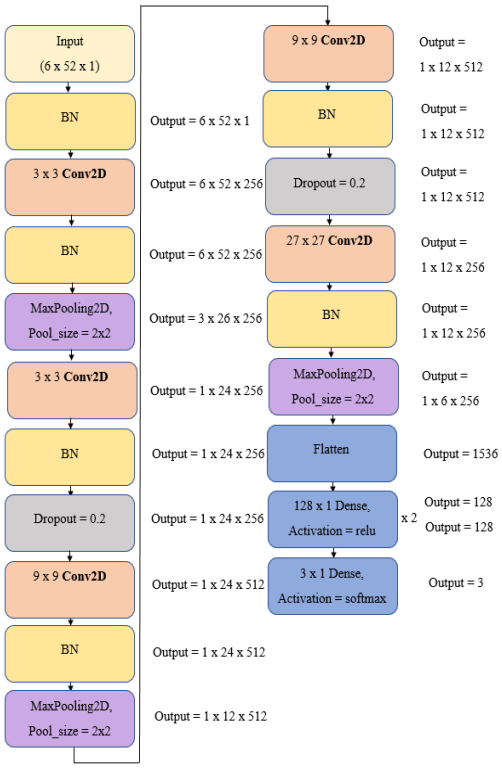


Fig. 10. Architecture of CNN.

BatchNormalization is a method for deep learning accelerating that solves a problem of learning efficiency. Normalization is implemented before every neural network layer [18]. Further, the convolutional neural network [19] receives normalized data at the input, and convolutional layers form 3x3 feature maps from it. During the experiment, it was revealed that a gradual twofold increase in the convolution core is two times more optimal for this architecture. With pooling (MaxPooling), the sample of the input space is being reduced by half (2x2), after that the "Dropout" layer is used to exclude a certain percentage of random neurons, since the neural network was overtrained during training. Then data is being converted to a one-dimensional vector using the Flatten layer. Classification is performed by the fully connected layer.

Based on the results obtained from experimental studies of the first model, several neural network models have been developed. The difference between the second and the third models (Fig. 11, 12) are that the data of the neural network model were presented in the form of a sequence, which were also converted using the fast Fourier transform.

2) RNN (LSTM) + CNN + RNN (LSTM) + Dense: A neural network model is shown on Fig. 11, the basis of which is a convolutional neural network (CNN) [20], that uses time convolutional layers (Conv1D). This layer creates a convolution core, which is convoluted with the input layer in one time dimension [21]. The architecture is as follows: in the second experiment, the input data comes to the recurrence layer (LSTM) with the maximum number of neurons, depending on the GPU capability, then the data goes to a time convolution layer in which a window of size 3 was specified empirically, after that the data is normalized by normalization layer (BatchNormalization). Based on the first experiment, the "Dropout" layer was applied, in which 20% of neurons are randomly turned off to exclude overtraining of the neural network. Then the data is transferred to the recurrent neural layer and converted into a one-dimensional vector using the "Flatten" layer. Classification is performed by the fully connected layer (Fig. 11).

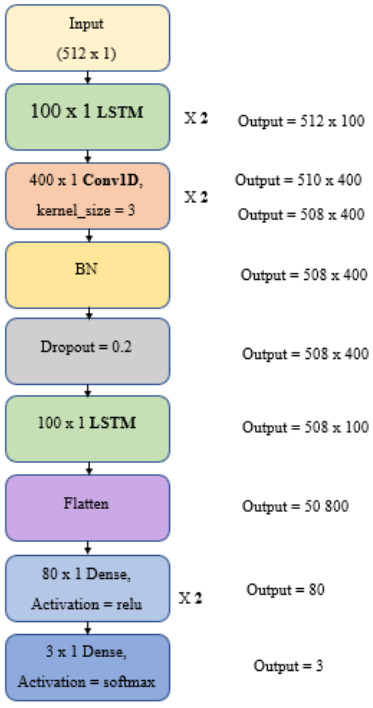


Fig. 11. Neural Network Architecture of the RNN (LSTM) + CNN + RNN (LSTM) + Dense Type using the Fourier Transform.

3) Batch_normalization + CNN + Dense: The difference between the architectures of the third model and the second model is that before the data is going to be transferred to the convolutional layer, it is being normalized. A normalization layer (BatchNormalization) was applied, before each time convolutional layer, then similarly, the data was converted into a one-dimensional vector, using the Flatten layer for fully connected layer and classification (Fig. 12).

4) LSTM + NN based on "U-net": fourth model was developed based on “Unet” (Fig. 13), using a time convolution, due to fact that database is small. The architecture of model 4 is shown in Fig. 14.

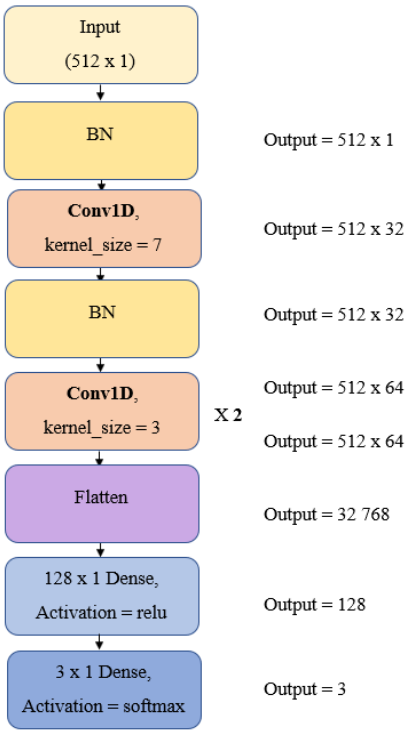


Fig. 12. Neural Network Architecture of Type Batch_Normalization + CNN + Dense using Fourier Transform.

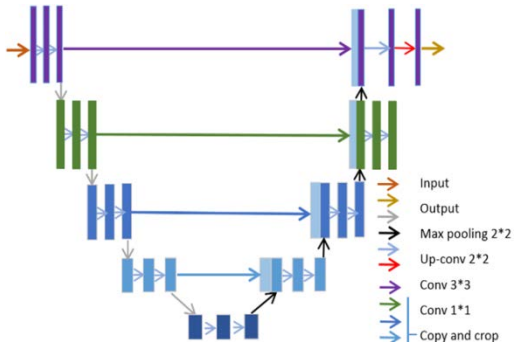


Fig. 13. The Architecture of the Neural Network “U-net".

For signal preprocessing, the fast Fourier transform method was used. Also, to improve the quality of training, a function was created (blink_augmentation), where several positions are generated for each "Blink” artifact, where this artifact will be recorded. The result is number of "Blink" samples with the same artifact but the different location in samples.

During the study, the U-net architecture was used (Fig. 13), which consists of an encoder (narrowing part), a bottleneck and a decoder (expanding part). This architecture is used for the analysis of R-grams, MRI and other medical images.

The first part of U-net is the classical architecture of a classification convolutional neural network [22]. It consists of repeated applications of two convolutional layers, with a 3–3 kernel, followed by the ReLU activation function and the MaxPooling operation, which reduces the input representation by the maximum value in the window (poolsize, in this case the value is 2).

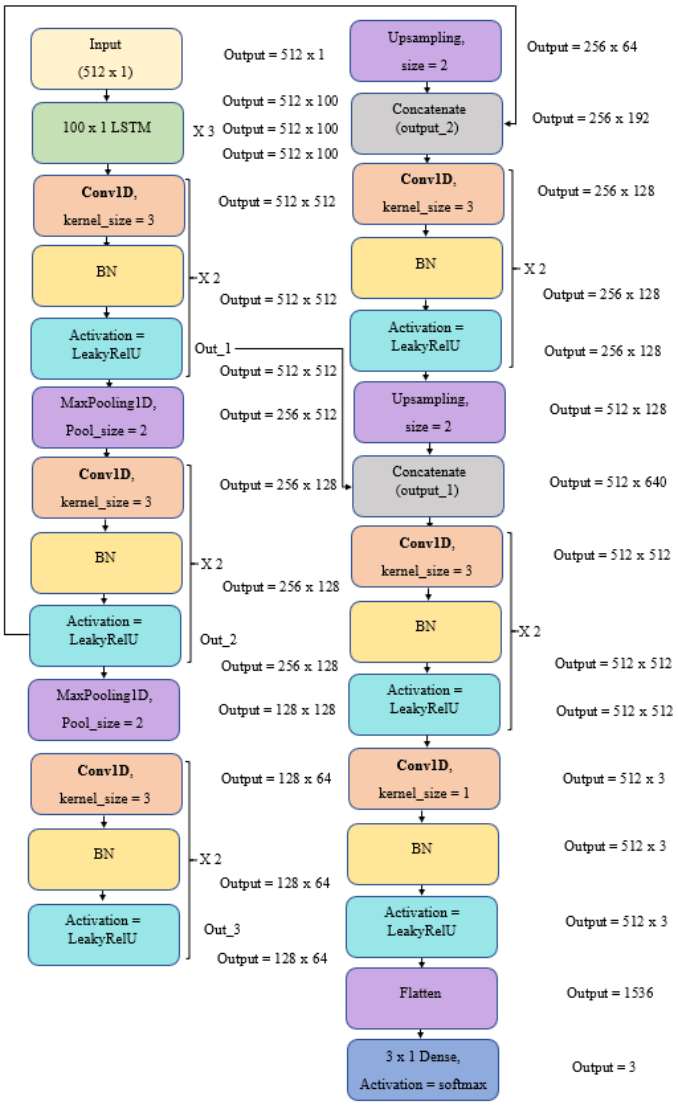


Fig. 14. The Architecture of the Neural Network Type LSTM + NN based on "U-net" using the Fourier Transform.

"Bottleneck" is a part of the network located between the contracting and expanding parts [23]. The second part consists of reverse convolution (deconvolution), which contains two convolutional layers with a 3–3 kernel and the Relu activation function, then concatenation is performed. At the last level,

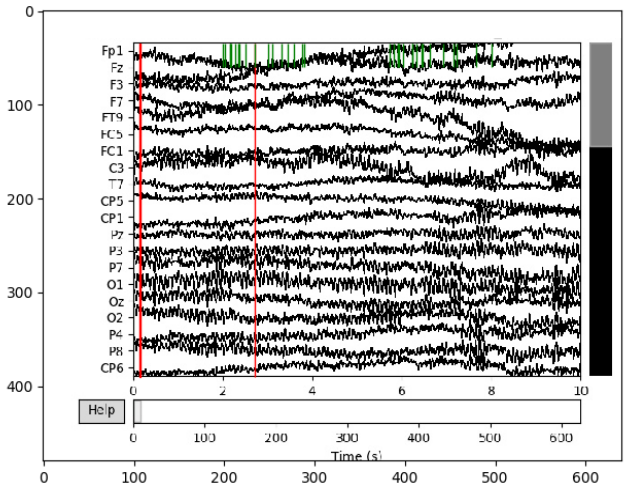
convolution 1x1 is used to match each vector with class attributes. Then the data is converted into a one-dimensional vector using the Flatten layer and the classification is performed by a fully connected layer [23].

In all experiments, the "fit" method was used for training. The number of samples, the gradient and the number of epochs for the model as well as compile method "Adam optimization function" and the error calculation function "categorical_crossentropy" were determined for each neural network model. It was revealed empirically that categorical_crossentropy is the most suitable error calculation function to optimize Adam parameters. To assess the quality of training, the Accuracy metric was chosen.

TABLE I. LEARNING OUTCOMES OF CLASSIFICATION MODELS WITH VARIOUS PARAMETERS

No.	Neural network architecture	Epochs	Batch_size	Accuracy	
				train	test
1	Batch_normalization + CNN + Dense with spectrogram	20	128	0.68	0.67
2	RNN (LSTM) + CNN + RNN (LSTM) + Dense	20	256	0.81	0.60
3	Batch_normalization + CNN + Dense	50	16	0.94	0.49
4	LSTM + NN based on U-Net model	10	300	0.70	0.70

A comparison of the results of an experimental study of four models is given in Table I. The training graph of neural network models was analyzed. The developed neural network based on the U-net architecture with recurrent layers demonstrates the best result of artifact recognition. 70% accuracy were acquired on test samples. Fig. 15 shows the results of automatic search for artifacts.



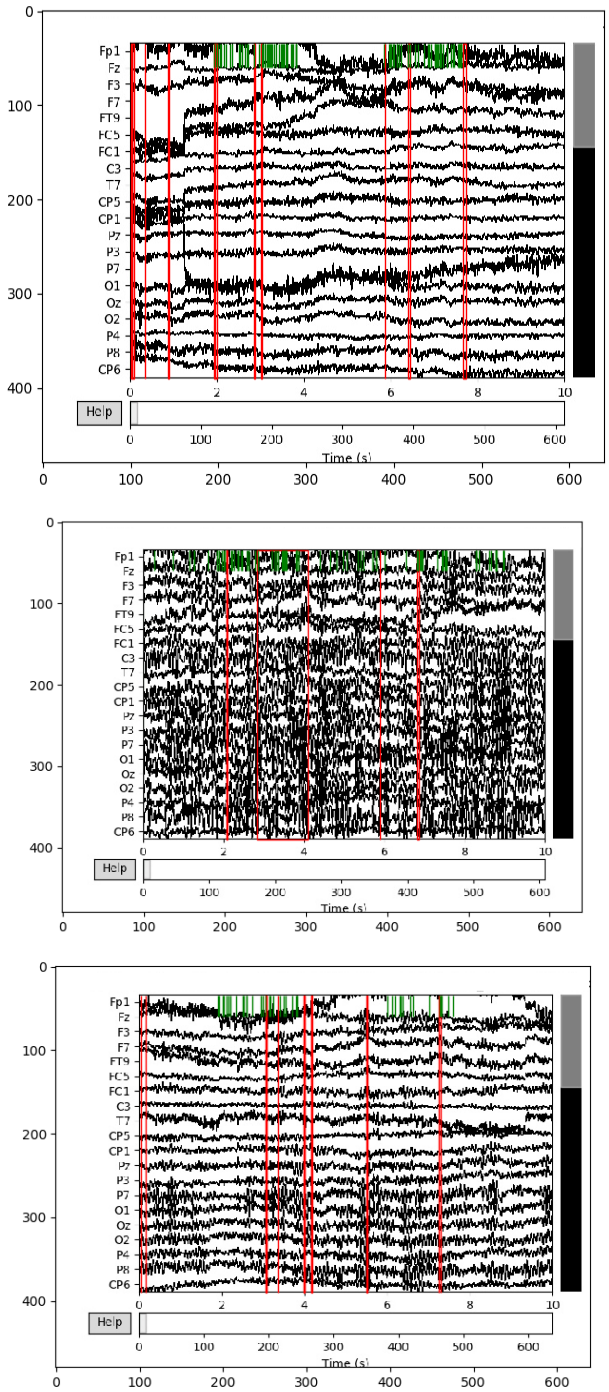


Fig. 15. The Result of Neural Network Activity. Examples of Highlighted Neural Network Artifacts: Blink Artifact is Green, Global Artifact Artifact is Red.

VI. CONCLUSION

A neural network model capable of recognizing artifacts in the process of recording EEG has been developed. Experimentally LSTM + U-net architecture was formed. To solve the problem, the U-net architecture, which is a two-dimensional convolution, was modified - a one-dimensional temporary convolution was used, the input of which received data from LSTM layers. Ensuring the required accuracy (70%)

is achieved due to the properties of the LSTM layers (trained to determine the signal state) and qualitative symmetric analysis (tension / compression) of the modified U-net layer.

Data analysis was carried out, in which the distance between artifacts in the signals, the maximum duration of each type of artifact was found. Using analytical functions, an optimal time window was allocated for artifact recognition, based on the maximum length of the "Blink" artifact. Since the data was manually filtered from the artifacts, and the database was small (9574 samples with artifacts), there was a problem with the quality of training of the neural network. The database was expanded using augmentation method, which partially influenced the learning process (28,722 samples with augmentation).

An analysis of existing architectures of neural networks, as well as an experiment with a training set was conducted. As a result, a neural network was developed based on recurrent neural network and U-net. The resulting neural network model is capable of detecting artifacts in the converted signal with an accuracy of 70%. The developed intelligent system can be used as an auxiliary tool for the analysis of the EEG signal.

During the study, the methods using libraries of applied software packages were developed for selecting an artificial neural network model of defects in digital signals, such as blinking artifacts in an EEG signal. Selected tools are able to create an environment for research and modeling of various signals.

The study showed the prospects for using the identified types of neural network architectures for analyzing EEG signals. The architectures selected during the study can be used in future studies aimed at modeling and clustering EEG signals.

REFERENCES

[1] J. N.Acharya, A. J. Hani, P. Thirumala, and T. N. Tsuchida, “American clinical neurophysiology society guideline 3: a proposal for standard montages to be used in clinical EEG,” The Neurodiagnostic Journal vol. 56, no. 4, pp. 253-260, 2016.

[2] A. S. Lundervold, A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” Zeitschrift für Medizinische Physik, vol. 29, no. 2, pp. 102–127, 2019.

[3] K. Nathan, and J. L. Contreras-Vidal, “Negligible motion artifacts in scalp electroencephalography (EEG) during treadmill walking,” Frontiers in Human Neuroscience, vol. 9, p. 708, 2016.

[4] B. Singh, and H. Wagatsuma, “A removal of eye movement and blink artifacts from EEG data using morphological component analysis,” Computational and Mathematical Methods in Medicine, vol. 2017, p. 1861645, 2017.

[5] N. Gebodh, Z. Esmaeilpour, D. Adair, …, and M. Bikson, “Inherent physiological artifacts in EEG during tDCS,” Neuroimage, vol. 185, pp. 408–424, 2019.

[6] I. Obeid, and J. Picone, “The temple university hospital EEG data corpus,” Frontiers in Neuroscience, vol. 10, p. 196, 2016.

[7] M. Golmohammadi, … and J. Picone, “Automatic analysis of EEGs using big data and hybrid deep learning architectures,” Frontiers in human neuroscience, vol. 13, p. 76, 2019.

[8] X. Jiang, G. B. Bian, and Z. Tian, “Removal of artifacts from EEG signals: a review,” Sensors, vol. 19, no. 5, p. 987, 2019.

[9] B.A. Zaikin, anf A.F. Kotov, “An estimation of efficiency of filtering algorithms of state vector of small-sized observed object with non-Markovian approximation of trajectory,” Russian Technological Journal, vol. 9, no. 4, pp. 38-48, 2021. https://doi.org/10.32362/2500-316X-2021-9-4-38-48

[10] F. Lotte, L. Bougrain, …, and F. Yger, “A review of classification algorithms for EEG-based brain–computer interfaces: a 10-year update,” Journal of Neural Engineering, vol. 15, no. 3, p. 031005, 2018.

[11] A. Kilicarslan, and J. L. C. Vidal, “Characterization and real-time removal of motion artifacts from EEG signals,” Journal of Neural Engineering, vol. 16, no. 5, p. 056027, 2019.

[12] Z. Tayeb, J. Fedjaev, …, and J. Conradt, “Validating deep neural networks for online decoding of motor imagery movements from EEG signals,” Sensors, vol. 19, no. 1, p. 210, 2019.

[13] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, “Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals,” Clinical Neurophysiology, vol. 130, no. 1, pp. 25–37, 2019.

[14] N. N. Astakhova, L. A. Demidova, and E. V. Nikulchev, “Forecasting method for grouped time series with the use of k-means algorithm,” Applied Mathematical Sciences, vol. 9, no. 97, pp. 4813-4830, 2015.

[15] W. Chen, Y. You, Y. Jiang, M. Li, and T. Zhang, “Ensemble deep learning for automated visual classification using EEG signals,” Pattern Recognition, vol. 102, p. 107147, 2020.

[16] C. Holdgraf, S. Appelhoff, and S. Bickel, “iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology,” Scientific Data, vol. 6, p. 102, 2019.

[17] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” In Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research (PMLR, vol. 48), pp. 507–516, 2016.

[18] K. Ramasubramanian, and A. Singh, “Deep learning using keras and tensorflow. In: Machine Learning Using R. Berkeley: Apress, pp. 667-688, 2019.

[19] R. T. Schirrmeister, J. T. Springenberg,. …, and T. Ball, “Deep learning with convolutional neural networks for EEG decoding and visualization,” Human Brain Mapping, vol. 38, no. 11, pp. 5391–5420, 2017.

[20] A. Saxena, “Convolutional neural networks: an illustration in TensorFlow. XRDS: Crossroads,” The ACM Magazine for Students, vol. 22, no. 4, pp. 56–58, 2016.

[21] M. Wang, J. Hu, and H. Abbass, “Stable EEG Biometrics Using Convolutional Neural Networks and Functional Connectivity,” Australian Journal of Intelligent Information Processing Systems, vol. 15, no. 3, pp. 19–26, 2019.

[22] G. Xu,; X. Shen, S. Chen, “A Deep Transfer Convolutional Neural Network Framework for EEG Signal Classification,” IEEE Access, vol. 7, pp. 112767–112776, 2019.

[23] Y. Weng, T. Zhou, Y. Li, and X. Qiu, “NAS-Unet: Neural architecture search for medical image segmentation,” IEEE Access, vol. 7, pp. 44247–44257, 2019.

Changing Communication Path to Maintain Connectivity of Mobile Robots in Multi-Robot System using Multistage Relay Networks

Ryo Odake, Kei Sawai
Graduate School of Engineering
Toyama Prefectural University, Toyama, Japan

***Abstract*—Mobile robots are being increasingly used to gather information from disaster sites and prevent further damage in disaster areas. Previous studies discussed a multi-robot system that uses a multistage relay backbone network to gather information in a closed space after a disaster. In this system, the mobile robot explores its search range by switching the connected nodes. Here it is necessary to maintain the communication quality required for the teleoperation of the mobile robot and to send and receive packets between the operator PC and the mobile robot. However, the mobile robot can become isolated when it is not able to maintain the communication quality required for teleoperations in the communication path after changing the nodes. This paper proposes a method to change the communication path of a mobile robot while maintaining its communication connectivity. In the proposed method, the mobile robot changes its route while maintaining communication connectivity without any communication loss time by connecting to two nodes.**

Keywords*—Multi-robot; multistage relay network; communication connectivity; changing communication path

I. Introduction

After a disaster occurs, disaster reduction activities are performed in the affected area to prevent the damage from spreading. In the implementation of disaster reduction activities, information needs to be gathered to determine the damage status [1–2]. Existing infrastructure, such as surveillance cameras, drone aerial photographs, and rescue teams, can be used to gather such information [3–6]. However, in some cases, the existing infrastructure cannot be used because of infrastructure malfunction or lack of power supply. Also, it is difficult for people to control the drone based on the camera images. Therefore, the use of drones is not effective in enclosed spaces after a disaster. Moreover, there is a risk of endangering human lives or inducing secondary disasters during information gathering by rescue teams. Therefore, the use of mobile robots is widely preferred for gathering information in enclosed spaces after disasters [7–11].

Two communication methods are adopted for mobile robots: wired and wireless. Wired communication helps maintain a stable communication quality and power supply to the mobile robot by using cables [12]. However, cables can get disconnected and communication with the mobile robots can be interrupted when cables become tangled with obstacles or the wheels of the mobile robot. Wireless communication has a high runnability because of the absence of physical restrictions using cables [13–14]. However, in wireless communication, mobile robots may become isolated when radio waves are hindered by obstacles. Therefore, in an enclosed space after a disaster, it is necessary to use the communication method that best matches the purpose and the situation of the disaster area [15–16]. This paper discusses a method for gathering information using wireless communication in environments where it is difficult to explore with a mobile robot using wired communication.

Robot wireless sensor networks (RWSNs) involve the teleoperation of mobile robots using wireless communication [17–19]. In an RWSN system, a mobile robot expands its search range by deploying a relay node called a sensor node (SN) in its path (Fig. 1). Therefore, the RWSN can gather information without depending on the existing infrastructure. In a network that uses multistage relaying, such as RWSN, the communication quality decreases as the number of relays and the distance between the nodes increases. Therefore, it is difficult to maintain communication connectivity of mobile robots in a multistage relay network; these networks are mainly operated by single robots. However, there is a limit to the range searched by a single robot in a large-scale facility. Therefore, this study discusses a multi-robot system that uses a multistage relay backbone network (Fig. 2).

In the proposed multi-robot system, a static multistage relay network is constructed by deploying SNs equipped with a single mobile robot first. Then, this system connects the mobile robot to the constructed network, and the mobile robot explores the search range by switching its connection between the nodes. Then, the mobile robot exhaustively explores within the network construction range while changing the search range by switching between the nodes. However, when changing the node to be connected, the operator experiences communication loss time with the mobile robot and cannot obtain environmental information from the mobile robot. Moreover, there is the risk of isolation when a mobile robot fails to connect to a node or is unable to reconnect. In addition, the operator cannot determine whether the communication quality required for the teleoperation of the mobile robot can be maintained in the communication path after changing the node to which the robot needs to connect. Therefore, there is a risk that the mobile robot becomes isolated without maintaining the communication quality.

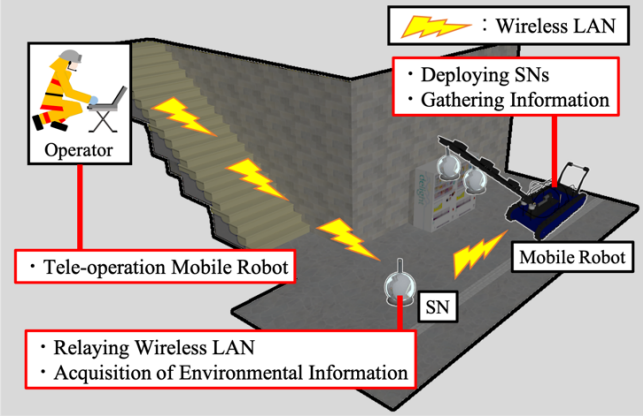


Fig. 1. Gathering Information by Mobile Robot (Robot Wireless Sensor Networks : RWSN).

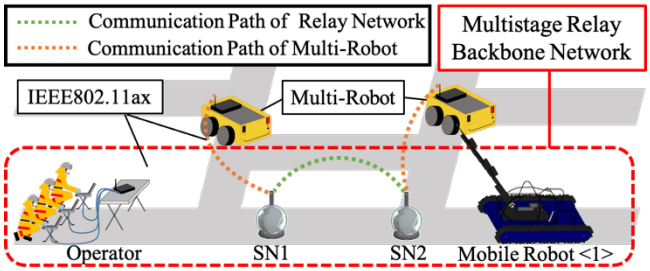


Fig. 2. Multi-Robot System using Multistage Relay Network.

To solve these problems, this paper proposes a method to change the communication path so that the communication connectivity of mobile robots is maintained. In the proposed method, the mobile robot teleoperates with two communication paths to eliminate any communication loss with the mobile robot. The communication quality required for the teleoperation of the mobile robot in the communication path is maintained after changing the nodes to be connected to; this method obtains the communication quality of the changed communication path before changing the communication path. The experiment in this paper confirmed that the required communication quality can be maintained in the communication path after changing the connected node using the proposed method. And the experiment showed the effectiveness of the proposed method.

II. Multi-robot System using Multistage Relay Network

A. Multi-robot Information Gathering

Many wireless teleoperation systems for mobile robots are based on the transmission control protocol/Internet protocol (TCP/IP). TCP/IP is highly compatible with mobile robot communication because most control systems of mobile robots use PCs. Therefore, socket communication is often adopted for mobile robot communication, and information communication is typically done by packet transmission and reception. Therefore, RWSN adopted the wireless LAN as the communication method. The operator receives packets containing information about the camera and sensor, which is acquired by the mobile robot, and this information is used to teleoperate the mobile robot (Fig. 3).

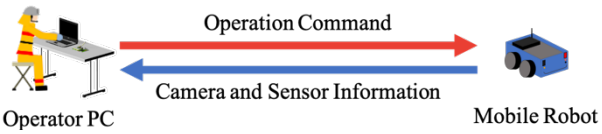


Fig. 3. Teleoperating Mobile Robot.

In previous research, there are systems that distribute and explore a large number of robots in the environment [20–21]. However, this system dynamically changes the route to send the information to the operator, which results in a misalignment of the reception intervals of the operation command packets. In addition, it is difficult to send large amounts of data, such as videos, using this system. These problems limit the operability and the ability of the mobile robot to gather information. Therefore, in this paper, the mobile robot is explored by connecting it to a static multistage relay network, such as the RWSN.

B. Flow and Requirements for Information Gathering by Multi-robot System using Multistage Relay Network

The process of the multi-robot operation using the multistage relay network is shown below (Fig. 4).

- Construction range of a multistage relay backbone network is expanded with the mobile robot <1> such as RWSN.
- Multiple mobile robots are connected to the constructed backbone network.
- Mobile robot <1> and the other mobile robots search within the network construction range while switching the nodes to be connected.

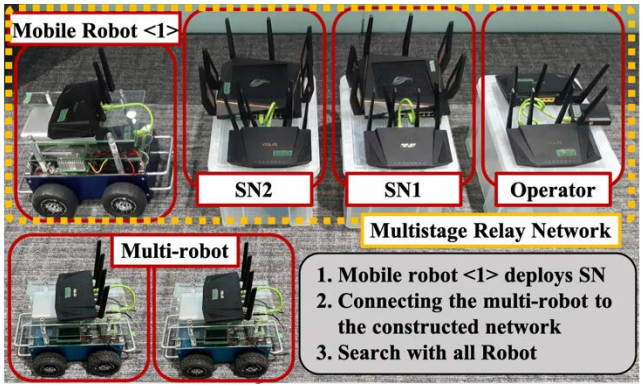


Fig. 4. Device of each Node that Constitutes Multi-robot System.

Based on the results of the operations of the robots at the damaged nuclear power plant, this study defined the network requirements of the multi-robot system as below.

- Teleoperation of a total of three or more mobile robots.
- Maintaining a throughput of 20 Mbps or higher in the communication path between each operator PC and each mobile robot.

The teleoperation of a multi-robot requires a throughput of 20.0 Mbps or higher in the communication path between each operator PC and each mobile robot. However, the theoretical value of IEEE802.11b/g used in RWSN is 54.0 Mbps, which is

insufficient throughput for multi-robot operations. This study constructs a network using IEEE802.11ax, which has a theoretical value of approximately 1,200 Mbps.

C. Topology of Multistage Relay Network

Based on the process flow and requirements of multi-robot systems, this paper proposed a multistage relay network topology for the construction of multi-robot environments. The network topology shown in Fig. 5 is characterized by the fact that each node in the multistage relay network has an access point (AP) and multiple mobile robots can be connected to a single node. Therefore, in this topology, each node constructs a network, and multiple mobile robots can explore the network construction range exhaustively while changing the APs to be connected.

D. Problems in Changing Communication Path

In this multi-robot system, the mobile robot changes the search range while switching the connected nodes, and exhaustively explores within the network construction range (Fig. 6). However, when the mobile robot changes the node to be connected, the below problems occur.

- There is communication loss between the operator PC and the mobile robot while changing the node to be connected to.
- There communication quality is unknown in the communication path after changing the node to be connected.

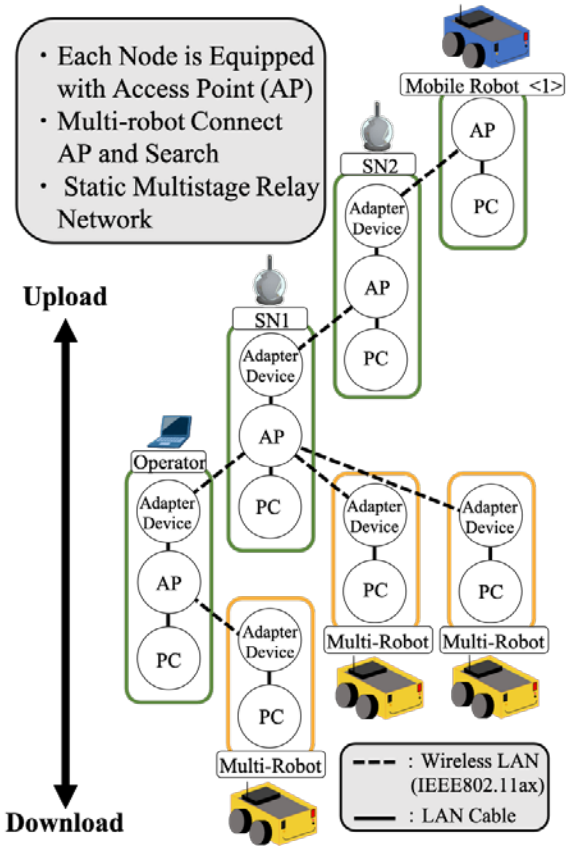


Fig. 5. Network Topology of Multi-robot System.

Fig. 6. Switching Nodes to Be Connected.

When the environment changes around the mobile robot during the communication loss time, the operator cannot receive sensor and camera information and cannot send operation command packets; consequently, it becomes difficult to respond to environmental changes. Moreover, when the connection change fails, the mobile robot becomes isolated. When the node to be connected is changed, the operator cannot determine whether the communication quality required for the teleoperation of the mobile robot can be maintained in the communication path after the node is changed. Therefore, there is a possibility of the mobile robot becoming isolated because of its inability to maintain communication connectivity after connecting to a new node. An isolated mobile robot is an obstacle to other mobile robots and affects their ability to explore the scene of disaster; they could also become a cause of secondary disasters because of their battery ignition. Therefore, this multi-robot system requires a method to change the node to which it is connected without losing the communication connectivity of the mobile robot. This paper proposes a communication path change method to maintain the communication connectivity of mobile robots.

III. Requirements for Changing Communication Path in Multi-robot System

Given the problems described in Section 2.2, the following specifications are required for the mobile robot to change the communication path in this multi-robot system.

1) The capability to send and receive packets between the operator PC and the mobile robot.

2) The capability to switch the communication path even when the operator does not know the exact location of the SN and the mobile robot.

3) Capable of obtaining the communication quality required for teleoperation of a mobile robot on the changed communication path, before switching the path.

During the exploration, it is possible to reduce the risk of isolation of the mobile robot by monitoring the changes in the environment and by being able to respond to them while the communication path was being changed. Therefore, as stated in specification (1), the system needs to be able to send and receive operation commands and camera/sensor information between the operator PC and the mobile robot during the communication path change.

After a disaster, a closed space would have an environment with multiple obstacles. In such cases, the SN and the mobile robot might not be able to see each other with a camera, or it might be difficult to measure the distance with a sensor because of the various disturbances. In such cases, there is a possibility of the mobile robot not being able to connect to the appropriate node and becoming isolated because of the lack of

location information. Therefore, as stated in specification (2), the communication path is changed without depending on the location information.

Some studies have used the electric field strength as a method for switching the routing [22]. However, the teleoperation of mobile robots sends and receives packets; therefore, it is necessary to maintain the communication quality at the packet level. Therefore, specification (3) requires the maintenance of the throughput required for the teleoperation of the mobile robot in the changed communication path before changing the node. As described in Section 2.2, the throughput required for the teleoperation of the mobile robot in this system is 20.0 Mbps; therefore, in this paper, the throughput to be maintained in the communication path after the change was set to 20.0 Mbps.

IV. Method for Changing Communication Path while Maintaining Communication Connectivity for Mobile Robots

This paper considers a method for changing the communication path of a mobile robot while maintaining communication connectivity in a multi-robot system using a multistage relay backbone network. This chapter proposes a method to obtain advance information about the communication quality of the changed communication path; this method will ensure that there is no communication loss time when the path is changed. In this method, apart from the main communication path, there is a sub-communication path that connects in advance to the next node. We also include a "judgment communication path" that monitors the communication quality. Section 4.1 describes a method to change the communication path without causing any communication loss; two communication paths were used here. Section 4.2 describes a method to determine in advance whether the communication quality required for the teleoperation of a mobile robot can be maintained in the communication path after changing the connecting nodes.

A. Communication Path Changing Method using Sub-communication Path

The proposed method used a mobile robot connected to two APs (IEEE802.11ax), as shown in Fig. 7. The steps for changing the communication path of a mobile robot using two communication paths are given below. Here, this paper assume that the node number is *n* (Fig. 8).

1) The mobile robots connect to the *n*th and *n+1*th nodes and teleoperate via the main communication path.

2) The communication quality of the judgment communication path is used to determine whether to switch between the main communication path and the sub-communication path.

3) The sub-communication path is changed to the main communication path.

4) The path that was originally the main communication path is changed to a sub-communication path. (The adapter device that was connected to the *n*th node is changed to the *n+2*th node).

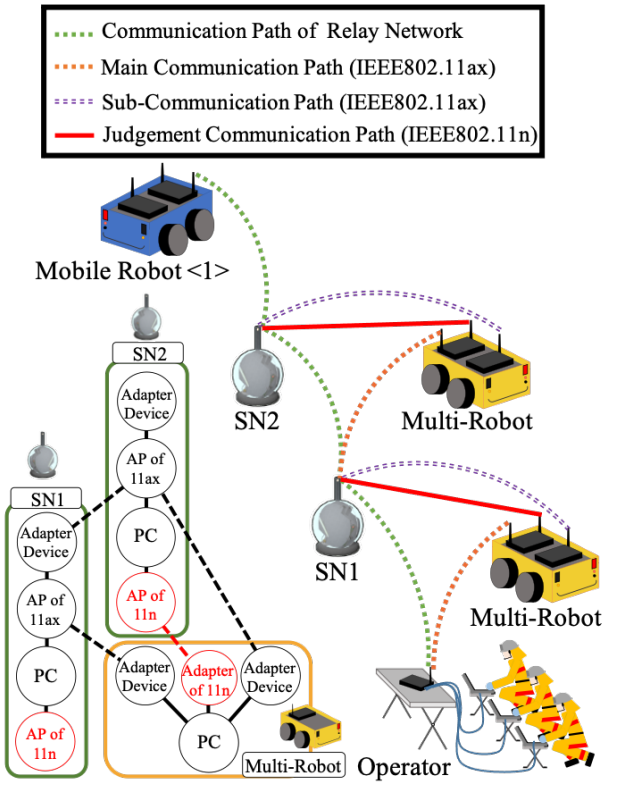


Fig. 7. Method for Changing Communication Path while Maintaining Communication Connectivity for Mobile Robots.

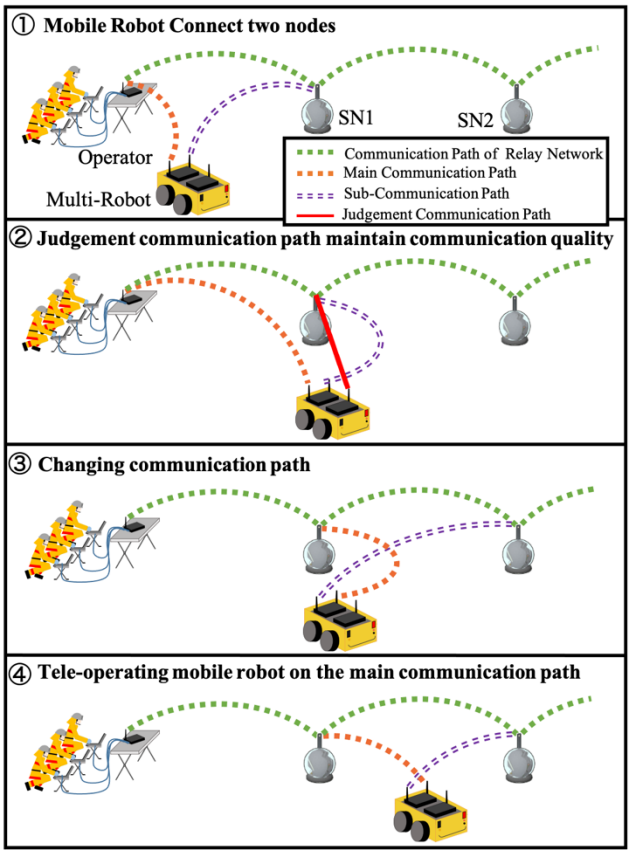


Fig. 8. Flow of Changing Communication Path.

Two communication paths were used: one with a large number of relays and the other with a small number of relays. This paper consider the decrease in communication quality due to the increase in the number of relays, and use the communication path with a small number of relays as the main communication path for sending and receiving teleoperation packets. Then, the sub-communication path becomes the next main communication path, as shown in Fig. 8. The packets are sent and received to confirm that the communication has not been interrupted. The operator can communicate with the mobile robot on the path via the *n*th node when changing the sub-communication path to the main communication path. Then, when changing the main communication path to the sub-communication path, it is possible to communicate with the mobile robot via the *n+1*th node. In this way, packets can be sent and received between the operator PC and the mobile robot while the communication route is being changed. Therefore, the mobile robots switch between the main communication path and the sub-communication path and can explore the disaster site without any communication loss even when there are changes in the communication path.

B. Monitoring Communication Quality using Judgment Communication Path

In the proposed method, in order to obtain in advance whether the communication path after changing the node to be connected can maintain the throughput required for teleoperation of the mobile robot, we construct a judgment communication path between the next node to be connected to and the mobile robot, as shown in Fig. 7. The judgment communication path is constructed by mounting an AP of IEEE802.11n, which is in the same frequency band as IEEE802.11ax, on the SN and an adapter device on the mobile robot. To obtain the throughput of this judgment communication path, the measurement packets are sent and received between the SN and the mobile robot <1>. This communication path is constructed in a separate network from the network of the multi-robot system, so that the throughput between the SN and the mobile robot can be measured without placing a load on the path used for teleoperation.

When the communication quality between an SN and a mobile robot of IEEE802.11n is maintained for the teleoperation of the mobile robot, the communication quality between an SN and a mobile robot of IEEE802.11ax can also be maintained for teleoperation of the mobile robot due to IEEE802.11n is in the same frequency band as IEEE802.11ax. This multi-robot system has also confirmed that the throughput between the operator PC and the SN is maintained required for the teleoperation of multiple mobile robots during the multistage relay network range is expanded. Therefore, if the communication quality required for the teleoperation of the mobile robot is maintained in the judgment communication path, it is assumed that the communication quality required for the teleoperation of the mobile robot is also maintained in the main communication path after switching the node to be connected. Therefore, in this method, the communication path is changed when the communication quality required for the teleoperation of the mobile robot can be maintained in the judgment communication path.

V. Evaluation of Communication Quality using Throughput Measurement

This chapter describes a method for measuring the communication quality characteristics when teleoperating a mobile robot. As described in Section 2, the teleoperation of a mobile robot sends and receives packets; therefore, it is necessary to evaluate the communication quality at the packet level. This paper evaluates the transmission speed in a TCP/IP-compliant communication path as the throughput at the packet level. The throughput [bps] specifies the transmission received per second by the PC. Therefore, the following experiments continue to send measurement packets from the mobile robot at a transmission speed of 25.0 Mbps and confirm whether a throughput of 20.0 Mbps or higher can be maintained at the receiving side. Therefore, the experiments in next chapter send measurement packets from the mobile robot side and confirm that the receiving side can maintain the throughput required for the teleoperation of the mobile robot.

VI. Changing Communication Path while Maintaining Communication Connectivity using Proposed Method

This experiment was conducted to confirm whether the proposed method can be used to change the communication path of a mobile robot while maintaining the communication connectivity. The experiment setting is shown in Fig. 9, and the equipment used is shown in Fig. 10. This experiment used Raspberry Pi 4 Model B as the PC, ASUS RT-AX3000 as the adapter device, GT-AX11000 as the AP of IEEE802.11ax, WN-AC433UA as the antenna of IEEE802.11n, and CAT8 LAN. The experiment was conducted in the following steps:

1) Deploying nodes equipped with AP of IEEE802.11n so that the distance between each node is 90 m. (Constructing a 270m network with 3hop).

2) Move the mobile robot and measure the throughput in the judgment path of IEEE802.11n at 10 m intervals (Between each SN and the mobile robot).

3) Move the mobile robot and measure the throughput in the communication path of IEEE802.11ax at 10 m intervals. (Between operator PC and mobile robot via each node).

This experiment selected 10 throughput values measured at each location and shows the average of them. In the multistage relay network constructed in this experiment, the nodes are deployed so that the distance between each node is 90 m (270 m in total), and the network is capable of maintaining a throughput of more than 60.0 Mbps between the operator PC and the mobile robot <1>. Fig. 11 shows the results of measuring the throughput on the judgment communication path (between SN and mobile robot). Fig. 11 shows that the throughput between SN1 and the mobile robot is more than 20.0 Mbps in the range of 60-130 m. Therefore, the proposed method changes the sub-communication path to the main communication path in that section. The throughput between SN2 and the mobile robot is more than 20.0 Mbps in the range of 150-210 m. Therefore, the proposed method changes the sub-communication path to the main communication path in that section.

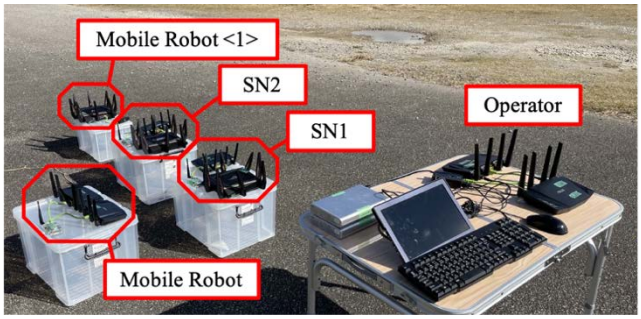


Fig. 9. Experimental Environment.

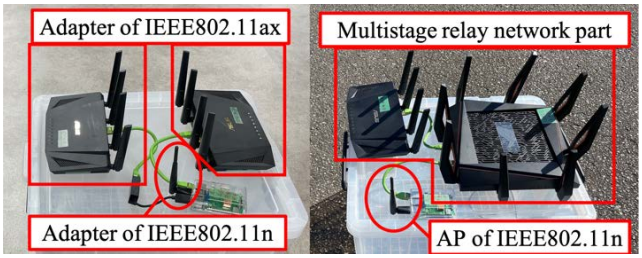


Fig. 10. Device of IEEE802.11n and IEEE802.11ax. (Left : SN, Right : Mobile Robot).

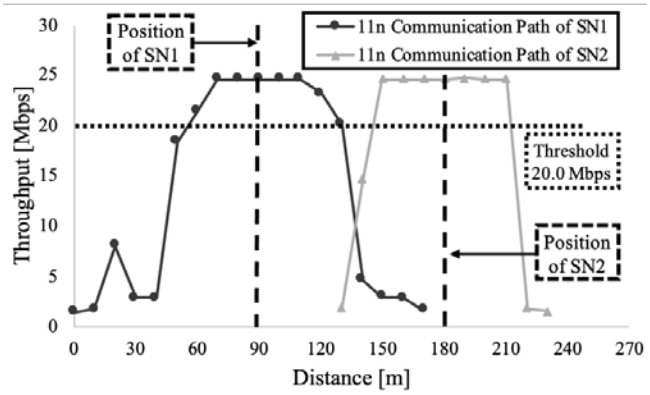


Fig. 11. Throughput of Judgment Communication Path (IEEE802.11n).

From the results shown in Fig. 11, we measured the throughput between the operator PC and the mobile robot so that the 60-130 m and 150-210 m sections overlap in this experiment, and confirmed whether the main communication path can be switched while maintaining the communication connectivity. The throughput in the communication path of IEEE802.11ax between the operator PC and the mobile robot via each node is shown in Fig. 12. As shown in Fig. 12, both communication paths have a throughput of more than 20 Mbps in the 60-130 m section, so it was possible to switch the main communication path while maintaining the communication connectivity. Also, in the 150-210 m section, both communication paths are maintained at more than 20 Mbps, so it was possible to switch the main communication path while maintaining the communication connectivity. As a result, when the proposed method was used to teleoperate a mobile robot, the communication path could be switched while maintaining the communication connectivity. Therefore, this method is effective for changing the connected nodes in a multi-robot system using a multistage relay backbone network.

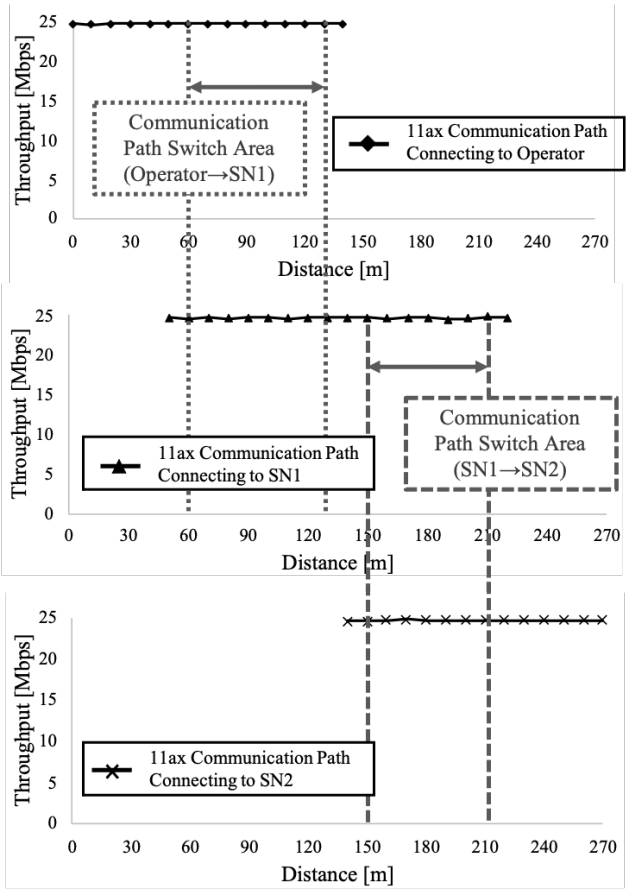


Fig. 12. Throughput between Operator and Mobile Robot. (Top : Mobile Robot Connecting to Operator, Middle： Mobile Robot Connecting to SN1, Bottom： Mobile Robot Connecting to SN2).

VII. Discussion

In the judgment communication path of this experiment, the throughput between SN1 and the mobile robot is more than 20 Mbps in the 60-130 m section, and the throughput between SN2 and the mobile robot is more than 20 Mbps in the 150-210 m section. The experimental results show that the IEEE802.11ax communication path can maintain a throughput of more than 20 Mbps in those sections, so it is possible to change the main communication path after confirming the communication quality in the judgment communication path in advance. Therefore, proposal method can obtain the communication quality of the changed communication path in advance, without communication loss time when the path is changed. Additionally, switching between the sub-communication path and the main communication path does not depend on distance information, but on the throughput of the judgment communication path, which is effective even in environments with many obstacles.

VIII. Conclusion

In a multi-robot system using a multistage relay network, multiple mobile robots were connected to the constructed network, and the mobile robots changed the search range by switching between the nodes. However, a mobile robot can be isolated when changing the node because of the loss of

communication or the poor quality of the communication in the new communication path. This paper proposed a method to change the communication path of a mobile robot using multiple paths so that communication connectivity is maintained.

The proposed method uses a sub-communication path, which connects in advance to the next node so that the communication path can be switched without any communication loss between the operator PC and the mobile robot. In addition, the proposed method constructs a judgment communication path between the nodes that constitute the sub-communication path; the main communication path and the sub-communication path are changed when the communication quality is maintained on the judgment communication path. Therefore, even when the sub-communication path is changed to the main communication path, the communication quality required for the teleoperation of the mobile robot can be maintained in the new main communication path. This study measured the throughput of the judgment communication path connected to each node and confirmed the section that can maintain the communication quality required for the teleoperation of the mobile robot. This experiment confirmed that the communication quality required for the teleoperation of the mobile robot could be maintained in the communication path of IEEE802.11ax in that section. The experiment results confirmed that it was possible to maintain the communication quality required for the teleoperation of the mobile robot in the new main communication path even when the main communication path was switched. The communication quality required for the teleoperation of the mobile robot was maintained in the judgment communication path. Therefore, the judgment communication path could monitor the communication quality of the changed path before changing the main communication path; this experiment proves the effectiveness of the proposed method.

In these experiments, the nodes to be connected to the mobile robot were changed manually. In the future, we will create a program that will automatically switch the communication path.

REFERENCES

[1] Yoshiaki Kawata, “The great Hanshin-Awaji earthquake disaster, damage, social response, and recovery,” Journal of Natural Disaster Science, Vol. 17, No. 2, pp.1-12, 1995.

[2] L. Ernesto Dominguez-rios, Tomoko Izumi, Yoshio Nakatani, “A disaster management platform based on social network system oriented to the communities self-relief,” IAENG International Journal of Computer Science, Vol. 42, No.1, pp.8-16, February 2015.

[3] Sabarish Chakkath, “Mobile robot in coal mine disaster surveillance,” IOSR Jourbal of Engineering, Vol. 2, No. 10, pp. 77-82, 2012.

[4] Keiji Sakuradani, Keigo Koizumi, Kazuhiro Oda, Satoshi Tayama,” Development of a sloap disaster monitoring system for expressway operation and maintenance control,” Jounal of GeoEngineering, Vol. 13, No.4, pp.189-195, December 2018.

[5] F. Kurz, D. Rosenbaum, J. Leitloff, O. Meynberg, P. Reinartz, “A real time camera system for disaster nad traffic monitoring,”. https://core.ac.uk/download/pdf/11146229.pdf.

[6] Jingxuan Sun, Boyang Li, Yifan Jiang, Chih-yung Wen, “A camera-based target detection and positioning UAV system for search and rescue (SAR) Purposes,” Sensors 2016, Vol. 16, No. 11, 1778. https://doi.org/10.3390/s16111778.

[7] Masataka Fuchida, Shota Chikushi, Alessandro Moro, Atsushi Yamashita, Hajime Asama, “Arbitrary viewpoint visualization for teleoperation of disaster response robots,” Journal of Advanced Simulation in Science and Engineering, Vol. 6, No. 1, pp.249-259, 2019.

[8] Hemanth Reddy A, Balla Kalyan, Ch. S. N. Murthy, “Mine Rescue Robot System – A Review,” Procedia Earth and Planetary Science, Vol.11, pp. 457-462, 2015.

[9] Trupti B. Bhondve, Prof.R.Satyanarayan, Prof. Moresh Mukhedkar, “Mobile rescue robot for human body detection in rescue operation of disaster,” International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, No.6, pp.9876-9882, June 2014.

[10] Zia Uddin, Mojaharul Islam, “Search and rescue system for alive human detection by semi-autonomous mobile rescue robot,” International Conference on Innovations in Science, Engineering and Technology, Octorber 2016.

[11] Xuewen Rong, Rui Song, Xianming Song, Yibin Li, “Mechanism and explosion-proof design for a coal mine detection robot,” Procedia Engineering, Vol. 15, pp.100-104, 2011.

[12] Tomoaki Yoshida, Keiji Nagatani, Satoshi Tadokoro, Takeshi Nishimura, Eiji Koyanagi, “ Improvements to the rescue robot Quince toward future indoor surveillance missions in the Fukushima Daiichi Nuclear Power Plant,” Field and Service Robotics, pp. 19-32, December 2013.

[13] Albert Ko, Henry Y. K. La, “Robot assisted emergency search and rescue system with a wireless sensor network,” International Journal of Advanced Science and Technology, Vol. 3, pp.69-78, Febuary 2009.

[14] Andrew Wichmann, Burcu Demirelli Okkalioglu, Turgay Korkmaz, “The integration of mobile (tele) robotics and wireless sensor networks: A survey,” Computer Communications, Vol. 51, No.15, pp. 21-35, September 2014.

[15] Yasushi Hada, Osamu Takizawa, “Development of communication technology for search and rescue robots,” Journal of the National Institute of Information and Communications Technology, Vol. 58, pp. 131-151, 2011.

[16] Carlos Marques, Joao Cristovao and Paulo Alvito, “A search and rescue robot with tele-operated tether docking system,” Industrial Robot: An International Journal, Vol. 34, No. 4, pp. 332-338, 2007.

[17] Yuta Koike, Kei Sawai, Tsuyoshi Suzuki, “A study of routing path decision method using mobile robot based on distance between sensor nodes,” International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 3, pp. 25-31, 2014.

[18] Kei Sawai, Ju Peng, Tsuyoshi Suzuki, “Throughput Measurement Method Using Command Packets for Mobile Robot Teleoperation Via a Wireless Sensor Network,” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, pp 348- 354, 2016.

[19] Tsuyoshi Suzuki, Ryuji Sugizaki, Kuniaki Kawabata, Yasushi Hada, Yoshito Tobe, “Autonomous deployment and restoration of sensor network using mobile Robots,” International Journal of Advanced Robotic Systems, Vol. 7, No. 2, pp. 105-114, 2010.

[20] M. Brett McMickell, Bill Goodwine, Luis Antonio Montestruque, “MICAbot: A robotic platform for large-scale distributed robotics,” 2003 IEEE International Conference on Robotics and Automation, pp. 14-19, 2003.

[21] Peng Zeng, Jiahong He, Bingtuan Gao, “Reliable robot-flock-based monitoring system design via a mobile wireless sensor network,” Sensor-Cloud Systems and Applications, Vol. 9, pp.47125-47135, 2021.

[22] Masayuki Tauchi, Tetsuo Ideguchi, Takashi Okuda, “Ad-hoc Routing Protocol Avoiding Route Breaks Based on AODV,” Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.

A Conceptual Design Framework based on TRIZ Scientific Effects and Patent Mining

E-Ming Chan[1]
Department of Mechanical, Materials and Manufacturing Engineering
University of Nottingham Malaysia, 43500 Semenyih, Selangor, Malaysia

Ah-Lian Kor[2]
School of Built Environment, Engineering and Computing
Leeds Beckett University, Leeds, United Kingdom

Mei Choo Ang[4]*
Institute of IR4.0, Universiti Kebangsaan Malaysia
43600 UKM, Bangi, Selangor, Malaysia

Kok Weng Ng[3]
Department of Mechanical, Materials and Manufacturing Engineering, University of Nottingham Malaysia, 43500 Semenyih, Selangor, Malaysia

Amelia Natasya Abdul Wahab[5]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM, Bangi, Selangor, Malaysia

***Abstract*—Conceptual design represents a critical initial design stage that involves both technical and creative thinking to develop and derive concept solutions to meet design requirements. TRIZ Scientific Effects (TRIZSE) is one of the TRIZ tools that utilize a database on functional, transformation, parameterization of scientific effects to provide conceptual solutions to engineering and design problems. Although TRIZSE has been introduced to help engineers solve design problems in the conceptual design phase, the current TRIZSE database presents general scientific concept solutions with a few examples of solutions from patents which are very abstract and not updated since its introduction. This research work explores the derivation of a novel framework that integrates TRIZ scientific effects to the current patent information (USPTO) using data mining techniques to develop a better design support tool to assist engineers in deriving innovative design concept solutions. This novel framework will provide better, updated, relevant and specific examples of conceptual design ideas from patents to engineers. The research used Python as the base programming platform to develop a conceptual design software prototype based on this new framework where both the TRIZSE Database and Patents Database (USPTO) are searched and processed in order to build a Doc2Vec similarity model. A case study on the corrosion of copper pipelines by seawater is presented to validate this novel framework and results of the novel TRIZSE Database and patents examples are presented and further discussed in this paper. The results of the case study indicated that the Doc2Vec model is able to perform its intended similarity queries. The patent examples from results of the case study warrant further consideration in conceptual design activities.**

Keywords*—TRIZ; patent mining; natural language processing; product design

I. Introduction

TRIZ is a Russian Acronym which translate to the "Theory of Inventive Problem Solving" which was first initiated by Genrich Altshuller in 1946 [1]. TRIZ represents an engineering theory consisting of many problem-solving tools derived from compiling and analyzing utility patents to find general solutions and trends, which will be used to solve complex engineering problems that can be further decomposed into their individual problems and components.

One of the main TRIZ tools is the contradiction matrix which is well-established and is used to improve an engineering system based on solving engineering contradictions. According to TRIZ, there will always be engineering contradictions that need to be resolved in all inventive problems [2]. Contradiction matrix will recommend inventive principles derived from years of study of patent information to assist engineers resolve these contradictions. However, these inventive principles are very general and abstract for most engineers and interpreting these inventive principles in the context of specific solution concepts are challenging because expertise knowledge and experience are required. The other TRIZ tools including TRIZ scientific effects (TRIZSE) and system of standard inventive solutions also recommend solution concepts which are very general and abstract [1]. Therefore, research have explored into extending and improving existing TRIZ tools and proposed new patent extraction system such as Inventive Design Method (IDM) and extracting design information from patent documents and focus on assisting engineers to overcome their psychological barrier [3]. The proposed IDM has its limitations and deficiencies in issues related to duplicated information. This research work acknowledges the importance and criticality of defining design problems described in the form of function, parameter, transform and a real-time link to patent documents in patent offices to provide accurate relevant examples to assist engineers.

Therefore, this research work will explore the derivation of a framework to link TRIZSE to patent documents to support engineers in conceptual design and investigate the potential of enhancing TRIZSE with patents database. By developing this

*Corresponding Author.

link, TRIZSE may provide a better and clearer solution to each of their effects. The main objectives of this paper are:

1) To introduce the preliminary conceptual design system using TRIZSE which is enhanced by utilizing a similarity model built on a patent database to provide sufficient examples suggested by TRIZSE.

2) To build a Graphical User Interface (GUI) to allow a user to provide their input and queries for a single functional statement.

3) Conduct a case study to assess the performance of the program in factors of user input, data accuracy and its relevance.

The rest of the paper will cover the extraction of information, development of Doc2Vec model, and a case study. The proposed process model will be presented in the methodology section. This is followed by validation of the Doc2Vec model for feature extraction from the patent database for a use case. The report will then be concluded with recommendation for future improvement.

II. Background

TRIZ Scientific Effects (TRIZSE) is one of the TRIZ tools which is applied in the conceptual design phase that utilizes a database based on three types of search query; Functional, Parameterization, and Transformation to recommend a list of solution concepts (for e.g., if a user desires to constrain gas, one suggestion effect from the Functional aspect is Glassy Carbon which has high temperature resistance, extreme resistance to chemical attack, and impermeability to gases and liquid [4]). However, these suggestions lack suitable examples to support the recommended list of solution concepts and fail to inspire conceptual design ideas. Nonetheless, these deficiencies of the TRIZSE database offer a unique opportunity to explore the possibility of applying appropriate data mining techniques to search for examples related to the recommended list of solution concepts based on available online patents information with the intention of retrieving relevant patents as examples for any given recommended solution concepts by TRIZSE. To link the recommended solution concepts of TRIZSE to patent information, it is necessary to develop a suitable framework (with a user-friendly user interface) that integrates data analysis and mining technique (also known as patent mining) based on TRIZSE.

Using similar initiative to enhance a TRIZ technique, TRIZ Scientific Effects (TRIZSE) are comprehensive descriptions of effects based on sciences which provide good conceptual solutions which can potentially link patent information to obtain sufficient relevant real examples from TRIZSE database. In conceptual design activities, TRIZSE plays a role of redefining each function in a system and provides a set of suggestions of functional solutions as shown in an online database developed by Martin [4]. TRIZSE itself serves as a basis of creative idea generation in solving an engineering problem. Access to TRIZSE databases is limited due to subscriptions or memberships [5, 6], and third party state-of-the-art patent mining and analysis software [7]. However, the TRIZSE database provided by Martin [4] is available for public usage.

For the conceptual design framework development, the user must first define the engineering problem to be solved. The user-defined problem should be decomposed from complex problems into simple sentence structure before being used as an input. This not only simplifies the problem but also provides a clearer view of expected results from the query. There are vast selections of problem analysis tools such as Causal Loop diagram [8] that identify the polarity influences, Functional Decomposition and Morphology [9] decompose complex systems and re-merges as well as Systems Dynamics [10] that can simulate potential design changes. However, a simple and direct tool should be selected as it can provide more direct input for the model to manipulate. Hence, Functional Analysis (FA) [11] which maps component relationships can be selected for deployment. FA fosters problem statement formulation that could be further decomposed into a simple solution statement consisting of Subject-Action-Object. This facilitates better transition to the parameters of the TRIZSE database.

The linkage between TRIZSE to a patent database can be implemented through text mining techniques. A text mining Python library called Gensim is widely popular amongst patent mining research and has diverse range of data mining models. One of the models is a similarity model Doc2Vec inspired by vector representations of words using neural networks known as Word2Vec [12]. The Word2Vec model has a disadvantage as it does not consider the word order of each text document [12]. It functions similarly to a bag-of-words model where only the frequency of each word in the patent document is accounted for [12]. The proposed Doc2Vec is an improved model that can assess the word and neighboring words in the sentence to properly gauge the semantics of the entire sentence. Doc2Vec is compared to other similarity models such as Term frequency-inverse document frequency (TF-IDF) which calculates the importance of a word based on the frequency in the patent document and Global Vectors for Word Representation (GloVe) just like TF-IDF but with word embeddings to discover word contexts [13]. The result of comparisons shows that Doc2Vec is superior and capable of handling higher amounts of data entries (in millions) [13].

The patents database is sourced online. There are major databases such as the United States Patents and Trademark Office (USPTO), the Japanese Patent Office (JPO), the European Patent Office (EPO) [14]. For the ease of database extraction, the USPTO database is selected.

III. Methodology

A. Data Extraction

The TRIZSE database is extracted from the online database [4]. A third party free-to-use software called ParseHub is shown in Fig. 1. It performs as an automated user-action and data collector. Parsehub is utilised to extract every combination of the database and its results. The data extracted is stored into a Comma-Separated Values (csv) file.

Patent documents are unstructured texts with sections of irregular word limits and formatting. Thus, this paper makes use of one of the consistent sections of a patent document, the abstract. Sections such as claims would range from few sentences to several pages.

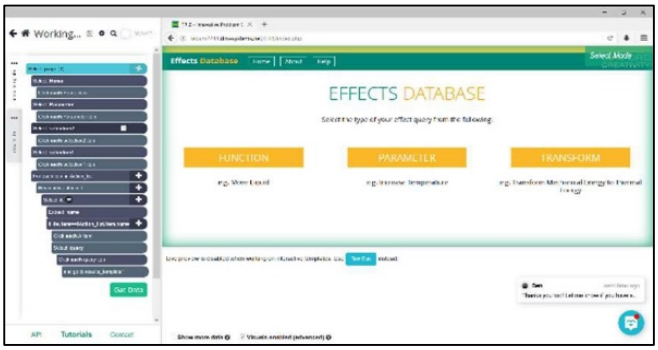


Fig. 1. Brief Interface of Parsehub.

B. Data Processing

The main data processing involves handling of millions of patents in the database. In USPTO, there are multiple types of patents such as utility, design, plant. Utility patents are patents that contain a feature or solution whereas design patents are typically drawings. Only utility patents are added value documents for this study. The total number of utility patents from the year 1971 to 2020 numbers is approximately 6.8 million. The abstracts of patents are useful for feature extraction in patent mining as there is a word limit and crucial information is present. Hence, each patent document's abstract undergoes a pre-processing stage, where words will be separated, and reformatted. The pre-processing steps involve removal of special characters, lowercasing capitals, and removal of redundant words [15, 16]. The pre-processing technique is adapted from the Gensim code examples [17]. The patent documents are tagged by simple indexing for accessing the documents in later stages.

C. Similarity Model Technique and Concept

Similarity models encompass supervised and unsupervised techniques, each with their own advantages and disadvantages. Supervised techniques involve high percentage of manual sorting and tagging that yields better overall results. On the other hand, unsupervised techniques do not have manual inspection. However, it is preferable over supervised techniques when dealing with massive number of documents. The Doc2Vec model is an unsupervised technique model. The data is fed into the algorithm and all unique words will be extracted from the patent documents into a vocabulary dictionary. The Doc2Vec model starts training with a vocabulary size and input texts. The Doc2Vec modelling can be simplified (as shown in Fig. 2).

```
Doc2Vec(abstracts, min_count=2,
window=3, vector_size=100,
epochs=20, workers=7)
```

Fig. 2. Modelling of Doc2Vec.

The description of the arguments for Doc2Vec is as follows: abstracts represent the input of patent abstracts; min_count the removal of words with frequencies less that the value f (in this example, 2); window is N (3 in this example), number of neighboring words that can affect the word of interest; vector_size shows the number of numerical elements that an abstract vector can attain. The higher the number of vector size, the higher the number of features of an abstract that can be captured; epochs are the number of training iterations of the model; workers are the number of computer cores for processing. With the vector generated, it can be compared numerically with a cosine similarity approach [14][17].

D. Process Chart of the Conceptual Design Framework

Once trained, the model can produce and predict a vector for new inputs to reflect similarity results. This can be used to evaluate similarity of a TRIZ Scientific Effect concept with real patent examples. The development of a Graphical User Interface (GUI) using PySimpleGUI library is to facilitate user input. The activities are depicted in Fig. 3. The processes and output is visible to the user through Graphical User Interface. The user inputs their problem in a Functional statement which will be a guide to invoke suitable TRIZSE settings. TRIZSE database will return results from the inputs which are fed into the similarity model individually by each effect. The model results are compiled in a descending order of patent similarity.

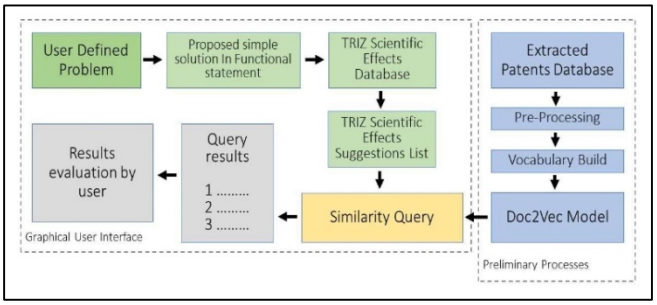


Fig. 3. Process Chart of the Conceptual Design System Framework.

IV. Results and Discussion

A. TRIZSE Database Extraction

This work starts with extraction and analysis of the TRIZSE database. The database extracted from the source website takes approximately 15 hours to parse through all combinations. The CSV file saved is checked and validated for accuracy. Fig. 4 shows the sample of the data extracted from the database. Each combination of TRIZ Scientific Effect yields a set of results which are suggestions of Effects along with brief description of the concept.

Absorb	Divided Solid	Effect	17 suggest	Amphiphiles	A chemical compound poss
Absorb	Divided Solid	Effect	17 suggest	Bingham Plastic	A viscoplastic material that
Absorb	Divided Solid	Effect	17 suggest	Diffusion	The movement of particles
Absorb	Divided Solid	Effect	17 suggest	Electret	An Electret is a dielectric m
Absorb	Divided Solid	Effect	17 suggest	Fractal Forms	A fractal is generally 'a roug
Absorb	Divided Solid	Effect	17 suggest	Gel	A gel is a solid, jelly-like ma
Absorb	Divided Solid	Effect	17 suggest	Ostwald Ripening	An observed phenomenon
Absorb	Divided Solid	Effect	17 suggest	Oxidation	A chemical reaction that in

Fig. 4. A Snapshot of Extracted TRIZSE Database.

The data extracted undergoes data elimination of duplicated data. For example, a combination of Absorb and Divided Solid in Function section can have three types of results which are

Framework, Effect, or Both. The results that are repetitive can be merged. For example, the Functional section containing 21,175 rows of effect suggestions is merged into 10,589 rows, which is 50% smaller. Similar process is applied to Parameter and Transformation section which yields similar reduction.

The Effects in TRIZSE database are obtained through the various combinations of inputs. Each Effects accommodates multiple combinations, and the total of Effects is approximately 1000. The concept description of each Effects will be assessed by the Doc2Vec model to identify similar concepts in patents.

B. USPTO Patents Extraction

The USPTO patents are downloaded from PatentsView website in a compressed ZIP file. A Python code parses through the ZIP file to extract the desired information which comprises the following: utility patents' number, title, and abstract. The process has taken 5 hours for 7.5 million patent documents and 6.8 million utility patents are extracted and stored.

C. Doc2Vec Model Code Development

The Doc2Vec modelling follows the stages of Preliminary processes in Fig. 3. The first stage is the pre-processing of each patent abstracts. In this stage, abstracts will undergo tokenization. Tokenization splits the abstracts, a paragraph of words, into individual words while keeping the order of the sentence. Each individual word is converted to lowercase and stop words are removed. Stop words are words that do not add to sentence context such as "a", "of", "and", "also". The duration of the pre-processing stage is 2660 seconds. Prior to Doc2Vec modelling, the abstracts undergo a vocabulary building phase which identifies unique words and stores it into a dictionary. This phase identifies the weight of each word. Heavier word weight shows higher importance. The process is 4340 seconds. The Doc2vec modelling uses the settings from Fig. 2, frequency of words less than 2 will be removed since they are uncommon words. The window size of 3 influences the target word with 3 nearby words. The vector_size is also known as feature size that captures the feature of a document and converts it into a vector of numbers [12] that is calculated from the weights of words in the vocabulary dictionary. The vector_size is set to 100 which balances between memory usage and feature uniqueness. Epochs are set at 20 with each iteration taking 40 minutes. By running multiple iterations on the data, the Doc2Vec model learns and re-calibrates features in a document. The Doc2Vec model is successfully built into 6.8 million abstract vectors in a process that has taken 17 hours. The model loaded within the Python workspace is displayed in Fig. 5 showing it has been successfully booted.

The unique tokens and the number of documents vectorized in the model is shown in Fig. 6.

D. Doc2Vec Model Validation

The model can be validated by a similarity query of a random patent abstract section from USPTO. Typically, the result of similarity in the model will proceed with high similarity documents, followed by lower similarity documents. The output in Fig. 7 shows that the behavior is as predicted.

The output seen in Fig. 7 uses a section of the abstract as input to find similar documents. The first result has a similarity of 88.57% and on online inspection, found to be the correct patent. Thus, the model can provide accurate similarity. A similarity tolerance test is conducted by running 10 times on the same query shown in Fig. 8. The graph shows that the average similarity value is at 0.885825 with a range of ± 0.000144. Since the value of the range is significantly small, it will not affect the order of top documents in similarity results.

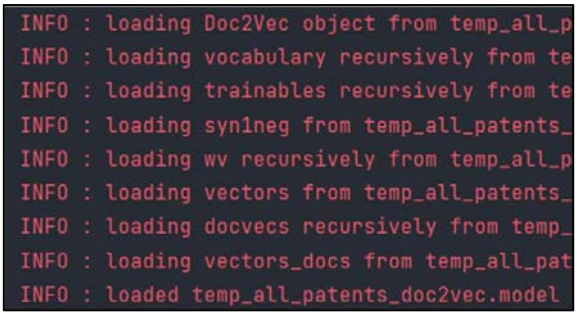


Fig. 5. Doc2Vec Model Logs.

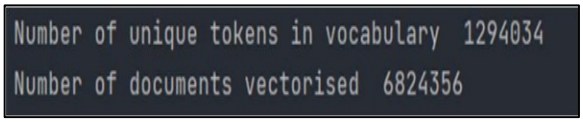


Fig. 6. Output of Token and Vectors Computed.

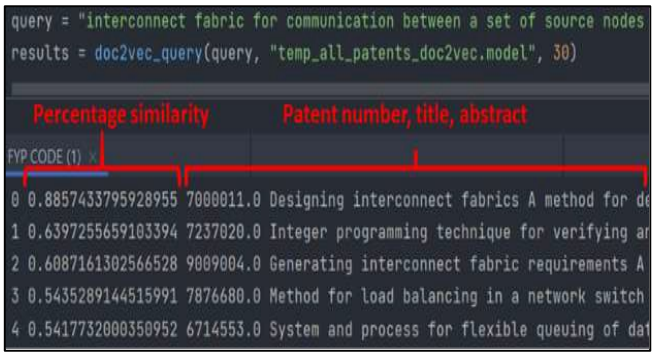


Fig. 7. Similarity Query of a Patent Abstract.

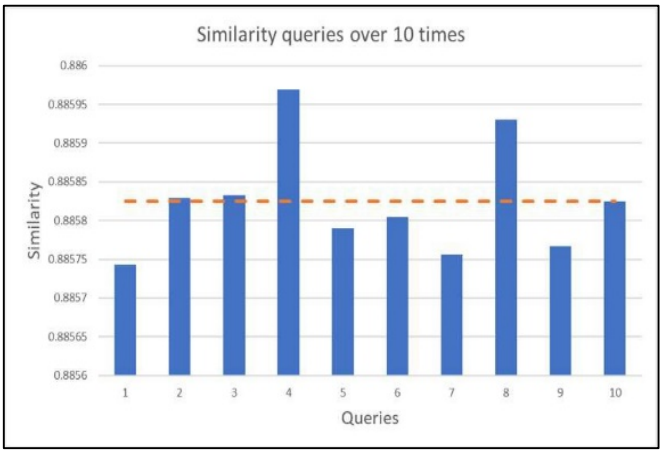


Fig. 8. Plotted Percentages of Similarity Query.

E. Graphical user Interface (GUI) Development

Once the fundamental operation of the program is developed, a GUI is built to accommodate user input. The interface contains the user input of their Functional Statement in three text boxes: Subject, Action, Object. With the guidance of the proposed Functional statement simple solution input, the rest of the TRIZSE database can be selected. The TRIZSE database is laid out with Function, Parameter and Transform sections clearly indicated. The user may select the applicable selections from the available options as well as the ability to exclude un-used sections. The GUI is coded and loaded as shown in Fig. 9, provides enough information to users to perform their input activities.

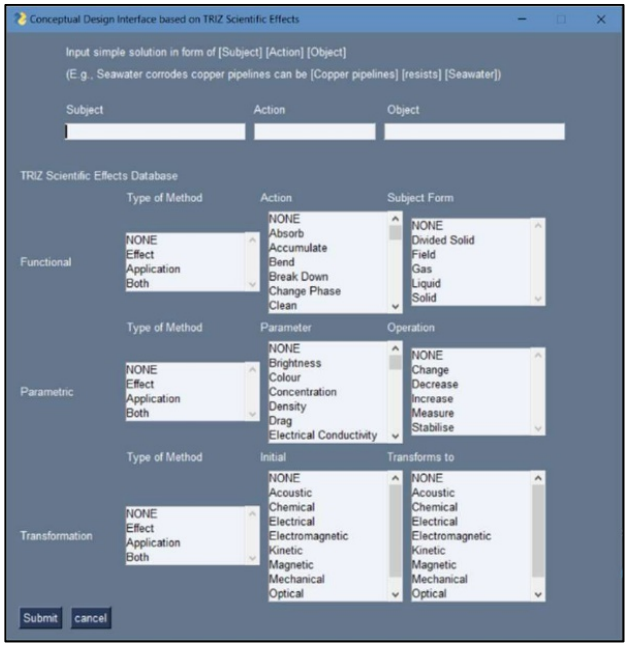


Fig. 9. Graphical user Interface of the Conceptual Design Software Prototype.

F. Case Study on the Corrosion of Copper Pipelines caused by Seawater

Copper alloys are used in ships as pipelines, heat exchanges, screw propellers, valves [18] and most components are often exposed to seawater which is corrosive. In marine engineering applications, copper alloys have an innate seawater corrosion resistance, but will be vulnerable under heavier load conditions [18] such as pressure generated from propeller torque or oscillating temperature in pipeline cooling systems. The case study aims to discover methods to prevent the corrosion of copper pipelines in a marine boat's cooling system. The problem first is approached by using Functional Analysis to identify and break down the problem into a simple problem statement of Subject-Action-Object as shown in Fig. 10.

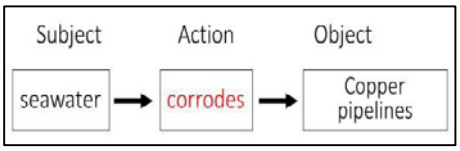


Fig. 10. Problem Statement in Functional Analysis Form.

The simplified problem statement is transformed into a simple solution statement by reversing the subject, object, and providing a counteraction. The solution statement is presented using the same format as shown in Fig. 11.

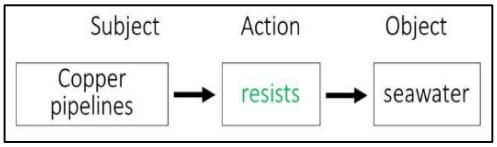


Fig. 11. Solution Statement in Functional Analysis Form.

With the defined solution statement, the GUI can be filled together with the TRIZSE options. Corrosion is an electrochemical reaction that requires surface contact and seawater is a form of liquid. Thus, in the Functional section, the best combination will be to resist liquid as an application. To prevent contact between copper and seawater, the surface finish should not degrade and should have high wear resistance. This can be used in the parameterization section with stabilize surface finish as an application. In the final section, it is unnecessary as the solution should be a preventive measure and thus, assumed to have no transformation of energy in copper pipelines. With the information gathered, the software prototype can be launched and have the input arranged in the GUI as displayed in Fig. 12. From the input screen, it is then submitted and the results of the TRIZSE database are shown in Fig. 13 and Fig. 14 which will be used with the Doc2Vec model to discover similar patent examples.

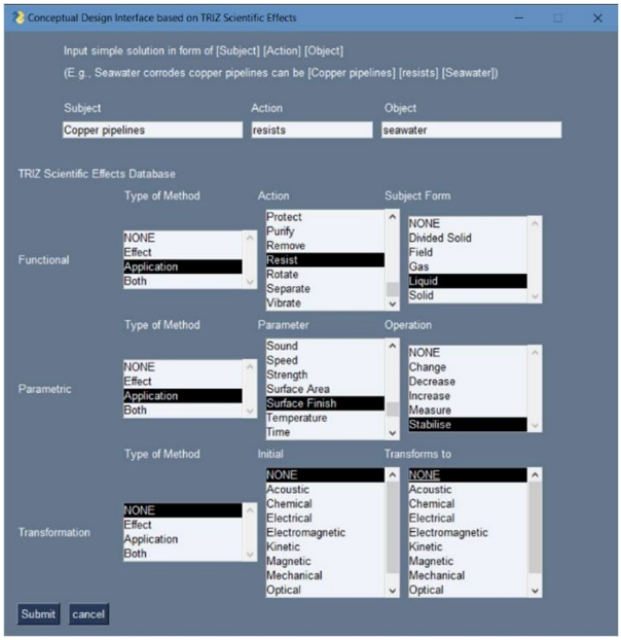


Fig. 12. GUI Interface with Case Study Settings.

As observed in Fig. 13, the list of suggestions from TRIZSE database provides a brief description of each effect with limited examples. A brief review will be conducted to reduce the redundant effects generated. Effects such as Valve and Tesla Valvular Conduit can be removed from consideration as it is an interest of internal volume. The effect Intumescent materials will also be removed as temperature should not be a factor in corrosion resistance. Another similar

review will be conducted on the results of the Parametrization section shown in Fig. 14. From the results of the Parametrization section, some effects can be removed to refine the patent results. Effects such as Redundancy may be considered as a last resort stage and will be removed. The effect Thin Films is also present in the list in Fig. 13 and will be removed since both lists will be compiled before loading into the Doc2Vec model. Using each TRIZSE obtained from the input, it is used together with the subject as a query input to the Doc2Vec model.

Input		TRIZ Scientific Effects	Effect Descriptions
Resist	Liquid	Diamond-like Carbon	(DLC) exists in seven different forms of amorphous carbon materials that display some of the unique properties of natural diamond. They are usually applied as coatings to other materials. All seven contain significant amounts of sp3 hybridized carbon atoms. As well as excellent hardness and wear resistance, DLC is claimed to have the lowest coefficient of friction of any known solid material.
Resist	Liquid	Ferrofluid	A liquid which becomes strongly magnetized in the presence of a magnetic field. Ferrofluids are colloidal mixtures composed of nanoscale ferromagnetic, or ferrimagnetic, particles suspended in a carrier fluid, usually an organic solvent or water. The ferromagnetic nanoparticles are coated with a surfactant to prevent their agglomeration (due to van der Waals and magnetic forces).
Resist	Liquid	Intumescent Materials	An intumescent is a substance which swells as a result of heat exposure, thus increasing in volume, and decreasing in density. Intumescent are typically used in passive fire protection.
Resist	Liquid	Polytetrafluoroethylene (PTFE)	A synthetic fluoropolymer. As fluorocarbons are not as susceptible to van der Waals force (due to the high electronegativity of fluorine) water and water-containing substances, and oil and oil-containing substances, like most foods do not wet PTFE. PTFE is very non-reactive. PTFE's coefficient of friction is 0.1 or less, which is the second lowest of any known solid material.
Resist	Liquid	Solenoid	A coil wound into a tightly packed helix. In physics, the term solenoid refers to a long, thin loop of wire, often wrapped around a metallic core, which produces a magnetic field when an electric current is passed through it. Solenoids are important because they can create controlled magnetic fields and can be used as electromagnets.
Resist	Liquid	Tesla Valvular Conduit	A one-way valve with no moving parts, using the geometry of the fluid path to redirect the flow of the fluid to oppose its own motion in one direction, but offering little resistance in the other.
Resist	Liquid	Thin Films	Thin material layers ranging from fractions of a nanometer to several micrometers in thickness. Electronic semiconductor devices and optical coatings are the main applications benefiting from thin film construction.
Resist	Liquid	Valve	A device that regulates the flow of a fluid (gases, liquids, fluidized solids, or slurries) by opening, closing, or partially obstructing various passageways.

Fig. 13. Output of Functional Section Effects Database.

Input settings		TRIZ Scientific Effects	Effect Descriptions
Stabilize	Surface Finish	Coatings	A covering that is applied to an object. The aim of applying coatings is to improve surface properties of a bulk material usually referred to as a substrate. One can improve amongst others appearance, adhesion, wettability, corrosion resistance, wear resistance, scratch resistance, etc. They may be applied as liquids, gases or solids.
Stabilize	Surface Finish	Electroplating	The process of using electrical current to reduce cations of a desired material from a solution and coat a conductive object with a thin layer of the material, such as a metal. Primarily used to bestow a desired property (e.g., abrasion and wear resistance, corrosion protection, lubricity, aesthetic qualities, etc.) to a surface that otherwise lacks that property. Also used to build up thickness on undersized parts.
Stabilize	Surface Finish	Epitaxy	The method of depositing a monocrystalline film on a monocrystalline substrate. Epitaxial films may be grown from gaseous or liquid precursors. Because the substrate acts as a seed crystal, the deposited film takes on a lattice structure and orientation identical to those of the substrate.
Stabilize	Surface Finish	Feedback	A circular causal process whereby some proportion of a system's output is returned (fed back) to the input. This is often used to control the dynamic behavior of the system.
Stabilize	Surface Finish	Ferromagnetic Powder	Ferromagnetic material in a powdered or finely divided form. Ferromagnetic materials (such as iron) form permanent magnets and/or exhibit strong interactions with magnets. Ferromagnetic materials lose their ferromagnetic properties above a characteristic temperature (the Curie Point).
Stabilize	Surface Finish	Physical Vapor Deposition	A variety of methods used to deposit thin films by the condensation of a vaporized form of the material onto various surfaces
Stabilize	Surface Finish	Preservative	A naturally occurring or synthetic substance that is added to products such as foods, pharmaceuticals, paints, biological samples, wood, etc. to prevent decomposition by microbial growth or by undesirable chemical changes.
Stabilize	Surface Finish	Redundancy	The duplication of critical components of a system with the intention of increasing reliability of the system, usually in the case of a backup or fail-safe.
Stabilize	Surface Finish	Sintering	A method for making objects from powder, by heating the material below its melting point (solid state sintering) until its particles adhere to each other. Traditionally used for manufacturing ceramic objects. Most, if not all, metals can be sintered - especially pure metals produced in vacuum which suffer no surface contamination. Many nonmetallic substances also sinter, such as glass, alumina, zirconia, silica, magnesia, lime, ice, beryllium oxide, ferric oxide, and various organic polymers.
Stabilize	Surface Finish	Thin Films	Thin material layers ranging from fractions of a nanometer to several micrometers in thickness. Electronic semiconductor devices and optical coatings are the main applications benefiting from thin film construction.

Fig. 14. Output of the Parametrization Section Effect Database.

Each query has been set to have an output of the first 10 patents and this setting can be customized as a preference. However, since the similarity of documents is highly saturated in the first few results, the similarity of the rest of the documents may be too broad to provide substantial design support. The output results are compiled, and the top 5 results are shown in Fig. 15. From displayed results in Fig. 15, the outcome of the top 5 highest similarity in patent documents have the effect Physical Vapor Deposition in majority. Analysis of the first patent (10115603) reveals that the patent describes a method of removing passivation films. Passivation film is a coat of protective material typically able to protect against corrosion [19]. However, the patent suggests removal of the said film instead. This displays the model inaccuracy of polarity checking, although being accurate to the topic of corrosion. The second patent (5514414) discusses the deposition of heated flux vapor by condensing onto the subject. The role of flux vapor in corrosion resistance is by removing impurities and oxidation layer [20]. The third patent (9859218) is not applicable towards pipelines as its application is focused on reinforcement of interconnected structures. The fourth patent (5352331) focuses on electroplating a thin film cermet; a composite of ceramic (cer) and metal (met). A thin layer of cermet is capable of producing corrosion resistance and protecting the internal material [21]. The fifth patent (4927472) discusses the formation of tin phosphate on metal surface which creates a layer of corrosion resistance. The idea of phosphating a base material with solution of phosphate ions can provide for corrosion resistance against highly corrosive environments such as seawater [22]. From the first 5 results of the Doc2Vec model, although not all results have the intention of offering corrosion resistance, it is still relevant to the topic scope of the model input.

Effect	Similarity Number	Patent	Title	Abstract
Physical Vapor Deposition	63.45%	10115603	Removal of surface passivation	Methods for removing a passivation film from a copper surface can include exposing the passivation film to a vapor phase organic reactant for example at a temperature of 100° C. to 400° C. In some embodiments the passivation film may have been formed by exposure of the copper surface to benzotriazole such as can occur during a chemical mechanical planarization process. The methods can be performed as part of a process for integrated circuit fabrication. A second material can be selectively deposited on the cleaned copper surface relative to another surface of the substrate.
Physical Vapor Deposition	61.86%	5514414	Solvent-less vapor deposition apparatus and process for application of soldering fluxes	Apparatus and method for condensing a solderless flux vapor onto a work surface to be soldered such as an electronic circuit board. The flux vapor is created by heating flux in a liquid state to a temperature greater than the temperature of the work surface. Flux is applied to the work surface without the use of any volatile organic chemicals.
Physical Vapor Deposition	58.86%	9859218	Selective surface modification of interconnect structures	Semiconductor structures including copper interconnect structures and methods include selective surface modification of copper by providing a CuxTiyNz alloy in the surface. The methods generally include forming a titanium nitride layer on an exposed copper surface followed by annealing to form the CuxTiyNz alloy in the exposed copper surface. Subsequently the titanium layer is removed by a selective wet etching.
Electroplating	58.29%	5352331	Cermet etch technique for integrated circuits	An etching process for patterning thin film cermet (14) on a semiconductor substrate (10) using a mild room temperature acid solution as the etchant. The semiconductor substrate (10) has a glass passivating layer (12) such as silicon dioxide deposited thereon. The cermet layer (14) is deposited on the silicon dioxide layer (12). A photoresist layer (16) is deposited and patterned on the cermet layer (14) followed by the deposition of a layer of aluminum (18). The cermet (14) is then preferentially etched with a mild room temperature hydrofluoric acid solution diluted with hydrochloric acid to form the desired cermet resistance pattern.
Physical Vapor Deposition	56.67%	4927472	Conversion coating solution for treating metal surfaces	A conversion coating predominantly consists of tin phosphate can be deposited on steel or tin-plated steel surfaces by contact with a solution containing phosphate ions tin ions an oxidizing agent such as chlorate and a chelating agent for the tin ions the latter to prevent the rapid loss of tin from the solution that otherwise would occur. A coating that confers excellent resistance to corrosion in hot water is formed on drawn and ironed thinly tin-plated cans.

Fig. 15. Output of the Doc2Vec Model.

V. Conclusion

This research has demonstrated the potential of linking TRIZSE to USPTO patents pool to provide additional support for conceptual design related activities. With the successful building of the Doc2Vec model, a case study has been conducted on the corrosion of copper pipelines in seawater. The concluded discussion of the results of TRIZSE and Doc2Vec model shows that:

- The model provides a majority of relevant examples that ought to be further investigated for the case study.
- The Doc2Vec model does not account for semantics and thus, could not differentiate polarity of relevant examples.
- The improvement of Doc2Vec model can be conducted in stages of pre-processing, feature size, and semantic analysis.

Although the results of patent relevancy are not sufficiently accurate, the linkage of TRIZSE database with USPTO patent pool shows the potential of assisting conceptual design activities. In conclusion, TRIZSE when combined with patent mining provides a viable support means to engineers in developing conceptual solutions by providing relevant examples of solutions in patent documents. Future explorations of the similarity model may be able to provide better accuracy and feature extraction along with semantic analysis. Enhanced accuracy is

necessary as the current search results from the similarity model include accurate relevant concept solutions as well as somewhat inaccurate concept solutions. Future work should only yield search outputs that are only directly relevant patents.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (which is part of the EMJMD Genial Consortium) for supporting this work through the Research University Grant GUP-2018-124.

REFERENCES

[1] K. Gadd and C. Goddard, TRIZ for engineers: enabling inventive problem solving. Chichester, West Sussex: Wiley, 2011.

[2] G. S. Altshuller, And suddenly the inventor appeared: TRIZ, the Theory of Inventive Problem Solving. Worcester, Massachusetts: Technical Innovation Center, Inc., 1996.

[3] A. Souili, D. Cavallucci, and F. Rousselot, "Identifying and reformulating knowledge items to fit with the Inventive Design Method (IDM) model for a semantically-based patent mining," Procedia Engineering, vol. 131, pp. 1130 – 1139, December 2015.

[4] A. Martin. "TRIZ Effects database." http://wbam2244.dns-systems.net/EDB/ (accessed 16 December, 2020).

[5] Creax. "Innovation is crucial - secure your company's future." https://creax.com/innovation-services/ (accessed 16 December, 2020).

[6] Samsung. "□□□ □□: TRIZ school." https://www.seri.org/forum/trizschool/ (accessed 16 December, 2020).

[7] Aulive. "Patent inspiration." https://www.patentinspiration.com/ (accessed 16 December 2020).

[8] J. Delgado-Maciel, G. Cortés-Robles, G. Alor-Hernández, J. García-Alcaraz, and S. Negny, "A comparison between the functional analysis and the causal-loop diagram to model inventive problems," Procedia CIRP, vol. 70, pp. 259-264, January 2018.

[9] L. Fiorineschi, F. S. Frillici, F. Rotini, and M. Tomassini, "Exploiting TRIZ tools for enhancing systematic conceptual design activities," Journal of Engineering Design, vol. 29, no. 6, pp. 259-290, June 2018.

[10] J. Delgado-Maciel, G. Robles, C. Sanchez-Ramirez, J. García-Alcaraz, and J. Méndez-Contreras, "The evaluation of conceptual design through dynamic simulation: a proposal based on TRIZ and system dynamics," Computers & Industrial Engineering, vol. 149, p. 106785, November 2020.

[11] C. Muenzberg, K. Michl, H. Heigl, T. Jeck, and U. Lindemann, "Further development of TRIZ function analysis based on applications in projects," in International Design Conference - DESIGN 2014, Dubrovnik, Croatia, May 19-22 2014, pp. 333-342.

[12] L. Quoc and M. Tomas, "Distributed representations of sentences and documents," in The 31st International Conference on Machine Learning, Beijing, China, June 2014, vol. 32, no. 2: PMLR, pp. 1188-1196.

[13] E. Tostrup and S. Mesic, "Massive patent data mining," Master's thesis, Department of Computer Science, LTH, Lund University, Sweden, 2019.

[14] J. Kim and S. Lee, "Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO," Technological Forecasting and Social Change, vol. 92, pp. 332-345, March 2015.

[15] C. K. Chan, K. W. Ng, M. C. Ang, C. Y. Ng, and A. Kor, "Sustainable product innovation using patent mining and TRIZ," in Advances in Visual Informatics, Lecture Notes in Computer Science, H. Badioze Zaman et al. Eds. Cham: Springer, 2021, pp. 287–298.

[16] M. Ghane, M. C. Ang, R. A. Kadir, and K. W. Ng, "Technology forecasting model based on trends of engineering system evolution (TESE) and big data for 4IR," presented at the 2020 IEEE Student Conference on Research and Development (SCOReD), 27-29 September, 2020.

[17] R. Řehůřek. "Doc2Vec Model — gensim." https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html (accessed 3 March, 2021).

[18] D. Féron, Corrosion behaviour and protection of copper and aluminium alloys in seawater. Padstow, Cornwall: CRC Press, 2007.

[19] J. Zhao et al., "Effect of surface passivation on corrosion resistance and antibacterial properties of Cu-bearing 316L stainless steel," Applied Surface Science, vol. 386, pp. 371-380, November 2016.

[20] J. Vimala, M. Natesan, and S. Rajendran, "Corrosion and protection of electronic components in different environmental conditions - an overview," The Open Corrosion Journal, vol. 2, pp. 105-113, October 2009.

[21] A. Tiwari, S. Seman, G. Singh, and R. Jayaganthan, "Nanocrystalline Cermet Coatings for Erosion–Corrosion Protection," Coatings, vol. 9, no. 6, p. 400, June 2019.

[22] D. Burduhos Nergis, P. Vizureanu, and C. Bejinariu, "Phosphate surface treatment for improving the corrosion resistance of the C45 carbon steel used in carabiners manufacturing," Materials, vol. 13, no. 15, p. 3410, August 2020.

On Validating Cognitive Diagnosis Models for the Arithmetic Skills of Elementary School Students

Hyejung Koh[1], Wonjin Jang[2], Yongseok Yoo[3]*
Uiduk University, Gyeongju 38004, Republic of Korea[1]
Incheon National University, Incheon, 22012, Republic of Korea[2, 3]

***Abstract*—Cognitive diagnosis models (CDMs) have been shown to provide detailed evaluations of students' achievement in terms of proficiency of individual cognitive attributes. Attribute hierarchy model (AHM), a variant of CDM, takes the hierarchical structure of those cognitive attributes to provide more accurate and interpretable measurements of learning achievement. However, advantages of the richer model come at the expense of increased difficulty in designing the hierarchy of the cognitive attributes and developing corresponding test sets. In this study, we propose quantitative tools for validating the hierarchical structures of cognitive attributes. First, a method to quantitatively compare alternative cognitive hierarchies is established by computing the inconsistency between a given cognitive hierarchy and students' responses. Then, this method is generalized to validate a cognitive hierarchy without real responses. Numerical simulations were performed starting from an AHM designed by experts and responses of elementary school students. Results show that the expert-designed cognitive attribute explains the students' responses better than most of alternative hierarchies do, but not all; a superior cognitive hierarchy is identified. This discrepancy is discussed in terms of internalization of cognitive attributes.**

Keywords*—Cognitive diagnosis model; attribute hierarchy model; cognitive hierarchy; model validation

I. Introduction

A. Cognitive Diagnostic Models

Learning analytics has attracted much attention recently, as more data become available for educators and learners [1]. Vast amounts of data are generated in the field of education, due to the widespread adoption of online education systems, such as massive open online courses [2]. By utilizing more data, a more accurate and detailed assessment of learning achievement is possible [3], resulting in enhanced learning experience [4]. For instance, teachers could offer individualized learning strategies tailored to needs of the target students, such as university students [5], under-represented students [6] or foreign language learners [7].

Cognitive diagnosis models (CDMs) have been actively studied as a useful tool for assessing students' knowledge states in terms of multiple cognitive attributes [8]. CDMs incorporate multiple cognitive attributes that are required to understand a concept or to perform a task. Items that require different combinations of cognitive attributes are developed, and the degree of proficiency for each cognitive attribute is estimated from the students' responses to these items. The quantitative assessment of a student's proficiency of individual cognitive attributes allows a more detailed evaluation of a student's achievement, compared to a total score-based learning diagnosis.

CDMs are largely divided into two groups: compensatory models and non-compensatory models [9]. Compensatory models assume that one cognitive attribute could compensate for another. Thus, even if a particular cognitive attribute is not mastered by the examinee, they may solve an item by using other cognitive attributes. For instance, reading comprehension requires numerous cognitive attributes, such as grammar and vocabulary. Even if there are a few words that a reader is not familiar within a given text, the reader could postulate the meanings of the words from the grammatical structure and solve related items correctly. In such cases, grammar plays a compensatory role for vocabulary.

In contrast, non-compensatory models assume that the lack of a cognitive attribute is not compensated for by other cognitive attributes [10]. Therefore, if one fails to master any of the cognitive attributes required, an item cannot be solved. For example, according to the non-compensatory model, an item requiring the concepts of logarithmic function and algebraic function could only be correctly solved by those who have mastered both concepts. If any one of the two concepts is lacking, they will not be able to correctly solve the item.

B. Related Work

Among early non-compensatory models, the rule-space model (RSM) was explored and established by Tatsuoka [11]. The relationship between an item and the cognitive attributes required to solve the item is represented by a matrix, called the Q-matrix. Each row of the Q-matrix corresponds to an item, while each column corresponds to a cognitive attribute; Q_{ij} is one if the j^{th} cognitive attribute is required to solve the i^{th} item. Otherwise, Q_{ij} is zero. The RSM was developed to estimate students' knowledge states from their item responses and a corresponding Q-matrix.

In this study, we adopt a variant of the RSM, called the attribute hierarchy model (AHM) [12] to quantify elementary school students' learning achievements in arithmetic operations. Similarly to CDMs, the AHM aims to represent the relationship between items and cognitive attributes. Furthermore, the AHM includes hierarchical relationships between the cognitive attributes in the model. This hierarchical structure of cognitive attributes is represented by a graph, in which each node corresponds to a cognitive

*Corresponding Author.

attribute, and an edge between two nodes implies that one cognitive attribute is a prerequisite for the other cognitive attribute. Therefore, the AHM is suitable for evaluating the achievement of mathematics subjects, where the hierarchy of cognitive attributes is important [13].

A precise relationship between cognitive attributes and items is crucial for the accurate measurement of learning achievement. However, in practice, developing a Q-matrix and corresponding items, requires months of intensive collaboration between modeling experts and teachers in the field.

To assist the development of the Q-matrix, several quantitative tools have been proposed. First, de la Torre [14] proposed to validate a given Q-matrix by the degree of agreement with the response data using the EM algorithm [15]. Such quantification provides an objective measure to compare multiple Q-matrices. However, it remains elusive which Q-matrices should be compared. To address this gap, DeCarlo proposed a Bayesian framework to validate individual elements of a given Q-matrix [16]. However, the flexibility of the Bayesian model comes with an increased complexity of the validation procedure because the reliability of each element of the Q-matrix must be provided in advance. Last, Chiu proposed a simpler nonparametric method to identify misspecified entries of a Q-matrix [17].

However, those validation methods are applicable only to RSMs, not to AHMs. The hierarchical structure of the cognitive attributes is not considered for the validation of a given Q-matrix. Therefore, adopting AHMs in practice requires extention of the validation framework and more careful consideration of the structure of the cognitive attributes of interest.

C. Contributions of the Study

We propose quantitative tools for validating the hierarchical structures of cognitive attributes, to aid the development of AHMs. Whereas the current validation methods quantify the validity of each element of the Q-matrix, we explore alternative hierarchical structure of the cognitive attributes. Thus, the search space of out method is a graph rather than a matrix.

Our approach would provide a more natural way to validate AHMs in conjunction with the experts' domain knowledge. For instance, the current methods provide a refined Q-matrix with some elements flipped from the initial Q-matix. However, such a refined Q-matrix may contradict to the hierarchical structure of cognitive attributes developed by the experts. Instead, our method starts from the experts' knowledge in terms of the hierarchical structure and validate individual associations of cognitive attributes.

The rest of this paper is organized as follows: the Methods section describes the AHM model, as well as the data collection and evaluation of alternative hierarchies of cognitive attributes. In the Results section, we show numerical simulations; finally, in the Conclusions and Discussions section, we discuss the results and their implications, with concluding remarks.

II. Methods

A. Modeling and Data Collection

An AHM model for arithmetic operations with natural numbers was designed as follows. First, seven cognitive attributes were chosen based on the elementary school mathematics curriculum – addition (A1), subtraction (A2), multiplication (A3), division (A4), carry (A5), borrow (A6), and '0' in multiplication (A7). Thereafter, the hierarchy of the seven attributes (H_0, Fig. 1A) was designed by experts. In brief, the root node (A1) represents addition, which is the prerequisite for all other cognitive attributes. The descendant nodes of the root node are A2, A3, and A5, which correspond to subtraction, multiplication, and carry, respectively. To understand a leaf node (A4, A6, or A7, corresponding respectively to division, borrow, and "0" in multiplication), one should master all the preceding cognitive attributes up to the root node. For example, A4 (division) requires A1 (addition), A2 (subtraction), and A3 (multiplication).

Based on the expert-designed hierarchical structure (H_0, Fig. 1A), alternative hierarchical structures ($H_1, H_2, \cdots, H_7$) were created by removing an edge from H_0 (Fig. 1B). For example, H_1 was generated by removing the edge between A1 and A2 (dashed line in Fig. 1B) from H_0. Similarly, the removed edges in H_k for $k = 2, 3, \cdots 7$ are indicated by (H_k).

Next, thirty items involving the seven cognitive attributes were developed. Our participant sample comprised 977 fourth graders who participated in the test; their responses to individual items were coded as either correct (1) or incorrect (0).

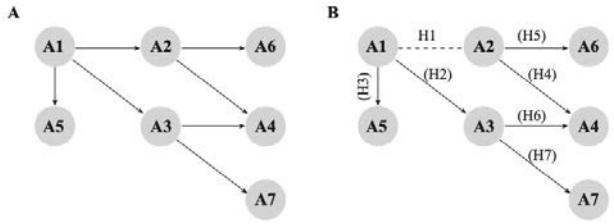


Fig. 1. (A) The Hierarchical Structure (H_0) of the Seven Cognitive Attributes (A1, A2, ⋯, A7) Designed by the Experts. (B) Alternative Hierarchical Structures ($H_1, H_2, \cdots, H_7$) were Created by Removing One Edge from H_0. For Example, the Graph in Panel B shows H_1, Generated by Removing the Edge between A1 and A2 (Dashed Line) from H_0. Similarly, other Hierarchical Structures (in Parentheses) are Generated by Removing the Corresponding Edges.

B. Validating the Hierarchies of Cognitive Attributes

The alternative hierarchies were validated by quantifying the degree to which a given hierarchy is in agreement with students' actual responses. The block diagram in Fig. 2 summarizes the quantification steps, each of which is explained as follows.

First, a different attribute hierarchy (H_k) implies a different structure of students' knowledge states. For the seven cognitive attributes, $2^7 = 128$ combinations of the seven attributes are enumerated. Among all the potential combinations, those that are in conflict with H_k are eliminated (Leighton et al., 2004) and the remaining attribute combinations comprise the set of valid knowledge states S_k.

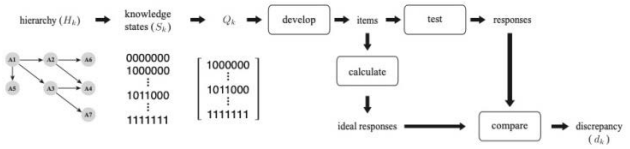


Fig. 2. A Schematic Diagram of the Quantitative Validation of a Hierarchical Structure. For any given Hierarchical Structure H_k, Corresponding Knowledge States S_k (Combinations of Attributes Consistent with H_k) and a Q-matrix Q_k are Generated. Items Corresponding to the Rows of Q_k are Developed and Students' Responses (r) to these Items are Collected. Ideal Responses (r_{ideal}) are Theoretically Calculated from the same Items. Comparing r and r_{ideal} Produces a Discrepancy Value d_k, which Quantifies the Validity of H_k; a Lower d_k Implies Higher Validness.

Next, the Q-matrix Q_k is defined by taking all the valid states in S_k—except the all-zero state (0000000)—as its rows.

Three statistical characteristics of each Q_k-matrix are then measured. First, the sum of each column of Q_k (column sum) indicates the number of items that require the corresponding cognitive attribute. Second, the sum of each row of Q_k (row sum) indicates the number of attributes included in the corresponding item. Third, the sparsity of Q_k is defined by the number of ones divided by the number of elements of Q_k, which corresponds to the frequency of attributes examined in the test set.

Next, we develop an item involving corresponding attributes for each row of Q_k, and students' item responses (r) to the items, which are collected through tests. The set of all the item responses is called R.

The ideal responses (r_{ideal}) to the items are theoretically generated. Here, *ideal* means that the response is solely based on the mastery of the cognitive attribute, excluding guessing or mistakes. For each knowledge state $s \in S_k$, an ideal response r_{ideal} is calculated by Equation 1.

$$r_{ideal} = \sim((\sim s){Q_k}^T), \quad (1)$$

where $\sim$ and T mean the logical NOT operator and the matrix transpose, respectively. The set of ideal responses for all the states in S_k is called I_k.

Last, the discrepancy between the ideal responses (I_k) and the actual responses (R) is calculated as follows. A lower discrepancy value indicates that H_k provides a better account for actual students' responses. First, the discrepancy for each response $r \in R$ is defined by the Hamming distance (the number of mismatches) between r and the closest r_{ideal} in I_k, presented as $h(r, I_k)$. The average discrepancy of responses is normalized by the number of items and defined as the discrepancy for H_k (Equation 2).

$$d(H_k) = \frac{\sum_{r \in R} h(r, I_k)}{N_R N_I}, \quad (2)$$

where N_R and N_I are the numbers of responses and items, respectively.

C. Generating Virtual Responses from an Existing Dataset

Ideally, a different test set would be developed and responses for each hierarchical structure H_k would be collected. However, this would be costly and require too much time. Developing a test set for H_k requires determining the combinations of cognitive attributes based on H_k and developing corresponding items; each of these steps requires repeated feedback from experts. Recruiting test takers for multiple test sets is also costly. Repeating multiple test sets involving the same set of cognitive attributes for a fixed target group would be impractical. Students might learn to recognize patterns during the sequence of similar tests or become bored by the similarity of multiple tests. In either case, collecting unbiased responses for each H_k is a challenging task. Furthermore, if each test is performed on a different group of students, it is unclear whether any difference in item responses is due to the different hierarchical structure or the heterogeneity of the student groups.

To overcome this limitation, we propose to generate virtual responses to each H_k by employing random sampling using a common dataset (Fig. 3). Specifically, for each item, corresponding attributes (each row of Q_k) are used as a query to retrieve responses to items that require attributes similar to the query from the existing database. To simplify, attributes that have the smallest Hamming distance to the query attribute are chosen as candidates first. Thereafter, if there are multiple candidates, one is randomly selected with equal probability. Otherwise, (if there is only one item with the smallest Hamming distance), the candidate response is the virtual response. Repeating the above procedure for all the items (all the rows of Q_k) comprises an iteration of simulation. Subsequently, the average discrepancy is measured for 1,500 iterations for each H_k. Additionally, the statistical significance of the discrepancy is measured by repeating the above calculations, starting from 10 different random seeds.

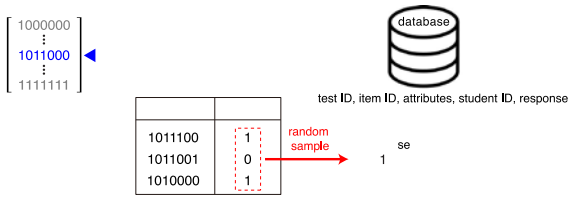


Fig. 3. Virtual Item Responses are Generated from an Existing Database. Each Row of Q_k is used as a Query to Retrieve Candidate Responses to Items with Similar Attributes. among these Candidate Responses, One Response is Randomly Sampled and used as a Virtual Response.

III. Results

A. Statistical Characteristics of the Q-matrix

Even with the same set of cognitive attributes, different hierarchical structures necessitate items with different combinations of cognitive attributes; these result in different numbers of rows in the corresponding Q-matrices. Removing an edge from H_0 increases the number of rows of Q_k (27 ~ 33) for $k = 1, 2, \cdots, 7$ (Fig. 4). This is because more items are needed to determine participants' mastery of independent attributes than what are needed for related attributes. More specifically, Q-matrices corresponding to H_5 and H_7 had the largest number of rows (33). Compared with H_0, H_5 and H_7 have independent nodes A6 and A7, respectively; these nodes had many preceding attributes in H_0.

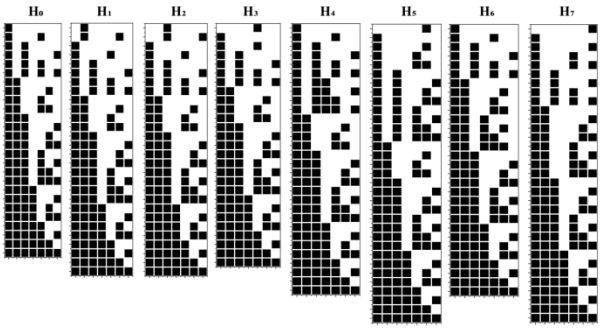


Fig. 4. The Q-matrices (Q_k) for different Hierarchical Structures (H_k). The Rows and Columns of each Q_k are Associated with Items and Attributes, Respectively. Each Row of Q_k Represents Cognitive Attributes Required for each Item in Black.

Even though the number of rows of Q_k differ, the frequencies of ones in Q_k are similar. The average numbers of items per attribute did not differ significantly ($p = 0.970$, one-way ANOVA). Similarly, the average numbers of attributes per item did not differ significantly ($p = 0.998$, one-way ANOVA). Finally, the average sparsity of Q_k was $0.571\ (\pm 0.012)$. Thus, on average, hierarchies (H_k) are homogeneous in terms of the number of items per cognitive attribute and the number of cognitive attributes per item.

B. Comparison of the Discrepancy for Each Hierarchy

The hierarchy designed by the experts (H_0) had a lower discrepancy than all the alternative hierarchies, except for H_2 (Fig. 5); this indicates that H_0 explains students' responses better than most alternative hierarchies, except for H_2. The increase in the discrepancy between the alternative hierarchies implies that the removed edge is important for explaining the actual students' responses.

In contrast, $d(H_2)$ was lower than $d(H_0)$ (Fig. 6). The *t*-test for 10 simulations with different random seed numbers confirmed that there was a significant difference between $d(H_0)$ and $d(H_2)$ ($p = 4.18 \times 10^{-18}$). This result implies that H_2 explains the students' real item responses better than H_0. There is thus a hierarchy of cognitive attributes that better describes responses than the hierarchy designed by the experts.

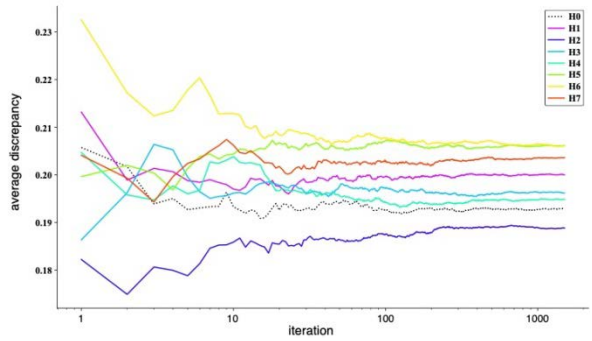


Fig. 5. Average Discrepancy (d_k) of each Hierarchy (H_k) as a Function of Iteration. During Early (< 10) Iterations, d_k Fluctuates Considerably; However, it Stabilizes after about 500 Iterations. After 1,500 Iterations, the Value of d_k was Measured for each H_k. The Hierarchy H_k, which was Designed by the Experts, showed Lower d_k (Black, Dotted) than Most of the other H_k, Except for H_2 (Purple).

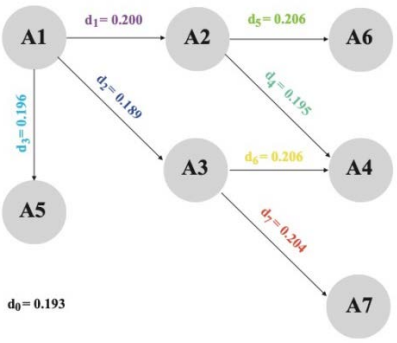


Fig. 6. The Discrepancy (d_k) of each Hierarchy (H_k) is shown on the Removed Edge (for $k > 0$). The Hierarchy without the Edge between A1 and A3 (H_2) showed the Lowest d_k, even Lower than that of the Expert-Designed Hierarchy (d_0).

IV. Conclusion and Discussion

In this study, we propose a method to quantitatively validate the hierarchy of cognitive attributes of a CDM and corresponding Q-matrices. The hierarchy designed by experts (H_0) was compared with alternative hierarchies ($H_1, H_2, \cdots, H_7$) with an edge removed. The discrepancy for each hierarchy was defined by the distance between the real and ideal responses (r and r_{ideal}), the inverse of which is interpreted as a quantitative indicator of how well a hierarchy and the corresponding Q-matrix describe the students' item responses.

Virtual responses were generated from an existing database, rather than directly collecting responses for each hierarchy and its corresponding Q-matrix. After generating a Q-matrix that corresponds to each hierarchy, we selected the items with the closest Hamming distance ($d_{Hamming}$) to each row of Q by comparison with the existing datasets, and one of the responses to these items was randomly selected as a virtual response.

The hierarchy of cognitive attributes designed by the experts (H_0) had generally lower discrepancy than alternative hierarchies; however, one hierarchy (H_2) had a lower discrepancy than H_0. The difference between H_0 and H_2 was the edge between addition (A1) and multiplication (A3), which is present in H_0, but absent in H_2. This implies that the link between addition and multiplication might be weaker than was expected by the experts.

Our interpretation of this gap is that multiplication, once acquired as a separate skill, may not require the concept of addition. The concept of multiplication can be divided into three categories: repetitive addition, multiples, and product set [18]. The first concept of multiplication—repetitive addition—is utilized to teach multiplication to first-time learners. It is therefore reasonable to assume that understanding or performing multiplication also relies on the knowledge of addition, in agreement with the hierarchy designed by the experts (H_0). However, the other aspects of multiplication may play more important roles for students who mastered the concept of multiplication. In other words, after acquiring the knowledge of multiplication, they may perform multiplication as an independent skill, rather than repeating

addition multiple times. Therefore, we posit that the relationship between addition and multiplication may be important for learning, but that it is less so for practicing the actual skill of multiplication.

We propose a scalable validation tool for comparing alternative hierarchies, which could encourage more teachers to utilize CDMs for learning achievement analyses. Selecting relevant cognitive attributes and designing the hierarchy among them requires experts' knowledge and experience, which could hinder wider uses of CDMs. When a user wants to validate a chosen hierarchy or explore alternative hierarchical structures, item responses are sampled from an existing database, without having to develop new items and collect responses for each candidate. The proposed quantitative validation of alternative hierarchies could be used as objective indicators of the validity of established hierarchies.

Our future work will generalize the proposed framework to more complex cases. In this study, we considered only seven attributes with rather simple associations. In general, presence of loop a graph hinders theoretical analysis as well as numerical calculations based on the graph. The hierarchical structure in this study has only one loop. It is of great interest to explore a more complex attributes structure with multiple loops.

ACKNOWLEDGMENT

This study was supported by Incheon National University Research Grant in 2019.

REFERENCES

[1] Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), 2020.

[2] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some scaling laws for MOOC assessments," KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014), 2014.

[3] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, and S. N. Liao, "Predicting academic performance: a systematic literature review," Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education, 2018.

[4] K. Mangaroska and M. Giannakos, "Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning," IEEE Transactions on Learning Technologies, 12(4), 516-534, 2018.

[5] H. Aldowah, H. Al-Samarraie, H., and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis." Telematics and Informatics, 37, 13-49, 2019.

[6] A. Cano and J. D. Leonard, "Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations," IEEE Transactions on Learning Technologies, vol. 12, no. 2, pp. 198-211, 2019.

[7] Bravo-Agapito, J., Bonilla, C. F., & Seoane, I, "Data mining in foreign language learning," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1287, 2020.

[8] M. von Davier and Y. S. Lee, Handbook of diagnostic classification models, Springer International Publishing, 2019.

[9] A. A. Rupp and J. L. Templin, "Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art," Measurement, 6(4), 219-262, 2008.

[10] M. Birenbaum, A. E. Kelly, and K. K. Tatsuoka. "Diagnosing knowledge states in algebra using the rule-space model," Journal for Research in Mathematics Education, 24(5), 442-459, 1993.

[11] K. K. Tatsuoka, "Rule space: An approach for dealing with misconceptions based on item response theory," Journal of educational measurement, 345-354, 1983.

[12] J. P. Leighton, M. J. Gierl, and S. M. Hunka, "The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach," Journal of educational measurement, 41(3), 205-237, 2004.

[13] M. J. Gierl, C. Alves, and R. T. Majeau, "Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment," International Journal of Testing, 10(4), 318-341, 2010.

[14] J. de la Torre, "An empirically based method of Q-matrix validation for the DINA model: Development and applications," Journal of Educational Measurement, 45, 343-362, 2008.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, 39(1), 1–38, 1977.

[16] L. T. DeCarlo, "Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model," Applied Psychological Measurement, 36, 447-468, 2012.

[17] C. Y. Chiu, "Statistical refinement of the Q-matrix in cognitive diagnosis," Applied Psychological Measurement, 37, 598-618, 2013.

[18] H. K. Kang, An Alternative Program for the Teaching of Multiplication Concept Based on Times Idea. Journal of Korea Society of Educational Studies in Mathematics, 11(1), 17-37, 2009.

Gabor Descriptor for Representation of Spatial Feature

Comparison Between the Proposed Descriptor and the Conventional Fourier Descriptor

Kohei Arai
Information Science Department
Saga University, Saga
Japan

***Abstract*—New spatial feature descriptor based on Gabor wavelet function is proposed. The proposed method is compared to Fourier descriptor. The experimental results with Advanced Earth Observing Satellite: ADEOS / Advanced Visible and Near-Infrared Radiometer: AVNIR image show an effectiveness of the proposed method. It is found that the restored image quality, in terms of root mean square error between the original and the restored images depends on the support length of the mother wavelet and is much better than that with the conventional Fourier descriptor method for spatial feature description.**

Keywords*—Spatial feature; Gabor wavelet descriptor; Fourier descriptor; wavelet transformation; ADEOS (advanced earth observing satellite); AVNIR (advanced visible and near-infrared radiometer)

I. Introduction

There are various examples of applying wavelet analysis to the processing analysis of earth observation satellite images. A method of superimposing a plurality of visible images after wavelet transform [1], a method of superimposing a plurality of SAR: Synthetic Aperture Radar images with different off-nadir angles after wavelet transform [2], and an annual fluctuation pattern of sea surface temperature estimated from satellite data. Method [3] that applies wavelet transform to extract its features, method [4] that applies wavelet transform to the extraction of surface roughness of sea ice, method [5] that extracts spatial features from images extracted from soil moisture, etc. There are [6], [7]. As a report on image processing analysis using the Gabor transform, a report on compression [8] – [11], a report on texture analysis [12] – [15], a report on reconstruction [16] – [21], and a report on feature extraction [22] – [25], there are reports on classification [26] – [28].

However, as far as the authors are aware, there is no example of applying wavelet analysis to the method of describing the geometrical features of an image. The Fourier descriptor is well known as a method of describing the geometrical features of an image [29] – [31]. The Fourier descriptor is the one that the contour line (closed curve) extracted from the image is Fourier expanded, and the closed curve is described by a finite number of Fourier expansion coefficients. The Fourier descriptor does not describe the characteristic points of the closed curve in detail but describes all points using the sin function (cos function). Also, some proposals have already been made for the method of applying the Fourier descriptor to open curves.

On the other hand, the Gabor descriptor proposed in this paper takes advantage of the characteristics of Gabor transformation and is devised so that it is possible to describe the complicated part of the contour line in detail. The author considered that the Gabor descriptor enables the description of geometrical features with less information than the Fourier descriptor with less deterioration [32] – [34]. Furthermore, remote sensing satellite image database system allowing image portion retrievals utilizing principal component which consists spectral and spatial features extracted from imagery data is proposed recently [35].

In this paper, the author tries to extend the descriptor by Fourier expansion and opening to the descriptor by Gabor expansion. The proposed method was applied to the Visible and Near Infrared Radiometer: VNIR data of the Advanced Spaceborne Thermal Emission and Reflection Radiometer: ASTER sensor mounted on the Terra satellite to confirm that it is superior to the existing method and to change the support length in Gabor conversion. The author also confirmed the effect of the proposed method.

The next section describes theoretical background and the proposed method for spatial feature descriptor based on Gabor wavelet function. Then some experiments are described followed by conclusion together with some discussions.

II. Theoretical Background and Proposed Method

A. Fourier Expansion and Gabor Wavelet Expansion

The Fourier transform of the complex function f given the contour line (closed curve etc.) extracted from the image is defined by the equation (1).

$$F(\omega) = \int_{-\infty}^{\infty} \overline{e^{\iota\omega t}} f(t)dt \tag{1}$$

The Gabor transform is defined by equations (2) and (3).

$$W(\omega,\sigma,k) = \int_{-\infty}^{\infty} \overline{\exp\left(-\frac{(t-k)^2}{\sigma^2}\right)e^{\iota\omega t}}) f(t)dt \tag{2}$$

$$= \int_{-\infty}^{\infty} \overline{e^{\iota\omega t}}\exp(-\frac{(t-k)^2}{\sigma^2} f(t)dt \tag{3}$$

However, $\overline{e^{\iota\omega t}}$represents the complex conjugate of $e^{i\omega t}$ and $\overline{\exp\left(-\frac{(t-k)^2}{\sigma^2}\right)e^{\iota\omega t}}$ expresses the complex conjugate of $\exp(-\frac{(t-k)^2}{\sigma^2})e^{i\omega t}$. ω is a parameter in the frequency axis direction, and σ and k are parameters in the time axis direction.

The Gabor transform is a Gaussian function with a window function introduced into the Fourier transform. That is, the Gabor transform is a kind of Fourier transform with window function. In particular, when *k*=0, equation (2) becomes equation (4).

$$W(\omega,\sigma,0)=\int_{-\infty}^{\infty}\overline{\exp\left(-\frac{(t)^2}{\sigma^2}\right)e^{\iota\omega t}}\,f(t)dt \quad (4)$$

When σ = ∞, equation (2) becomes equation (5).

$$W(\omega,\infty,k)=\int_{-\infty}^{\infty}\overline{e^{\iota\omega t}}\,f(t)dt \quad (5)$$

Therefore, it can be seen that it does not depend on the parameter *k*.

According to Eq. (2), if the value of the parameter σ is large, the influence of the parameter *k* on *W(ω,σ,k)* is small. If the value of the parameter σ is small, the influence of the parameter *k* on *W(ω,σ,k)* is large.

The Gabor transform is not restored to the original data when the inverse transform is performed because it is not an orthogonal transform. However, even in the Fourier transform which is the orthogonal transform, for example, when the inverse transform is performed through the low-pass filter, the original data is not restored. The reason why Gabor transform is not orthogonal transform is that it introduces window function into Fourier transform. In other words, it is necessary to consider the effect of the window function when performing Gabor transformation.

B. Fourier Descriptor and Gabor Wavelet Descriptor

In the Fourier descriptor, the extracted spatial feature information is Fourier expanded and the expansion coefficient is used as the descriptor. In the Gabor descriptor, the extracted spatial feature information is Gabor expanded, and the expansion coefficient is used as the descriptor. The Fourier descriptor requires operations from -∞ to ∞ according to Eq. (1). The Gabor descriptor proposed in this paper does not require operations from -∞ to ∞ by introducing the window function (Gaussian function) in Eq. (2). However, it is assumed that the tail of the Gaussian distribution is regarded as zero. From the viewpoint of the above calculation amount, the author thinks that the Vega ball descriptor is superior to the Fourier descriptor.

The number of non-zero Gabor expansion coefficients is called the support length. It can be seen from Eq. (2) that the support length depends on the parameter σ.

C. Gabor Descriptor for Closed Curve

Let *Z(l)=(x(l), y(l))* be the closed curve. *Z(l)* is a point that has advanced a distance *l* on the closed curve from a certain start point on the closed curve as shown in Fig. 1.

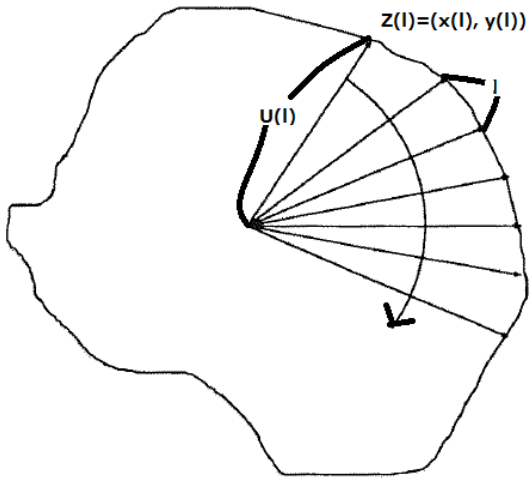


Fig. 1. Concept of the Proposed Wavelet Descriptor.

Here, if the complex function *u(l)* is defined as follows,

$$u(l)=x(l)+iy(l),\left(i=\sqrt{-1}\right) \quad (6)$$

The complex function *u(l)* is the total length *L* of the closed curve which is a periodic function. The Gabor descriptor is realized by performing Gabor expansion on the complex function *u(l)*.

D. Feature of the Proposed Gabor Wavelet Descriptor

The proposed method considers the information of the edge of the closed curve of the image as a periodic complex function and describes and saves the geometrical feature of the image using the information of the position of a certain starting point on the closed curve and the Gabor expansion coefficient. The proposed method has the following features,

1) When compared with the method of saving edge information as a binary image (method of saving as "face" information), it is superior in terms of the amount of saved data. That is, information of pixels that are not edges is not included.

2) The Fourier descriptor requires operations from -∞ to ∞, but the Gabor descriptor does not require operations from -∞ to ∞.

III. Experiments

A. Experimental Method

The data used this time are actual observation images near Mt. Usu acquired by the sensor AVNIR onboard the ADEOS satellite. Fig. 2 shows the aerial photo of Aza-Nakanoshima island near the Mt. Usu while Fig. 3 shows the acquired Advanced Visible and Near Infrared Radiometer: AVNIR (onboard on Terra Satellite) image.

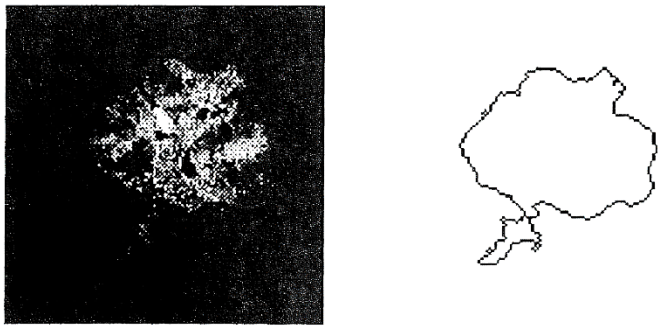
Fig. 3(a) shows an actual observation image, and Fig. 3(b) shows an image resulting from contour extraction. In this experiment, the experiment is performed using Fig. 3(b). Fig. 4 shows the distance from a certain point inside the closed curve in Fig. 3(b) to a point on the closed curve. The purpose of this experiment is to compare the effectiveness of the wavelet descriptor with that of the Fourier descriptor and to

confirm it by evaluating the restoration accuracy. Therefore, the result obtained by halving the number of Fourier expansion coefficients by a low pass filter is used in the Fourier descriptor, and the result obtained by only the low frequency component is used in the wavelet descriptor. At the same time, the impact of the support length of the wavelet will be examined. In addition, equation (7) is used to evaluate the restoration accuracy.

$$J_1 = \sqrt{\frac{\sum_{s=1}^{N}\{(\widehat{x(s)}-x(s)\}^2+\{(\widehat{y(s)}-y(s)\}^2}{N}} \qquad (7)$$

Note that from Fig. 4, N=256.

Fig. 2. Google Map (Aerial Photo Image) of Aza Nakanoshima near the Mt.Usu of Intensive Study Area.



(a) Original AVNIR Image. (b) Detected Edge.

Fig. 3. Acquired AVNIR Image of Aza-Nakanoshima Island near the Mt. Usu and the Detected Edge Image.

Fig. 4 is an example of constructing a complex function *u(l)* based on Fig. 3(b). The vertical axis in Fig. 4 is the time axis. The result of projecting the complex function *u(l)* on the two-dimensional plane orthogonal to the time axis agrees with Fig. 3(b). Since the complex function *u(l)* is composed of a closed curve, Fig. 4 shows one cycle of the complex function *u(l)*.

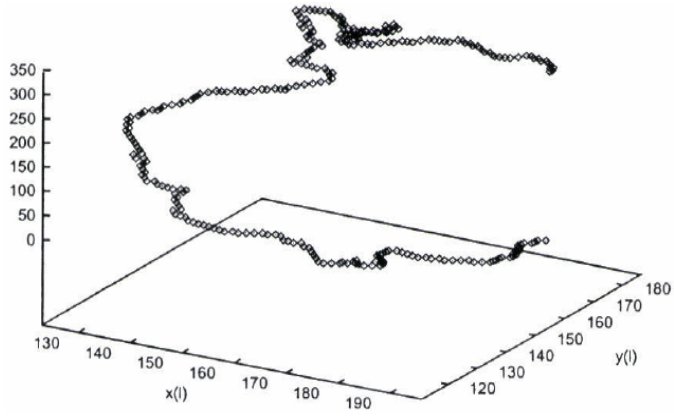
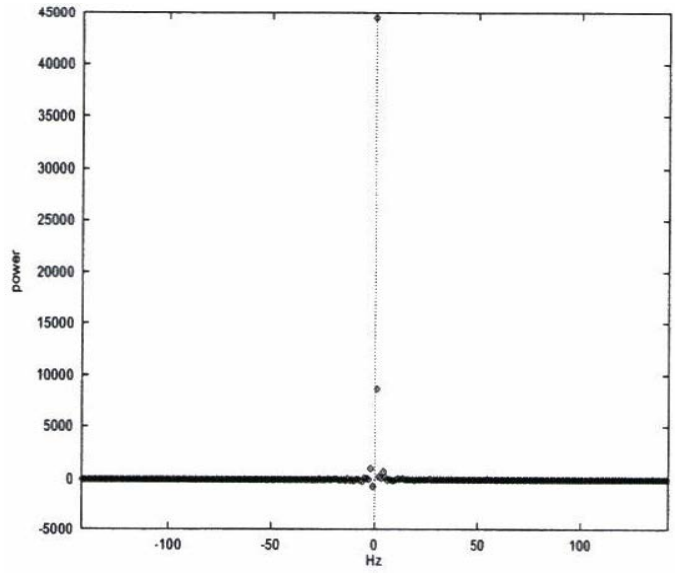


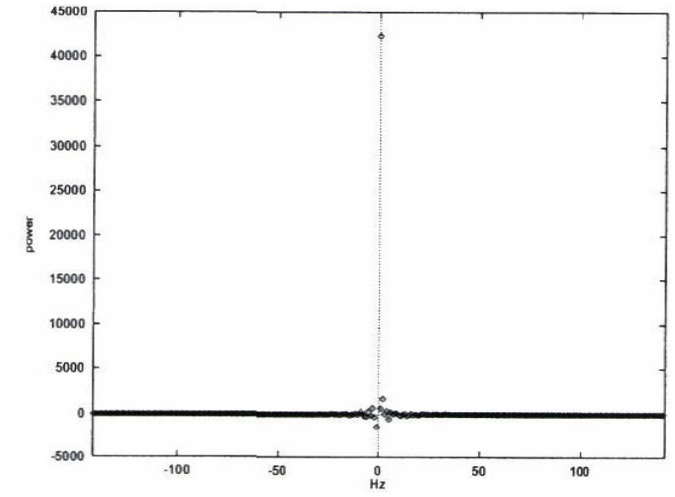
Fig. 4. The Complex Function u(l).

B. Experimental Result

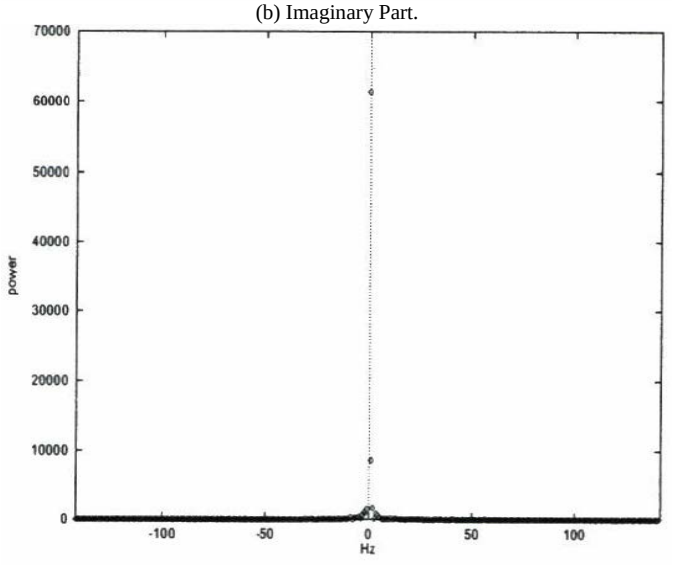
Fig. 5 shows the real and imaginary parts of Fourier spectrum and the power spectrum when the Fourier expansion is applied to the complex function *u(l)*. Fig. 6 corresponds to the result of the Gabor expansion coefficient when σ=5-40. Fig. 7 shows the number of non-zero Gabor expansion coefficients (support length) when the parameter σ is changed. Fig. 8 shows the restoration accuracy J_1 when the parameter σ is changed. Fig. 6 shows an example of a restored image when the parameters σ = 10 and 20. Table I shows the restoration accuracy J_1 when the parameter σ is fixed and the Gabor expansion coefficient with a large parameter ω is forcibly set to zero (the restoration accuracy J_1 is shown when the number of coefficients forcibly set to zero is changed.



(a) Real Part.



(b) Imaginary Part.



(c) Power Spectrum.

Fig. 5. Real and Imaginary Parts of the Fourier as well as Power Spectrum of *u(l)*.

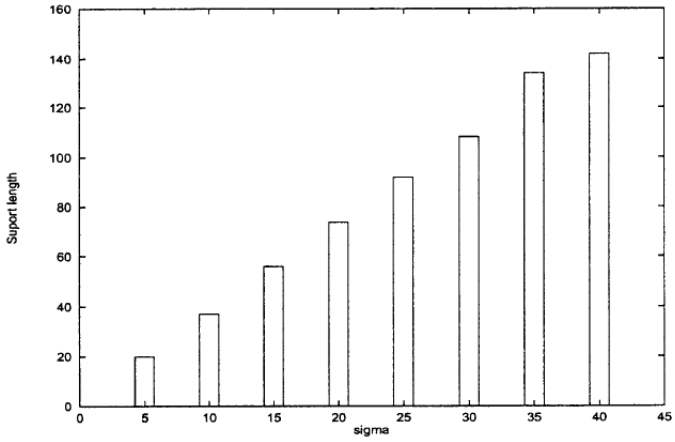


Fig. 6. Support Length for the Designated Parameter σ.

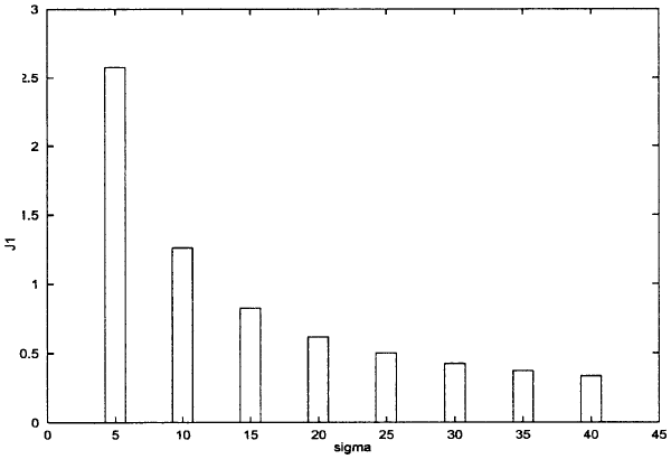
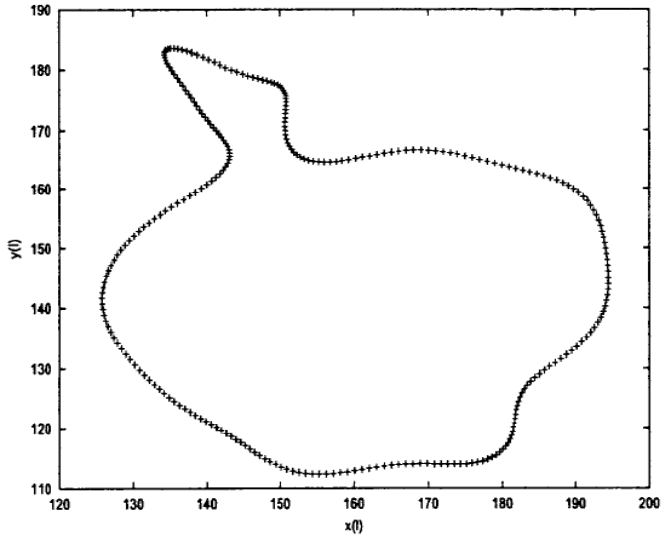
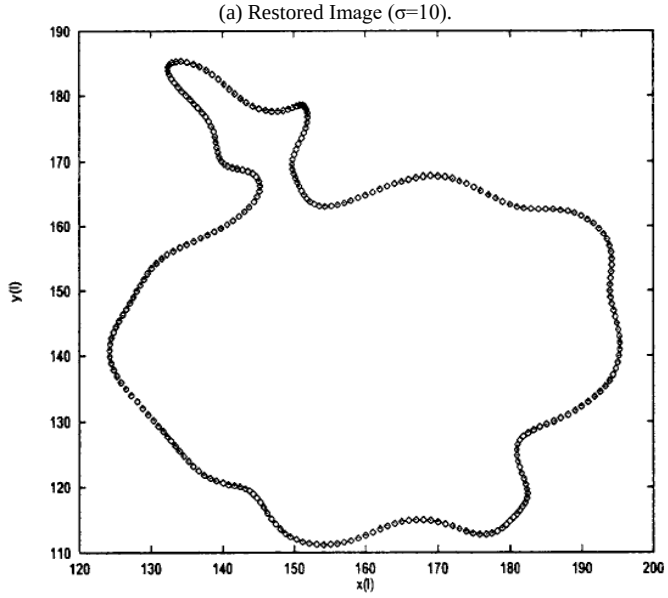


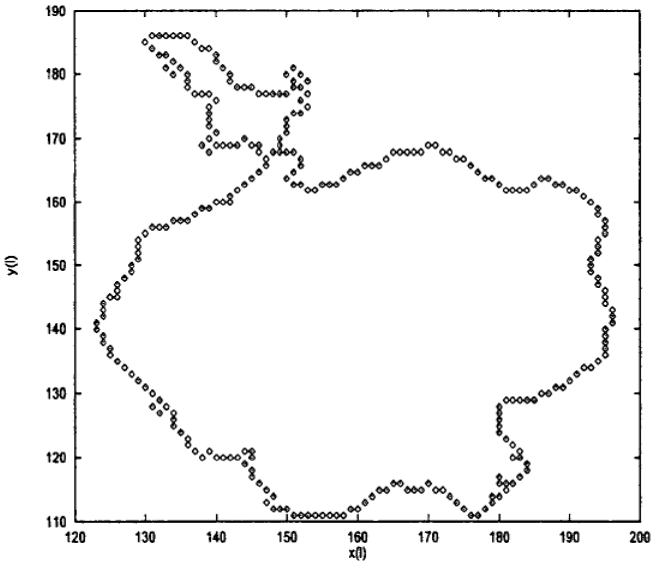
Fig. 7. Error J_1 for the Designated Parameter σ.



(a) Restored Image (σ=10).



(b) Restored Image (σ=20).



(c) True Contour (the Detected Edge).

Fig. 8. Example of Restored Image for σ=10, 20.

TABLE I. RESTORATION ACCURACY J1 BY THE NUMBER Γ OF COEFFICIENTS

σ|γ	20	40	60	100	160
10	1.374	1.267	1.266	1.266	1.266
20	1.08	0.656	0.622	0.618	0.618
30	1.056	0.531	0.453	0.427	0.424
40	1.051	0.5	0.401	0.348	0.335
50	1.05	0.491	0.382	0.312	0.282
80	1.05	0.485	0.369	0.28	0.216
100	1.049	0.484	0.368	0.276	0.195
150	1.049	0.483	0.367	0.273	0.183

That is, the accuracy J_1 is shown which is obtained by applying the low pass filter to the result of the Gabor transform of the parameter σ and restoring it. Table I shows the restoration accuracy J_1 by the number γ of coefficients that are not forced to zero. The parameter γ is a parameter in the frequency axis direction for the low-pass filter, and the parameter σ is a parameter in the time axis direction.

C. Remarks

As for the relationship between the restoration accuracy and the parameter σ, it can be seen from Fig. 4 that the support length depends on the parameter σ. That is, the support length increases as the parameter σ increases. From Fig. 7, the restoration accuracy J_1 depends on the parameter σ. That is, the restoration accuracy J_1 improves as the parameter σ increases. From Fig. 8, the restoration accuracy of the restored image depends on the support length.

On the other hand, regarding the relationship between the restoration accuracy and the parameter ω, from Table I, the restoration accuracy J_1, when the number of coefficients that force the Gabor expansion coefficient with a large parameter ω in each parameter σ to zero, is changed is zero. The smaller the number of coefficients to be set, the better.

IV. CONCLUSION

New spatial feature descriptor based on Gabor wavelet function is proposed. The proposed method is compared to Fourier descriptor. The experimental results with an ADEOS (Advanced Earth Observing Satellite) / AVNIR (Advanced Visible and Near-Infrared Radiometer) image show an effectiveness of the proposed method.

Through experiments with actual remote sensing satellite imagery data, it is found that the proposed description of geometric features of images based on the wavelet transform based on the Daubechies basis is more reproducible than that by existing Fourier descriptors under the condition of the same bandwidth. This is because, when the image space domain is transformed into the Fourier frequency domain and the wavelet frequency domain by the Fourier transform and the wavelet transform, respectively, if the band is halved to the maximum image frequency, the high frequency component disappears in the former case.

On the other hand, in the latter case, it can be said that the high frequency reproducibility is excellent because the high frequency component is preserved even when the latter is divided into the low frequency component. Moreover, it was found that this reproducibility depends on the support length of the basis function in the case of the wavelet descriptor, and when the support length is 2, it exceeds the Fourier descriptor in all cases of 4 or more. From the above, it can be said that the reproducibility of the geometrical description of the image by the wavelet descriptor under the same band limitation exceeds that of the Fourier descriptor.

V. FUTURE RESEARCH WORK

Further investigations are required for the validation of the proposed spatial feature extraction method with a variety of imagery data.

ACKNOWLEDGMENT

The author would like to thank Dr. Kaname Seto (former student of the Saga University) and Mr. Taroh Nakao (former student of the Saga University) for their contributions on this study. The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

[1] J.P. Djamdji, A. Bijaoui, and R. Maniere:•gGeo- metrical Registration of Images: The Multiresolution Approach,Photogrammetric Engineering & Remote Sensing, Vol. 59, No. 5, pp.645-653, (1993).

[2] W.M. Moon, J.S. Won, V. Singhroy, and P.D. Lowman Jr.:ERS-1 and CCRS C-SAR data integration for look-direction bias correction using wavelet transform, Canadian Journal of Remote Sensing, Vol. 20, No. 3, pp.280-285, (1994).

[3] M. Mak:•gOrthogonal Wavelet analysis: Interannual Variability in Sea Surface Tempera - ture, Bulletin of the American Meteorological Society, Vol. 76, No. 11, pp.2179-2186, (1995).

[4] R.W. Lindsay, D.B. Percival, and D.A. Roth rock,:The Discrete Wavelet Transform and the Scale Analysis of the Surface Properties of Sea Ice, IEEE Transactions on Geoscience and Remote Sensing, Vol. 34, No. 3, pp.771-787, (1996).

[5] Z. Hu, Y. Chen, and S. Islam: Multiscaling properties of soil moisture images and decomposi tion of large-and small-scale features using

wavelet transforms,•hInternational Journal of Remote Sensing, Vol. 19, No. 13, pp.2451-2467,(1998).

[6] Kohei Arai, Kaname Seto, L.M. Jameson: "Extraction of water mass features from satellite image by polar coordinate wavelet, Journal of Visual Information Society, Vol.19, Suppl.No.1, pp.99-102, (1999).

[7] Kohei Arai and Kaname Seto: "Change point extraction of multi-temporal earth observation satellite image based on wavelet decomposition and tiling," Journal of Japan Society for Visual Information, Vol.20, Suppl.No.1, pp.285-288, (2000).

[8] T. Ebrahimi, and M. Kunt:•gImage compres sion by Gabor expansion,•hOptical engineering, Vol. 30, No. 7, pp.873-880, (1991).

[9] H. Wang and H. Yan:•gAdaptive Gabor discrete cosine transforms for image compression, Electronics letters, Vol. 28, No. 18, pp.1755-1756, (1992).

[10] M.P. Anderson, M.H. Loew, and D.G. Brown: Gabor function-based medical image compres - sion,•hImage and vision computing, Vol. 13, No. 7, pp.535-541, (1995).

[11] R.A. Baxter : SAR image compression with the Gabor transform, IEEE Transactions on Geoscience and Remote Sensing, Vol. 37, No. 1, pp.574-588, (1999).

[12] A. Teuner, O. Pichler, and B.J. Hosticka: Unsupervised texture segmentation of images using tuned matched Gabor filters, IEEE Transactions on image processing, Vol. 4, No. 6, pp.863-870, (1995).

[13] D. Dunn, and W.E. Higgins :Optimal Gabor filters for texture segmentation, IEEE Transactions on image processing, Vol. 4, No. 7, pp.947- 964, (1995).

[14] P.P. Raghu, and B. Yegnanarayana: Segmen tation of Gabor-filtered Textures using deter ministic relaxation, IEEE Transactions on image processing, Vol. 5, No. 12, pp.1625-1636, (1996).

[15] R. Porter, and N. Canagarajah: Robust rotation-invariant texture classification: wavelet, Gabor filter and GMRF based schemes, TEE Proc. -Vis. Image Signal Process., Vol. 144, No. 3, pp.180-188, (1997).

[16] J. Yao: Complete Gabor transformation for signal representation, IEEE Transactions on image processing, Vol. 2, No. 2, pp.152-159, (1993).

[17] G. Cristobal, and R. Navarro: Space and frequency variant image enhancement based on a Gabor representation, Pattern recognition letters, Vo1.15. No. 3, pp.273-277, (1994).

[18] T.S. Lee: Image representation using 2D Gabor wavelets, IEEE Transactions on pattern analysis and machine intelligence, Vol. 18, No. 10, pp.959-971, (1996).

[19] R. Navarro, A. Tabernero, and G. Cristobal: Image representation with Gabor wavelets and its applications,•gAdvances in imaging and electron physics, Vol. 97, pp.1-84, (1996).

[20] R. Navarro, A. Tabernero, and G. Cristð®bal: Image representation with Gabor wavelets and its applications, Advances in imaging and elec- tron physics, Vol. 99, pp.329, (1998).

[21] X. Wu, and B. Bhanu: Gabor wavelet representation for 3-D object recognition, IEEE Transactions on image processing, Vol. 6, No. 1, pp.47-64, (1997).

[22] B.J. Super, and A.C. Bovik: Localized mea - surement of image fractal dimension using Gabor filters,•hJournal of visual communication and image representation, Vol. 2, No. 2, pp.114-128, (1991).

[23] J.G. Teti, Jr., and H.N. Kritikos: SAR ocean image decomposition using the Gabor expansion, IEEE Transactions on Geoscience and Remote Sensing, Vol. 30, No. 1, pp.192-196, (1992).

[24] F. Smeraldi, O. Carmona, and J. Bigun: Saccadic search with Gabor features applied to eye detection and real-time head tracking, Image and vision computing, Vol. 18, pp.323-329, (2000).

[25] D.M. Weber, and D.P. Casasent: Quadratic Gabor filters for object detection, IEEE Transactions on image processing, Vo1.10. No. 2, pp.218-230, (2001).

[26] P.P. Raghu, and B. Yegnanarayana: Multispectral image classification using Gabor filters and stochastic relaxation neural network, Neural Networks, Vol. 10, No. 3, pp.561-572, (1997).

[27] L. Wang, C.T. Chen, and W.C. Lin:•gAn ecient algorithm to compute the complete set of discrete Gabor coecients, IEEE Transactions on image processing, Vol. 3, No. 1, pp.87-92, (1994).

[28] J. Yao, P. Krolak, and C. Steele: The generalized Gabor transform,•h IEEE Transactions on image processing, Vol. 4, No. 7, pp.978-988, (1995).

[29] C.T. Zahn and P.Z. Roskies : Fourier descriptors for plane closed curves, IEEE Trans. Computer, Vol. C-21, pp.269-281, (1972).

[30] Takafumi Miyatake, Takashi Matsuyama, Makoto Nagao: "Recognition of curves invariant to affine transformation using Fourier descriptor," Transactions of Information Processing Society of Japan, Vol.24, No.1, pp.64-71, (1983).

[31] Yoshinori Uesaka: "A new Fourier descriptor applicable to open curves," IEICE Transactions, Vo1.J67-A, No.3, pp.166-173, (1984).

[32] Kohei Arai, Kaname Seto and Taro Nakao: "Description of geometrical features of images by wavelet descriptor based on Gabor transform," Proc. of the 30th Annual Meeting of the Remote Sensing Society of Japan, (2001).

[33] Kohei Arai: "Fundamental Theory of Wavelet Analysis" Morikita Publishing, Nov., (2000).

[34] Kohei Arai, Leland Jameson: "How to use earth observation satellite data by wavelet," Morikita Publishing, July, (2001).

[35] Kohei Arai, Remote sensing satellite image database system allowing image portion retrievals utilizing principal component which consists spectral and spatial features extracted from imagery data, International Journal of Advanced Research in Artificial Intelligence, 2, 5, 32038, 2013.

AUTHORS' PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html.

Comparative Analysis of National Cyber Security Strategies using Topic Modelling

Minkyoung Song, Dong Hee Kim, Sunha Bae, So-Jeong Kim
National Security Research Institute, Daejeon, South Korea

***Abstract*—Comprehensive comparative analyses of national cyber security strategies (NCSSs) have thus far been limited or complicated by the unique nature of cybersecurity, which combines various areas such as technology, industry, economy, and defense in a complex manner. This study aims to characterize the NCSSs of major countries, quantitatively considering the time series, and identify further cybersecurity agendas for the benefit of NCSS revision in South Korea, by applying topic modelling to the analysis of eight NCSSs from the US, UK, Japan, and EU. As a result, fifteen agendas were identified and grouped into four sectors. We determined from the agenda distribution that the approach of each country to cybersecurity was different. In addition, additional agendas worthy of consideration for future NCSS revisions in South Korea were proposed, based on a comparison of the 15 aforementioned agendas with those of South Korea. This study is significant for cybersecurity policy in terms of enabling quantitative analysis in a single framework via latent dirichlet allocation (LDA) topic modelling, and deriving further cybersecurity agendas for future NCSS revisions in South Korea.**

Keywords*—*Cybersecurity policy; national cyber security strategy (NCSS); policy analysis; quantitative analysis

I. Introduction

Even in the midst of the COVID-19 pandemic, many of us were able to maintain our daily lives and national activities through the use of cyberspace. Cyberspace has become a new existential dimension for individuals and society to access the world via the transcendence of the physical limits of time and space. However, as dependence on cyberspace increased through digitalisation across the world, cyber threats have also become more diverse, complex, and menacing. Furthermore, malicious cyber activities on critical infrastructure such as electrical utilities, banking systems, and telecommunications networks could threaten the national security of almost any country. Therefore, at least 100 countries have established national strategies to secure their cyberspace.

A national cybersecurity strategy (NCSS) is one of the most concise documents for understanding the national approach to securing cyberspace. Analysing NCSSs is essential for determining the response stances at a national level and for understanding international cybersecurity trends. However, NCSSs endogenously include multidimensional agendas such as technology, industry, economy, and defence, which make it difficult to perform consistent and systematic analysis of NCSSs and to discover which agenda should be addressed in the development or revision of an NCSS. Meanwhile, text analysis has emerged as a new method for the analysis of large amounts of descriptive data, such as in NCSSs. Text analysis is useful for enabling empirical and quantitative analysis of descriptive data by identifying the keywords of documents and understanding content from relationships between words. Recently, even though text analysis has begun to be introduced into the analysis of NCSSs, there is still room for improvement, such as through the inclusion of time series in the analysis, which can lead to further agendas for future NCSS development.

This study aims to determine the cybersecurity agendas of leading countries and derive their implications using topic modelling, a text analysis technique, to prepare their NCSSs for development and/or revision. To accomplish this objective, we focused on the eight NCSSs of the US, UK, Japan, and EU, who have attempted to assert their leadership in cyberspace by establishing their NCSSs earlier, and by constantly improving their NCSSs in consideration of the changing threat environment. Furthermore, the aforementioned four countries are suitable for inspiring the design of a domestic cybersecurity agenda for South Korea because these countries have similar cybersecurity approaches and similar or higher technology levels with respect to South Korea.

The remainder of this paper is organised as follows. The first part consists of a literature review on topic modelling and NCSS analysis. The second part presents an analysis of target NCSS documents. The third part provides information on 15 cybersecurity agendas and their trends by nation and period as a result of topic modelling analysis, and the last part reveals the conclusion and future direction of research.

II. Literature Review

A. Topic Modelling

Unstructured data such as NCSSs have been analysed mainly using qualitative methods rather than quantitative methods because of the nonlinear relationship between cause and effect, the importance of historical reasons, and path-dependent development. However, as natural language processing (NLP) techniques are applied to existing data mining processes, empirical and quantitative analyses of unstructured text data increasingly gain attention in the field of policy analysis. In its early stages, this type of text analysis was used primarily for library and information science and computer engineering, whereas nowadays, it is used for a greater variety of purposes as part of quantitative content analysis.

Topic modelling is a statistical technique used to discover hidden structures from collections of documents. In policy research, topic modelling has been used to discover policy

agendas or issues from news articles, speeches, and petitions, and to monitor research trends through the analysis of research papers in time series. Furthermore, research applying topic modelling to policy evaluation has recently emerged. Table I shows a list of these studies with descriptions of the methodology.

TABLE I. A LIST OF POLICY ANALYSIS STUDIES APPLYING TOPIC MODELLING

Research Area	Study	Year	Purpose	Dataset	Algorithm
Policy Agenda Analysis	[1]	2013	Monitoring public opinion	all posts by top 2,000 'LiveJournal' (blog platform in Russia) users (2011–2012)	LDA
	[2]	2015	Understanding citizens' direct policy suggestions	2,850 petition texts (2011–2014)	LDA
	[3]	2017	Analysing political agenda of European Parliament	English language legislative speeches in European Parliament plenary (1999–2014)	NMF
	[4]	2019	Policy requirement at citizens' level	173 posts in social media	LSA
	[5]	2019	Identifying relation between mass media and public attention	news articles, Google trends query, Twitter keywords (Jul. 31st to Nov. 5th, 2017)	NMF
Research Trend Analysis	[6]	2017	Discovering themes and trends in transportation research	17,163 papers (1990–2015)	LDA
	[7]	2018	Exploring research trend of smart factory	2,488 international papers and 404 Korean papers (1995–2016)	LSA
	[8]	2019	Analysing research topics in cybersecurity and data science	48 papers (2012–2018)	LDA
Impact Analysis	[9]	2021	Assessing temporal patterns of newspaper coverage	6,645 articles on German Renewable Energy Act (2000–2017)	LDA

As shown in Table I, researchers could determine policy implications by selecting appropriate datasets and algorithms according to their research purposes and interpreting the topic modelling results. Specifically, Table I shows that policy agendas could be discovered from SNS postings, petitions, speeches, articles, etc., research trends from research papers, and policy impact from the contents of articles on specific issues. On the other hand, algorithm selection does not depend on the research purpose or area. Some algorithms for categorising topics from words in documents include latent semantic analysis (LSA), non-negative matrix factorisation (NMF), and latent dirichlet allocation (LDA), among which LDA is the most widely used for topic modelling in social science. This is because LDA assumes that multiple topics exist in a single document, which is in harmony with the social science assumption that a single body of text does not reflect only a single point of view, but that multiple competing points of view can appear within the same document. Therefore, this study attempted topic modelling using LDA, the algorithm that is most widely used for policy analysis.

B. National Cybersecurity Strategy

A national cybersecurity strategy (NCSS) is a document that reflects cybersecurity policy direction and stance on cyber threats at the national level. Because the NCSS sets national strategic objectives and priorities for a specific period, it is essential to consider the evolving cyber threat environment and the national approach to cybersecurity in a timely manner. For example, Japan has a cybersecurity policy structure that is revised every three years, and the EU every seven years. However, because of rapid technological changes and the short technological life cycle of information and communications technology (ICT), NCSS revision cycles need to be shorter in the future. For nations that want to properly establish or revise their NCSSs, analysis of the NCSSs of countries that have leadership in cyberspace or similar approaches to cybersecurity is important. This strategy will help with identifying new policy agendas that have not yet been considered and with uncovering any issues that may require cybersecurity cooperation.

Studies on NCSSs have usually aimed to discover common structures or identify further agendas that need to be considered. However, prior to the application of data analysis such as topic modelling to the cybersecurity policy area, qualitative methodologies, which forced reliance on the opinions of experts, were used in the analysis of NCSSs. Qualitative analysis not only consumes large amounts of time, but also is prone to inconsistencies because of the likelihood of differing opinions among these experts. Therefore, it is necessary to establish an automated quantitative analysis system to work alongside qualitative analysis. NCSS analysis using topic modelling has thus far focused only on analysing more countries and more data, which is unsuitable for studies aiming to discover cybersecurity agendas for the establishment or revision of NCSSs. Of course, it is important to examine global cybersecurity trends practically and academically, but in any research for the purpose of establishing or revising an NCSS, it is necessary to limit the scope of analysis to NCSSs in like-minded countries or in advanced countries. Table II outlines prior studies on NCSS analysis.

TABLE II. A LIST OF NCSSS ANALYSIS STUDIES USING TOPIC MODELLING METHOD

Study	Year	NCSSs	Methodology	Description
[10]	2013	19 NCSSs	(Qualitative) comparison based on 11 categories	Identifying formal structures for NCSS development
[11]	2015	3 NCSSs	(Qualitative) comparison based on 7 categories	Finding NCSSs, in general, changed from voluntary self-regulation to enforced self-regulation
[12]	2016	10 NCSSs	(Qualitative) content analysis	Finding 8 main components of NCSSs
[13]	2019	6 NCSSs	(Qualitative) cross-section analysis using 8 comparison elements	Evaluating robustness of existing cyber security strategy of Bangladesh
[14]	2017	60 NCSSs	(Quantitative) clustering and topic modelling	(Initial attempt to compare NCSSs using topic modelling method) Identifying 10 topics in NCSSs
[15]	2020	101 NCSSs	(Quantitative) topic modelling	Identifying 4 critical agendas in NCSSs

Topic modelling has contributed to the understanding of international trends in cybersecurity by enabling massive data analysis and extending NCSS analysis to quantitative and empirical areas. However, the limitations of not considering the time series and the scope of analysis remain, rendering these past analyses insufficient for deriving policy implications for future NCSSs. Therefore, this study focuses on characterising the NCSSs of the US, UK, Japan, and EU, tracking changes in their topic distributions over time, and then identifying critical national cybersecurity agendas through comprehensive comparative analysis of the results of topic modelling.

III. DATASET

South Korea launched its first national cybersecurity strategy in 2019. Although this strategy is not the first official document to reveal the response stance of South Korea to cyber threats, it is the first cybersecurity strategy document established in accordance with the national security strategy. This strategy contains six strategic tasks, the titles of which are listed in detail in Table III.

To implement the NCSS, South Korea has announced an action plan at the agency level to support these six strategic tasks until 2022. This suggests that the policy demand for NCSS revision would increase, such as in identifying additional policy agendas worthy of consideration but not covered by existing NCSSs. Therefore, this study selected the NCSSs of the US, UK, Japan, and EU for comparative analysis to derive additional considerations for revising the NCSS of South Korea. Two criteria were considered in the selection of the target of analysis. For the first criterion, the target must have similar approaches to cyberspace in terms of international relationships, while also having similar ICT research and development level, to that of South Korea. For the second criterion, the target should have published an NCSS more than once, such that its NCSS transition in time series can be tracked. This target selection enables a direct comparative analysis of cybersecurity agendas derived from topic modelling results and the strategic tasks of the South Korean NCSS; it is also suitable for examining NCSS trends by country and period, which have not been provided by prior studies that used topic modelling. The dataset for this study is presented in Table IV. Prior to analysis, the aforementioned eight NCSSs were subjected to pre-processing: synonyms were extracted into single words, and unnecessary words with general meaning were eliminated. As a result, 1,287 words remained for the actual LDA topic modelling analysis.

TABLE III. STRATEGIC TASKS PRESENTED IN NCSS OF SOUTH KOREA

1. Increase Safety of National Core Infrastructure	
1-1	Strengthen security of national information and communications networks
1-2	Improve cybersecurity environment for critical infrastructure
1-3	Develop next-generation cybersecurity infrastructure
2. Enhance Cyber Attack Response Capabilities	
2-1	Ensure cyber attack deterrence
2-2	Strengthen readiness against massive cyber attacks
2-3	Devise comprehensive and active countermeasures for cyber attacks
2-4	Enhance cybercrime response capabilities
3. Establish Governance Based on Trust and Cooperation	
3-1	Facilitate public–private–military cooperation system
3-2	Build and facilitate nation-wide information sharing system
3-3	Strengthen legal basis for cybersecurity
4. Build Foundations for Cybersecurity Industry Growth	
4-1	Expand cybersecurity investment
4-2	Strengthen competitiveness of cybersecurity workforce and technology
4-3	Foster growth environment for cybersecurity companies
4-4	Establish principle of fair competition in cybersecurity market
5. Foster Cybersecurity Culture	
5-1	Raise cybersecurity awareness and strengthen cybersecurity practice
5-2	Balance fundamental rights with cybersecurity
6. Lead International Cooperation in Cybersecurity	
6-1	Enrich bilateral and multilateral cooperation systems
6-2	Secure leadership in international cooperation

TABLE IV. ANALYSIS TARGET DOCUMENTS

Nation	Year	Document (NCSS)	Version
U.S.	2003	National Strategy to Secure Cyberspace	Previous
	2018	National Cyber Strategy	Current
U.K.	2011	The UK Cyber Security Strategy	Previous
	2016	National Cyber Security Strategy 2016–2021	Current
Japan	2015	Cybersecurity Strategy	Previous
	2018	Cybersecurity Strategy	Current
EU	2013	EU Cybersecurity Strategy: An Open, Safe and Secure Cyberspace	Previous
	2020	EU Cybersecurity Strategy for the Digital Decade	Current

IV. Data Analysis and Result

A. Result of Topic Modelling

A total of 15 agendas were identified, as shown in Table V. Although there are a few tools available for determining topic sets, such as the minimum perplexity approach, the suitable approach is not yet clear. Therefore, we designated an optimal number of topics by reviewing the keywords constituting each agenda in a way that minimised duplication, and maximising the explanatory power of the agendas from a holistic point of view.

Table V consists of the columns Sector, Agenda, Keywords, and Proportion. The naming of each agenda is based on its constituent keywords. Furthermore, the agendas are grouped into four sectors: Infra Stability (I), Protection and Response Capability (II), Industry and Technology (III), and International Cooperation (IV), in accordance to their strategic or operational objectives. Lastly, the rightmost column of Table V presents the proportion of each agenda.

TABLE V. Fifteen Cybersecurity Agendas of US, UK, Japan and EU and their Proportion

Sector	Agenda	Keyword (Top 15)	Prop.
Infra Stability (I)	① Network and System Vulnerability	**cyber-attack**, **network**, **system**, **vulnerability**, **software**, computer, internet, actor, **damage**, attacker, critical infrastructure, **malware**, **hardware**, attention, **disruption**	7.93
	② Cyber Security Role and Responsibility	**security**, **agency**, **system**, **responsibility**, cyber space, **role**, **risk**, state, investment, **control**, IT, procurement, **administration**, **asset**, effectiveness	5.86
	③ Risk Assessment and Management	**risk**, **cyber threat**, **vulnerability**, cyber-attack, **critical infrastructure**, **assessment**, **priority**, operation, challenge, company, nation, damage, **risk management**, **resource**, opportunity	7.13
	④ Information Communication Network Access Control	**system**, **information**, **network**, **security**, infrastructure, **communication**, **access**, **control**, **information system**, computer, AI, internet, trustworthiness, knowledge, integrity	6.29
Protection and Response Capability (II)	① Privacy and Intellectual Property Security	**internet**, information, **right**, **freedom**, **citizen**, **privacy**, **protection**, **security**, business environment, **society**, **DNS**, **online**, **intellectual property**, **breach**, **human right**	7.04
	② Cyber Defence Capability	**capability**, **cyber-attack**, **defence**, **cyber threat**, **national security**, nation, **critical infrastructure**, **state**, actor, **cyber terrorism**, **adversary**, network, infrastructure, **ability**, **operation**	7.85
	③ Incident Response and Information Sharing	**cyber-attack**, **incident**, **response**, information, **cyber threat**, **capability**, **coordination**, **information sharing**, damage, sharing, **detection**, **recovery**, knowledge, **situational awareness**, **monitoring**	8.40
	④ Cyber Crime Law Enforcement and Investigation	**cyber-crime**, **law**, **enforcement**, **capability**, agency, cyber threat, **intelligence**, **response**, **investigation**, **authority**, tool, force, child **protection**, resource, **capacity**	8.61
Industry and Technology (III)	① Standard, Certification and Supply Chain Security	**system**, **security**, **operation**, **IoT**, **business environment**, **critical infrastructure**, **standard**, information, **industry**, **assurance**, safety, **certification**, **supply chain**, connection, collaboration	6.75
	② ICT Innovation	**information**, **internet**, **society**, **security**, **economy**, infrastructure, **progress**, **innovation**, **multi-stakeholder**, governance, **ICT**, **market place**, communication, country, culture	4.22
	③ Public Private Partnership (PPP)	**industry**, cyber awareness, **research**, **R&D**, security, **coordination**, **standard**, public, role, **partnership**, collaboration, company, information, state, **innovation**	4.61
	④ Security Awareness and Knowledge	**business environment**, **market place**, **company**, **investment**, **personnel**, **cyber awareness**, risk, cost, judiciary, solution, opportunity, **knowledge**, role, **human resource**, demand	7.90
International Cooperation (IV)	① International Norm and State Behaviour	**state**, **rule, behaviour**, **principle**, **national security**, **peace**, **law**, **norm**, **stability**, **international community**, **international law,** society, safety, actor, **alliance**	8.18
	② EU Member State Cooperation	**member**, **state**, **cooperation**, **defence**, authority, progress, agency, **NIS directive**, **ENISA**, **coordination**, **incident**, **resilience**, role, framework, capability	3.99
	③ International Partnership	**country**, **partner**, **cooperation**, cyber threat, **partnership**, **industry**, challenge, capability, information, **ally**, border, network, communication, **participant**, NATO	5.26

B. NCSS Agenda Transition

1) Topic distribution by nation: The results of topic distribution for the four nations are presented in Table V. This section aims to identify the differences in cybersecurity approaches by nation. In Fig. 1, which was derived from the current NCSS of each country that was analysed, the blue bar represents the percentage of each agenda, whereas the orange line represents the percentage of each sector, which is the sum of the percentages of agendas constituting that sector (I–IV). According to the results, the NCSSs of the US and UK focused on improving cybersecurity response capability, whereas those of Japan and the EU vitalised the cybersecurity industry and international cooperation, respectively.

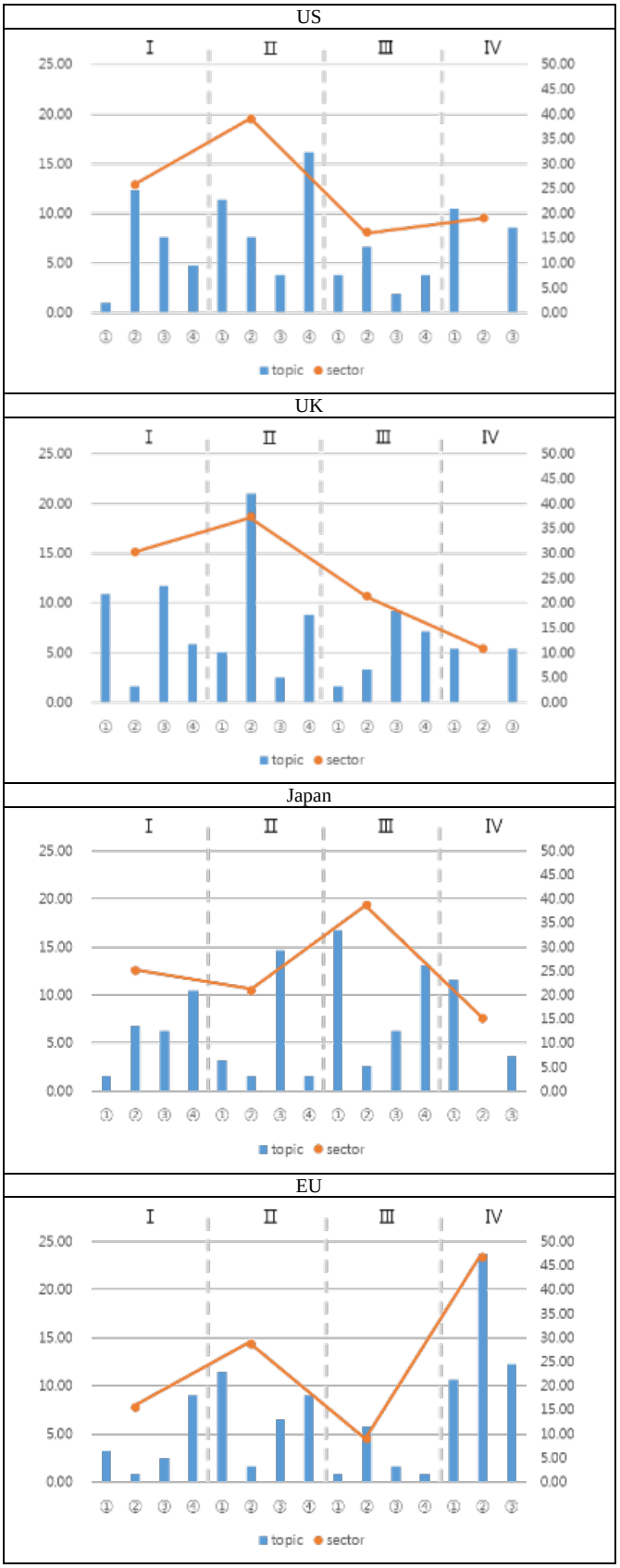


Fig. 1. Topic and Sector Distribution Derived from the Current NCSS of each Country.

The NCSS of the US (2018) emphasised improving incident response capabilities (II), especially cybercrime law enforcement and investigation capabilities (II-④), and establishing cybersecurity governance (I-②). In addition, intellectual property security (II-①) had a relatively higher proportion compared with in other NCSSs, which means that the US, with its world-class technology levels in a wide range of emerging technology areas such as IT, aerospace, and defence industries, likely regard the technological and economic aspects of its cyberspace from a national security perspective.

Similar to the US, the UK (2016) prioritised protection and response capability (II). However, unlike the US, the UK highly concentrated on cyber defence capability (II-②). This may reflect the concept of traditional defence power in cyberspace, and is consistent with the creation of a National Cyber Force, which is known to provide offensive and defensive capabilities in pursuit of national security objectives, and operation of an Active Cyber Defence (ACD) programme, which is meant to reduce harm from commodity cyber-attacks by providing necessary tools and services. Moreover, the UK was also shown to be discussing both vulnerability mitigation (I -①) and risk assessment and management (I -③) to improve infrastructure stability. A risk-centric approach to cybersecurity could be the basis for establishing concretised security measures according to asset or information-specific importance and the level of risk exposed; therefore; the NCSSs of the US, Japan, and UK covered this type of approach at high proportions.

On the other hand, the NCSS of Japan (2018) was characterised by a relatively high proportion (39%) for the cybersecurity industry and technology sector (III). This observation is consistent with the objective of its strategy; the first objective of Japan, unlike in the other analysed countries, was to enable socio-economic vitality and sustainable development. Their suggested policy approach to achieving this objective was to advance cybersecurity, establish a secure supply chain, and build a secure IoT system. This approach to cybersecurity was clearly different from those of the other analysed countries, which prioritised the protection of critical infrastructure and enhancement of deterrence in cyberspace. In addition, their National Information Security Center (NISC), which is responsible for information security policy, announced the necessity to protect the supply chain against dependence on excessive foreign technologies, to drive data accumulation and utilisation using emerging technologies such as AI, and to accomplish international standardisation of related technologies. Based on a comprehensive view of these considerations, the focal point of the cybersecurity policy of Japan seemed to be the revitalisation of future technological industry and economy.

Finally, the topic modelling results characterised the NCSS of the EU as emphasising international cooperation. In particular, the EU sought to improve levels of cyber resiliency and consistency across Europe through cooperative responses in cyberspace based on the NIS Directive, as shown by the word composition of the topic regarding EU member state cooperation. In particular, according to the contents of the EU

cybersecurity strategy, the EU would strengthen the interoperability of information systems, establish a security operations centre (SOC) network, and expand the use of the EU Cyber Diplomacy Toolbox to achieve their objectives of improving the level of resiliency and consistency in cyberspace. Therefore, the strategy of the EU would have been devised based on a very high proportion of international cooperation.

2) Topic distribution by period: The topic distributions of the analysed NCSSs are shown in Table VI.

As presented in Table VI, there were no significant increases or decreases in the NCSS agenda transitions in time series. This observation indicates that in setting their cybersecurity agendas, each country considers its own threat environment and its geopolitical characteristics rather than the agenda trends at that time. In other words, in the establishment or revision of an NCSS, an understanding of the threat environment facing the country should be obtained, and a clear analysis of their own approach to solving it should be conducted.

The agenda that exhibited the biggest distribution gap in the US strategies was network and system vulnerability. Although the previous strategy of the US prioritised this agenda (which had the highest proportion, at almost 20%), that proportion was significantly reduced to less than 1% in the current strategy. This observation could reflect a change in response posture against cyber-attacks, crimes, or even threats, from passive protection that mitigates critical vulnerabilities in their own network or system, to an active response posture that includes law enforcement in anti-cybercrime efforts and cooperative responses with like-minded countries.

These trends could also be observed in the UK, which addressed strengthening defence and response capabilities in both their 2011 and 2016 strategies. In particular, in their current strategy, the weight of the agenda on cyber defence capability has increased (+12.68%), suggesting that defensive and even offensive operations could be conducted based on an understanding of cyberspace as a military domain.

Meanwhile, the NCSSs of Japan had the smallest change in topic distribution over time, because the three-year NCSS establishment cycle of Japan is not only short compared to those of other countries but also established with the basic act on cybersecurity as a legal basis. On the other hand, because Japan is constantly emphasising the revitalisation of the cybersecurity industry, it is necessary for them to continuously grasp related standards and supply chain security trends in the future.

The EU also had a small change in their distribution of cybersecurity agendas by period. However, a noticeable difference was that the proportion of EU member state cooperation slightly decreased, whereas the proportion of agenda on international partnership somewhat increased. For context, Europe has recently continued to discuss European capability building from the security and defence standpoint and the 'strategic autonomy' based on it. As the need to work together with international partners to achieve these goals is emphasised, it is necessary to observe how Europe will strengthen its international partnerships to secure strategic autonomy in the future.

TABLE VI. TOPIC DISTRIBUTION IN EACH NCSS DOCUMENT

Topic (Agenda)	Topic distribution (%)							
	US 2003	UK 2011	EU 2013	JP 2015	UK 2016	JP 2018	US 2018	EU 2020
Network and System Vulnerability	19.42	5.30	1.15	4.04	10.92	1.57	0.95	3.28
Cyber Security Role and Responsibility	13.04	0.00	1.15	4.04	1.68	6.81	12.38	0.82
Risk Assessment and Management	12.17	3.03	1.15	1.79	11.76	6.28	7.62	2.46
Information Communication Network Access Control	8.70	0.76	2.30	2.69	5.88	10.47	4.76	9.02
Privacy and Intellectual Property Security	7.54	12.12	10.34	3.14	5.04	3.14	11.43	11.48
Cyber Defence Capability	8.99	8.33	1.15	2.24	21.01	1.57	7.62	1.64
Incident Response and Information Sharing	6.96	2.27	11.49	17.94	2.52	14.66	3.81	6.56
Cyber Crime Law Enforcement and Investigation	5.22	29.55	13.79	1.79	8.82	1.57	16.19	9.02
Standard, Certification and Supply Chain Security	2.90	2.27	2.30	18.83	1.68	16.75	3.81	0.82
ICT Innovation	2.61	5.30	4.60	6.28	3.36	2.62	6.67	5.74
Public Private Partnership(PPP)	3.48	3.03	1.15	4.48	9.24	6.28	1.90	1.64
Security Awareness and Knowledge	6.38	14.39	5.75	9.87	7.14	13.09	3.81	0.82
International Norm and State Behaviour	0.87	10.61	9.20	14.80	5.46	11.52	10.48	10.66
EU Member State Cooperation	0.00	0.76	31.03	0.00	0.00	0.00	0.00	23.77
International Partnership	1.74	2.27	3.45	8.07	5.46	3.66	8.57	12.30

C. Comparative Analysis of NCSS Agendas with Coverage in South Korea

As discussed earlier, NCSS agendas had different distributions depending on the threat environment and approach to cybersecurity of each country. However, the US, UK, Japan, and EU have close cooperation in cybersecurity and in related technologies and research areas, and thus identifying the cybersecurity agendas of these countries is essential for future cooperation or diplomacy. Furthermore, analysing the agendas of like-minded countries is valuable as a method for determining suitable NCSS agendas for a given country because it provides an understanding of global cybersecurity trends in the context of cooperative response. Therefore, this section identifies agendas worthy of consideration for future NCSS revisions in South Korea by comparing the strategic task of the current NCSS with the 15 agendas previously derived.

Table VII is the result of comparing the 15 agendas derived from this analysis with the contents of the NCSS of South Korea. This analysis reveals two agendas that were not covered (marked with ×) and one agenda partially covered (marked with △) in the NCSS of South Korea.

First, one agenda on risk assessment and management in sector I was not covered in the NCSS of South Korea. Because cyber threats tend to be increasingly diverse and sophisticated, a single way of managing security vulnerabilities in systems or networks may not be sufficient for preventing cyberattacks that use social engineering techniques. However, the current NCSS of South Korea has been focused on vulnerability management in the infra stability sector, and not on risk assessment and management. Here, cyber risk is a concept that considers not only vulnerabilities in the system itself but also the possibility of manipulation, disruption, or destruction of specific assets [16].

Moreover, cyber risk management refers to a series of actions that identify the value and importance of individual assets, evaluate the impact of vulnerabilities or risks of exploiting them, and prepare and implement appropriate countermeasures for the assessed risk. Therefore, efforts should be made to ensure the stability of critical infrastructure in a dynamic cyber threat environment through the establishment of a framework for assessing and managing risk to critical assets in addition to vulnerability management [17].

Furthermore, the NCSS of South Korea has no discussion on the protection of intellectual property rights. Competitiveness in science and technology is becoming more important in both cybersecurity and economic aspects compared to in the traditional security perspective. Whereas many countries are making great efforts to secure technological competitiveness, the number of malicious cyber activities targeting the intellectual property (IP) of research institutes or universities has been increasing. Accordingly, countries with high levels of technology, such as the US and UK, are implementing strict measures against such technology theft to maintain their technological and economic superiority [18]. In particular, the US government is using name-and-shame processes, such as public indictments on IP theft, to inform countries about these malicious activities and continue efforts to strengthen relevant law enforcement capabilities. For future NCSS revisions in South Korea, there is a necessity for multilateral discussions to protect future cybersecurity R&D achievements through close cooperation between science, technology, and industry, to secure the technological advantage of the country.

Finally, a discussion on supply chain security is necessary. The supply chain refers to the overall system of organisations, resources, human resources, and information in the process of providing products or services to customers. The supply chain is particularly complex for ICT products and services, and includes processes of S/W and H/W design, deployment, acquisition, operation, and maintenance. Supply chain security issues, which began to be discussed in earnest after the US sanctions against Huawei, are currently being embodied in policies for developing supply chain risk assessment tools or systems, and diversifying or internalising 5G suppliers [19]. However, in the case of the NCSS of South Korea, discussions on overall supply chain risk management, including all ICT products and services such as 5G, IoT devices, and cloud services, are limited, and are covered only through standards and certification systems and the security-by-design concept. Therefore, it is necessary to establish and realise a supply chain security system across the country to analyse supply chain risk and prepare for global supply chain reorganisation under US–China trade tension.

TABLE VII. THE RESULT OF COMPARING THE 15 AGENDAS WITH THE CONTENTS OF NCSS OF SOUTH KOREA

Sector	Agendas	Comparison Result	
		(○/△/×)	Related tasks # of Table III
Infra Stability	Network and System Vulnerability	○	1-1,2
	Cyber Security Role and Responsibility	○	3-1
	Risk Assessment and Management	×	-
	Information Communication Network Access Control	○	2-2
Protection and Response Capability	Privacy and Intellectual Property Security	×	-
	Cyber Defence Capability	○	2-1,3
	Incident Response and Information Sharing	○	2-2, 3-2
	Cyber Crime Law Enforcement and Investigation	○	2-4
Industry and Technology	Standard, Certification and Supply Chain Security	△	1-3
	ICT Innovation	○	4-3
	Public Private Partnership (PPP)	○	3-1, 4-1,3
	Security Awareness and Knowledge	○	4-1,2
International Cooperation	International Norm and State Behaviour	○	6-2
	EU Member State Cooperation	○	6-1
	International Partnership	○	6-1

V. Conclusion

Cybersecurity has more complex and multidimensional characteristics compared to those of traditional security, and involves a combination of hyper-connected cyberspace, rapid development of ICT, and double-use issues of cyber technology. In addition, differences in the approaches to cyber space and cyber threat environments in different countries contribute to further increasing this complexity, which in turn complicate the macroscopic perspective analysis of national cybersecurity policies. Therefore, this study aimed to derive the national cybersecurity policy agendas of major countries from a macro perspective by using the topic modelling method.

The study was divided into two parts. The first part was to use a topic modelling method to identify national cybersecurity policy agendas in major countries, and the second part was to determine policy agendas that could be further considered for future NCSS revisions in South Korea. Thus far, policy research in the field of cybersecurity with the use of topic modelling has focused on expanding the scope of analysis to observe the global cybersecurity landscape. Therefore, this study is meaningful in that it used topic modelling to explore critical agendas and quantitatively compare the focal points of various NCSSs for the benefit of future NCSS revisions in South Korea.

As a result of this study, 15 agendas were derived from words that compose the NCSSs of the US, UK, Japan, and EU. These agendas were grouped into infrastructure stability, response capability, industrial revitalisation, and international cooperation, in accordance to their attributes. Based on the agenda distribution, we observed that the approach to cybersecurity differed by country: the US and UK focused on response capability, whereas Japan and the EU focused on the cybersecurity industry and international cooperation, respectively. Furthermore, the distribution of NCSS agendas depended only on the perceived cyber threat environment and approach to cybersecurity by country, and no agenda exhibited a significant increase or decrease in proportion over time, regardless of country. On the other hand, we highlight the necessity for discussions on risk assessment and management systems, intellectual property theft, and supply chain security systems, to diversify cyber security management systems at a national level, based on a comparison of the 15 agendas with the NCSS strategic task of South Korea.

This study provides a comprehensive understanding of the cybersecurity policy agenda from the perspective of South Korea. However, because the scope of the analysis was limited to NCSSs and to deriving implications for future NCSS revisions, we propose discovering policy agendas from a wider variety of sources and comparing them in future research. As presented in the previous literature review, policy agendas could be derived from a variety of sources, including publicly published reports, news articles, research papers, petitions, and even SNS postings. In particular, because of the multidimensional nature of cybersecurity policy, multilateral cooperation efforts across society, government, science, technology, industry, and academia are essential for building global cybersecurity resiliency beyond national security. Therefore, it would be meaningful to comprehensively compare cybersecurity policy demands from various perspectives.

References

[1] O. Koltsova and S. Koltcov, “Mapping the public agenda with topic modeling: the case of the Russian Livejournal”, *Policy & Internet vol. 5, no. 2*, pp. 207-227, 2013.

[2] L. Hagen, O. Uzuner, C. Kotfila, T. M. Harrison and D. Lamanna, “Understanding citizens’ direct policy suggestions to the federal government: a natural language processing and topic modeling approach”, *2015 48th Hawaii International Conference on System Sciences*, 2015.

[3] D. Greene and J. P. Cross, “Exploring the political agenda of the European parliament using a dynamic topic modeling approach”, *Political Analysis vol. 25, no. 1*, pp. 77-94, 2017.

[4] O. B. Driss, S. Mellouli and Z. Trabelsi., “From citizens to government policy-makers: Social media data analysis”, *Government Information Quarterly vol. 36, no. 3*, pp. 560-570, 2019.

[5] S. Pinto and F. Albanese, “Quantifying time-dependent Media Agenda and public opinion by topic modeling”, *Physica A: Statistical Mechanics and its Applications vol. 524*, pp. 614-624, 2019.

[6] L. Sun and Y. Yin, “Discovering themes and trends in transportation research using topic modeling”, *Transportation Research Part C: Emerging Technologies vol. 77*, pp. 49-66, 2017.

[7] H. L. Yang, T. W. Chang and Y. Choi, “ Exploring the Research Trend of Smart Factory with Topic Modeling”, *Sustainability vol. 10, no. 8*, pp. 2779-2793, 2018.

[8] T. Bechor and B. Jung, “ Current State and Modeling of Research Topics in Cybersecurity and Data Science”, *systemics, cybernetics and informatics vol. 17, no. 1*, pp. 129-156, 2019.

[9] J. Dehler-Holland, K. Schumacher and W. Fichtner, “Topic Modeling Uncovers Shifts in Media Framing of the German Renewable Energy Act”, *Patterns vol. 2, no. 1*, 2021.

[10] E. Luiijf, B. Kim and P De Graaf, “Nineteen national cyber security strategies”, *International Journal of Critical Infrastructures vol. 9, no. 1-)*, pp. 3-31, 2013.

[11] K. S. Min, S. W. Chai and M. Han, “An International Comparative Study on Cyber Security Strategy”, *International Journal of Security and Its Applications vol. 9, no. 2*, pp. 13-20, 2015.

[12] R. Sabillon, V. Cavaller and J. Cano, “National Cyber Security Strategies: Global Trends in Cyberspace”, *International Journal of Computer Science and Software Engineering vol. 5, no. 5*, pp. 67-81, 2016.

[13] K. Sarker, H. Rahman, K. F. Rahman, M. S. Arman, S. Biswas and T. Bhuiyan, “A Comparative Analysis of the Cyber Security Strategy of Bangladesh”, *International Journal on Cybernetics & Informatics vol. 8*, pp. 1-21, 2019.

[14] F. Kolini and L. Janczewski “Clustering and Topic Modelling: A New Approach for Analysis of National Cybersecurity Strategies”, Pacific Asia Conference on Information Systems, 2019.

[15] J. An, S. Kang and H. Im, “An Analysis of National Cybersecurity Strategies using Topic Model”, *Korean Journal of International Relations vol. 60, no. 4*, pp. 119-169, 2020.

[16] R. R. Perols and U. S. Murthy “The Impact of Cybersecurity Risk Management Examinations and Cybersecurity Incidents on Investor Perceptions and Decisions” , *A Journal of Practice & Theory vol. 40, no. 1*, pp. 73-89, 2021.

[17] A. Buzdugan, “Review on use of decision support systems in cyber risk management for critical infrastructures”, *Journal of Engineering Science, vol. 27, no. 3*, pp. 134-145, 2020.

[18] V. K. Aggarwal and A. W. Reddie, “New economic statecraft: Industrial policy in an era of strategic competition”, *Issues & Studies vol. 56, no. 2*, 204006, 2020.

[19] O. Osunji, “Know your suppliersL A review of ICT supply chain risk management efforts by the US government and its agencies”, *Cyber Security: A Peer-Reviewed Journal, vo. 4, no. 3*, pp. 232-242, 2021.

Digital Transformation of Human Resource Processes in Small and Medium Sized Enterprises using Robotic Process Automation

Cristina Elena Turcu[1]
Computers Department
Stefan cel Mare University of Suceava
Suceava, Romania

Corneliu Octavian Turcu[2]
Computers, Electronics and Automation Department
Stefan cel Mare University of Suceava
Suceava, Romania

***Abstract*—The aim of this paper was to obtain data and information on the digital transformation of human resource (HR) processes in small- and medium-sized enterprises (SMEs) with the help of robotic process automation (RPA), in order to increase competitiveness in the digital age. Romanian businesses are attempting to close the gap with companies in developed countries by implementing projects that allow the adoption of emerging technologies in HR departments. This paper presents some of the preliminary findings, resulted from a collaboration between a university and an SME, for the efficient implementation of specific HR processes using RPA. The paper provides a brief introduction of the RPA concept as well as a list of HR processes that can be automated within enterprises, with the benefits brought to the enterprise and employees presented in both qualitative and quantitative terms for each HR process. In addition, a case study for the automatic collection of candidates' documents and extraction of primary information about them was considered. Further on, the problems encountered during implementation were listed, along with potential solutions. Given the benefits offered, RPA could play an important role in transitioning HR functions into the digital era.**

***Keywords*—Robotic process automation (RPA); small- and medium-sized enterprises (SME); human resource (HR); digital HR; recruitment**

I. Introduction

We are currently witnessing multiple challenges for companies, such as demographic and social changes, technological advances, etc. In order to meet these challenges, companies need to be agile and adaptable. An important role in fulfilling these expectations, in the conditions of ongoing and massive disruptions, is played by human resources. The good organization of human resources allows the optimization of the employees' work, even in the conditions of the disturbances introduced by the COVID pandemic.

If before the global pandemic, "digital technologies did not find a strong and widely based application in the small- and medium-sized enterprises (SMEs) sector", "due to the COVID-19 disruption, SMEs are now trying to avoid a total shut down of economic activities by introducing digital technologies" [1].

In all US industries, according to the research conducted by the McKinsey Global Institute in 2016, approximately 17% of work consists of data collection and 16% of data processing - tasks that at the time of the study were largely performed by human workers [2].

In small- and medium-sized businesses, human resource (HR) departments typically suffer from aging IT systems. Employees of these departments often need to:

- Enter data into multiple systems;
- Toggle between different applications, entering in one application data retrieved from another one.
- Reconcile data across two or more systems.

To quickly streamline these routine tasks without upgrading or replacing existing old systems, Robotic Process Automation (RPA) technology offers solutions. Thus, much of the manual work involved in these tasks can be automated using RPA.

RPA adoption is growing every year. According to Gartner, in 2020 RPA was, for the second year in a row, the fastest-growing segment in the enterprise software market, with a 38.9% increase to $1.9 billion in revenues [3]. Several studies worldwide estimated an increase in the use of RPA in various fields. Thus, for example, Transforma Insights expects total market spending for RPA will increase from $ 1.2 billion in 2020 to $ 13 billion in 2030 [4].

Over the past few years, we are witnessing the digital HR transformation, “the application of advanced technologies and analytics, digital traits and behaviors, and HR customer centricity through the lens of the organizations HR Operating Model to optimize HR to deliver sustainable organizational performance” (Fig. 1) [5].

Analysts from McKinsey & Company called robotic process automation technology a “third arm” for HR organizations because it works with HR to amplify the department’s capacities [6].

According to a survey conducted by PwC on HR technologies, 45% of respondents (approx. 600 HR and HR information technology leaders on six continents) intend to invest in hyper-automation or robotic process automation (RPA) technology in the following 12-24 months [7].

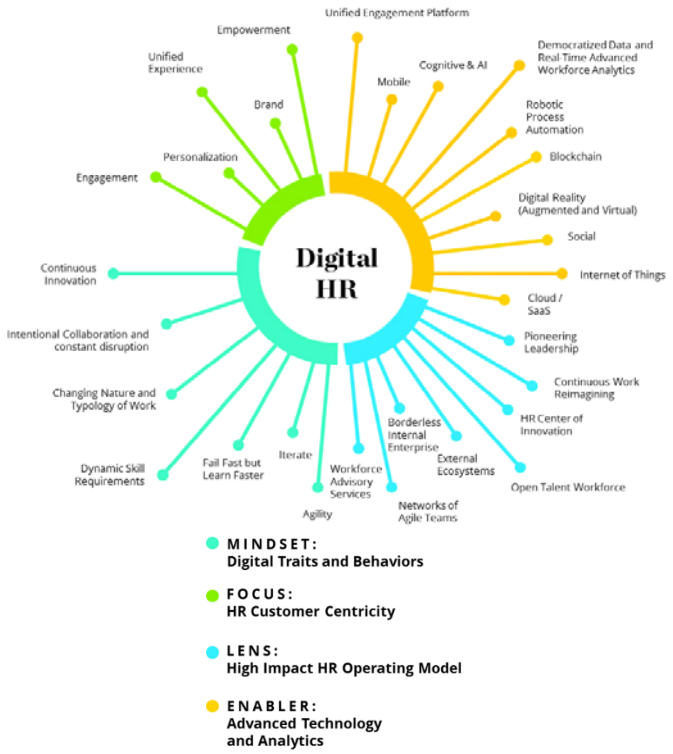


Fig. 1. Digital HR [5].

Published studies on the adoption of RPA in companies refer mainly to large companies. Small- and medium-sized enterprises are often overlooked. Our research seeks to fill a gap in the scientific literature on robotic process automation in human resource departments of SMEs, aiming to ensure the potential transfer of knowledge from academia to industry.

In this regard, we try to provide answers to the following research questions (RQ):

RQ1. What is Robotic Process Automation (RPA)?

RQ2. What are the HR processes in SMEs that can be automated with RPA?

RQ3. What advantages would the adoption of RPA bring in the HR departments of small- and medium-sized enterprises (SMEs)?

RQ4. What are the challenges HR departments are faced with when adopting RPA-based solutions?

To capture relevant knowledge on the topic, we conducted a literature review considering six of the databases frequently used by researchers: Web of Science, Springer, Science Direct, IEEE Xplorer, Scopus, and Google Scholar. In order to gain more insights for this paper, we also conducted a backward-and-forward search.

A further step in our research was the expert interviews, used to confirm and strengthen the list of use cases identified in the literature. Thus, we conducted a semi-structured in-depth interview with senior HR personnel from various SMEs. During the interviews, we considered the following topics: RPA objectives in HR, types of projects, benefits, impact on jobs, challenges. We reviewed the interview responses to obtain key insights that are summarized in the present paper.

The remainder of this paper is organized as follows. In the next section we present some definitions of the term Robotic Process Automation. Section III refers to several achievements in the field, while Section IV presents the processes of HR departments in SMEs where RPA can be incorporated along with their benefits that SMEs and employees can obtain by using RPA tools. The next section presents the flow of an application under development within a national project, mentioning some issues encountered during implementation. Further on, authors highlight the future scope of the technology, and in the last section of the paper we formulate the conclusions of our research.

II. Theoretical Background

A. Robotic Process Automation (RPA)

In order to answer the first research question and to clarify the concept of Robotic Process Automation (RPA), we present some selected definitions that we believe are relevant to our discussion.

IEEE Corporate Advisory Group defines RPA as the use of a "preconfigured software instance that uses business rules and predefined activity choreography to complete the autonomous execution of a combination of processes, activities, transactions, and tasks in one or more unrelated software systems to deliver a result or service with human exception management." [8].

RPA could be described as an "emerging form of business process automation technology based on the notion of software robots or artificial intelligence (AI) workers. RPA has become the new language of business. This technology is more powerful among the 21st century technologies" [9].

RPA technology is non-invasive for existing systems; it does not replace them, but interacts with them. In fact, it's a new layer above the organization's applications and cloud services that must integrate easily and efficiently with all of these.

The concept of robotic process automation is becoming increasingly integrated into various domains as a means for improving productivity, compliance, product quality, etc. In the gray literature, many key use cases are examined, presenting the experience of various organizations that have adopted RPA. The use of this technology allows the emulation of actions taken by human users, so that work processes can be automatically implemented in many business functions, such as: Human Resource, Finance & Accounting, Production, Sales & Marketing, Supply Chain, and Information Technology.

In this paper, authors focus on human resource departments. HR professionals deal with many processes and sub-processes, such as hiring, onboarding, employee benefits administration, managing complex workplace situations, enforcing regulatory changes. This usually means working with multiple software platforms, spreadsheets, documents, etc. Various studies highlight that 33% of companies have more than ten HR systems. 47% of companies have HR software that is over seven years old [10]. According to Center for Effective

Organizations and a study conducted by G&A Partners, HR professionals spend about 73.2% of their time dealing with tedious administrative tasks [11]. But some of these HR operations could be automated with RPA-based solutions. But a successful implementation depends on several factors, including the RPA tool.

B. RPA Tools

Currently, there are many RPA tools available on the market, some of them free, others commercial, such as, Automation Anywhere, Blue Prism, EdgeVerve, Microsoft, NICE, UiPath, WorkFusion, etc.

RPA market solutions offer various capabilities, as presented in numerous studies and scientific papers, such as, for example, [12]. It should be noted that providers of robotic process automation platforms quickly and continuously update their offers to meet the developing needs of their clients. Thus, for example, various RPA providers announced further development of their RPA platforms, by incorporating new breakthroughs in artificial intelligence and machine learning. Given the multitude of existing RPA platforms that offer powerful functionality, as well as their continued development, choosing the best automation tool for an organization could be a real challenge.

III. Related Work

Recently, because of the difficulties caused by the COVID-19 pandemic, many companies have turned to robotic process automation solutions to help them overcome the challenges of the pandemic. Thus, they can "automate repetitive tasks across multiple business applications without altering existing infrastructure and systems" [13].

Many publications present the strengths and weaknesses of the RPA. However, the reviewed literature highlights a predominantly positive assessment of RPA, given that its strengths outweigh its weaknesses [14]. Adopting RPA in enterprises offers many benefits, including higher productivity, improved business efficiency and accuracy, data security, scalability, auditability, low printing and storage [13]. The promises of RPA for enterprises include easy implementation at a relatively low cost compared to other solutions.

Some authors consider that RPA is suitable for processes with high levels of i) standardization, ii) volume of transactions, iii) maturity, iv) approach to business rules ([13], [15], [16], [17]). Other authors recommend the RPA approach for standardized and repetitive processes that i) follow business rules, ii) take longer to complete, iii) are performed on a regular basis, and iv) require manual interaction with information systems [18]. Furthermore, many authors advocate RPA adoption for processes involving structured data, low variance and logic-driven procedures [19].

IV. Discussion

The analysis of scientific and industry publications shows that RPA has been used moderately in the human resource departments of large enterprises and less in SMEs. A single enterprise can have hundreds or thousands of processes and sub-processes and the selection of those that could be automated with RPA is a difficult task.

We examined case studies that are related to the aforementioned research questions. Following an examination of these case studies, and an analysis of the academic and industry papers it can be highlighted that the processes suitable for RPA have several common features. Thus, the considered process should have clearly defined inputs and outputs. Workforce tasks that are rule-based, predictable, high-volume, time-consuming, repetitive, or prone to human error are good candidates for RPA.

Despite the vast potential of RPA technology, experts point out that it is not suitable everywhere. Thus, if a process is dynamic, requires creative thinking, deals with unexpected events or involves decision-making on a case-by-case basis, RPA can fail.

Various studies have shown that RPA projects failed initially in a proportion of 30% to 50% [20]. In order to determine which processes are suitable for RPA, an evaluation can be performed considering the criteria specified above.

As we can see, there are many opportunities to use RPA in human resource management. Below we present some HR processes in SMEs that can be automated with RPA, in response to the RQ2. Table I summarizes the HR use cases found in the literature and validated by some HR experts.

TABLE I. The RPA Opportunity for HR Departments

HR Process to be automated	Benefits	SME	Employee
CV Screening and recruitment	Facilitating employment	X	
	Identifying the best candidates	X	
	Keep candidates informed with better, automate communication	X	X
	Faster cross verification of candidates' details	X	
	Automatic data concatenation from multiple input sources	X	
	Behavior reference check	X	
	Criminal record checks	X	
	Correct evaluation of candidates	X	X
	Minimizing delays in hiring process	X	
	Reduction in hiring costs	X	
Onboarding	Smooth onboarding process		X
	Data integration capacity	X	
	Faster onboarding process		X
	Low-cost onboarding process	X	
Employee training	Accelerated employee skill acquisition		X
	Better alignment of employee skills with the organizational certification requirements	X	
	Enhanced learning experience		X
Employee data management	Better management of data of current/former employees etc.	X	
	Consistent actions across various systems/ databases/departments	X	
	Elimination of data entry errors	X	

	Employee data protection	X	X
	Efficient allocation of office space in the hybrid-work conditions	X	
Tracking attendance	Better calculation of employee salaries		X
	Workflow disruption prevention	X	
	Accurate time records	X	
Payroll management	Avoiding delay in salary payments		X
	Lower payroll costs	X	
	Reduced risk of multiple errors	X	
	Simplified payroll processing	X	
Expense management	Fast processing of expenses		X
	Correlation of individual expenses with company regulations and spending rules	X	
Maintaining compliance	High accuracy of the maintaining compliance process	X	
	Ensuring an error-free compliance maintaining process	X	
Exit management	Exit process consistency	X	
	Ensuring data privacy	X	
Performance management	Faster data collection for the calculation of performance indicators	X	
	Automatic report generation/ updating/ deletion	X	
Qualitative	Quantitative		

The general hire-to-retire (H2R) process integrates several disparate systems that often require swivel-chair work to migrate employee data between different systems. These processes are good candidates for RPA. Various studies have shown that recruitment is one of the most time-consuming key tasks for human resources. On average, recruiters spend 60% of their time for candidates sourcing and screening. Implementing an RPA-based solution can take over manual and repetitive work involved, such as filtering resumes, scheduling interviews, and simplifying the integration process.

RPA can significantly reduce inefficiencies on management reporting. Thus, an RPA-based solution can gather information from multiple sources and generate a report or save it into consolidated XLS.

The implementations we studied showed a significant decrease in process time and in errors, and a high potential for scalability.

Successful use cases will continue to emerge with the increasing development of RPA platforms and as more businesses expand their use of RPA, demonstrating the wide variety of issues that can be addressed.

Further on, we attempt to answer the RQ3 research question, presenting the benefits of adopting RPA. RPA has great value to offer for improving the human resource management system in small- and medium-sized enterprises. The benefits we mean are not only for the company, but also for employees. Thus, RPA-based solutions could offer SMEs important benefits, including increased productivity, improved accuracy, faster digital transformations, higher employee engagement, improved data security and quality, etc. Once an RPA layer has been added within an SME for HR processes, each individual employee can access and exploit it. RPA-based solutions could offer employees [13] a wide variety of benefits in laborious or tedious manual processes. We presented some of the qualitative and quantitative benefits identified following the analysis of the academic and industry literature.

After conducting the overview of various possibilities, the main conclusion is clear. RPA could play an important role to shift the HR functions to enter the digital era.

Next, based on literature research, we present some of the challenges that SMEs face in implementing RPA within HR departments (RQ4).

An important challenge for many SMEs is how to add automation to workflows. There can be a large number of processes and sub-processes within the HR departments of SMEs [21], so choosing which ones could be automated with RPA, as well as the order in which they can be automated can prove to be a big challenge.

Other challenges focus on human resource management, given that workforce need to adapt to new work, as RPA-based solutions can take on important parts of their daily tasks [22].

The challenges can also be considered from a legal viewpoint, analyzing, for example, questions such as "who has the control over the intellectual property robots handle and generate" and "who is responsible if the robot fails" [23]. Unfortunately, these two questions have not yet been answered.

V. Case Study

Recruitment is one of the most promising use cases for adopting RPA-based solutions. Processes in this area involve a large volume of repetitive, time-consuming tasks that are still manually operated by human force. Thus, a significant amount of time is spent screening resumes and application forms submitted by candidates for open positions. Software robots can make this process considerably easier by gathering applications quickly and comparing all of the data to a list of precise job requirements. These requirements can be viewed as predetermined rules that influence the whole selection method while using RPA technology.

Candidates for a position in an SME scan their documents and send them to the SME for processing. Most of the times the documents sent by the candidates are in pdf format. These include diplomas, identity card, certificates, CV, letter of intent, etc. Some documents may have a format imposed by SME, in this case the centralization of information from several candidates is much simplified. For example, an SME can request the completion of important information in editable pdf files with PDF tagged order features. But, retrieving information from candidates' scanned documents is a tedious, repetitive, time-consuming process that involves a multitude of copy-paste activities and requires manual interaction with information systems.

To eliminate these disadvantages, we proposed an RPA-based solution. For this case study, we have opted for the

UiPath platform. Thus, we developed a solution based on RPA enhanced with Optical Character Recognition (OCR) that can recognize and extract the information needed by the HR department from scanned documents and then enter it into the HR database. HR staff can run this solution on their laptop while handling more important activities. Then, the only operation to be performed by the HR department staff is to verify the veracity of the information entered in the system by the RPA solution.

The flow chart of this application is depicted in Fig. 2.

Fig. 2. The Flow Chart of the RPA-based Solution.

The proposed solution has been verified to ensure proper processing. Thus, we evaluated the implemented application with documents of different formats. In our experiments we used pdf files and image / picture files (JPEG, BMP, GIF, PNG, etc.). The extraction of information from the files sent by the candidates is one of the most significant difficulties to be handled. This problem becomes even more acute when the documents are edited in a language with special characters, as the Romanian language. Thus, the information on the applicants' documents may contain diacritics in Romanian. The correctness of the extracted information depends a lot on the OCR engine.

We have developed and tested several OCR engines available in UIPath, such as: Microsoft OCR, Google OCR, Abbyy OCR and other OCR software like: Convertio [24], Online OCR [25], Free Online OCR [26], i2OCR [27]. The best results were obtained using Convertio and Abbyy FineReader. Even if the developed application is in a beginning form, we consider that it is necessary to test as many OCR engines as possible and to select the one that provides the best recognition rate for Romanian characters.

One way to reduce the number of errors in extracting information from scanned files is to use OCR engines that are based on artificial intelligence. One such example is Abbyy FineReader, which has the latest OCR technology based on artificial intelligence and performs a variety of operations such as digitizing, retrieving, editing, protecting and sharing documents of various types in the same workflow [28].

A limitation of the research is the use of only one RPA platform (UIPath) for the case study. To create meaningful solutions for HR departments in SMEs, further research employing different RPA platforms is required.

The authors of this paper intend to continue research in the near future, considering further explorations of RPA in the field of human resource management in SMEs. Attention will be paid to combining solutions developed using RPA tools with solutions based on blockchain, Internet of Things, etc.

VI. Conclusion

This paper aimed to explore the robotic process automation technology and to identify the promising application of RPA in human resource management within SMEs. In the near future, human resource departments that rely on manual, paper processes will need to re-examine their processes and consider whether RPA technology can benefit them. Through this paper, the authors want to support SMEs that aim to streamline the HR department by adopting RPA.

In the analysis of numerous publications and case studies, various authors' observations on how companies and business people are trying to reorganize the company's human resources for an accelerated economic recovery after being affected by the pandemic crisis, we tried to find answers to some of the basic problems of HR digitization, which would contribute to a profitable and sustainable business for SMEs. The adoption of RPA can bring important benefits in the HR departments within SMEs for laborious or tedious, repetitive, time-consuming manual processes.

Through the case study implemented and presented, it was demonstrated that RPA can fulfill some of the basic pre-processing tasks undertaken by human resources departments in recruiting candidates.

The RPA platform used has an important role in the successful implementation of an RPA-based solution. There are various vendors on the market that offer RPA solutions. Choosing the best solution to meet the specific requirements of an SME's HR department is not an easy task, an in-depth analysis must be performed by taking into account several criteria.

Thus, for example, a developer must consider the type of analyzed documents when implementing RPA-based solutions for managing candidates' documents (scanned or generated, with or without images, with or without tagged order features, etc.). The correctness of the extracted data is also dependent on the ability of OCR tools to recognize the specific characters of the language used in the elaboration of documents. In many cases, processing data that has been extracted incorrectly can be time-consuming.

Although some HR processes require human intervention to correct extraction errors, RPA can be considered a viable candidate for streamlining HR processes, which can lead to a rapid improvement in a business's overall value.

ACKNOWLEDGMENT

This research was funded by the project "119722/Centru pentru transferul de cunoștințe către întreprinderi din domeniul ICT—CENTRIC, Contract subsidiar 15569/01.09.2020, Platformă inteligentă pentru gestionarea resurselor umane - HR ASSISTant", contract no. 5/AXA 1/1.2.3/G/13.06.2018, cod SMIS 2014+ 119722 (ID P_40_305).

REFERENCES

[1] Gregurec, M. Tomičić Furjan, and K. Tomičić-Pupek, "The impact of COVID-19 on sustainable business models in SMEs," Sustainability, vol. 13, no. 3, p. 1098, 2021.

[2] M. Chui, J. Manyika, and M. Miremadi, "Where machines could replace humans-and where they can't (yet)," 2016.

[3] F. Biscotti, V. Mehta, A. Villa, B. Bhullar, and C. Tornbohm, "Market share analysis: robotic process automation," worldwide, 2019. Technical report, 2020.

[4] "Robotic Process Automation 101: a primer on automating IT-based tasks via bots imitating human behaviour - Reports & Insights." https://transformainsights.com/research/reports/robotic-process-automation-technology-insight-report (accessed Nov. 19, 2021).

[5] "Digital HR | Deloitte | Human Capital," Deloitte Danmark. https://www2.deloitte.com/dk/da/pages/human-capital/articles/digital-hr.html (accessed Sep. 20, 2021).

[6] "How bots, algorithms, and artificial intelligence are reshaping the future of corporate support functions | McKinsey." Avaliable: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-bots-algorithms-and-artificial-intelligence-are-reshaping-the-future-of-corporate-support-functions?cid=other-eml-alt-mip-mck-oth-1811&hlkid=a214ffef34fa4418b3bacbffe56be12d&hctky=9653147&hdpid=e7907fc7-7f94-4d63-a845-e2913105cc0d (accessed Nov. 27, 2021).

[7] PricewaterhouseCoopers, "2020 HR Technology Survey: Key findings," PwC. https://www.pwc.com/us/en/services/consulting/workforce-of-the-future/library/hr-tech-survey.html (accessed Oct. 19, 2021).

[8] I. C. A. Group, "IEEE Guide for Terms and Concepts in Intelligent Process Automation." IEEE New York, NY, 2017.

[9] R. Syed et al., "Robotic process automation: contemporary themes and challenges," Computers in Industry, vol. 115, p. 103162, 2020.

[10] Josh Bersin, "15 Two Major Marketplace Issues," Dec. 2014. [Online]. Available: https://www.slideshare.net/jbersin/ten-disruptions-in-hr-technology-for-2015-ignore-at-your-peril/12-15Two_Major_Marketplace_Issues1_Too

[11] "HR's Time-Consuming Toll On Your Company," G&A Partners, Jan. 15, 2015. https://www.gnapartners.com/resources/infographics/hrs-time-consuming-toll-company (accessed Nov. 15, 2021).

[12] P. Hofmann, C. Samp, and N. Urbach, "Robotic process automation," Electronic Markets, vol. 30, no. 1, pp. 99–106, Mar. 2020.

[13] A. Leshob, A. Bourgouin, and L. Renard, "Towards a process analysis approach to adopt robotic process automation," in 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), 2018, pp. 46–53.

[14] J. Hindel, L. M. Cabrera, and M. Stierle, "Robotic process automation: Hype or hope?," in Wirtschaftsinformatik (Zentrale Tracks), 2020, pp. 1750–1762.

[15] M. Lacity and L. Willcocks, "Paper 16/01: Robotic Process Automation: The Next Transformation Lever for Shared Services." Retrieved from The Outsourcing Unit, LSE: http://www. umsl. edu/~ lacitym …, 2016.

[16] M. Lacity, L. P. Willcocks, and A. Craig, "Robotic process automation at Telefonica O2," 2015.

[17] L. Willcocks, M. Lacity, and A. Craig, "Robotic Process Automation at Xchanging," The Outsourcing Unit Working Research Paper Series.

[18] P. Lowes, F. R. Cannata, S. Chitre, and J. Barkham, "The business leader's guide to robotic and intelligent automation," Deloitte Development LLC, 2017.

[19] C. Mendis, C. Silva, and N. Perera, "Moving ahead with Intelligent Automation," 2016.

[20] "Can robots help your business be more human?" https://www.ey.com/en_gl/digital/can-robots-help-your-business-be-more-human (accessed Oct. 27, 2021).

[21] K. McCandless, "7 Essential Human Resources Processes for Small Businesses." https://www.fool.com/the-blueprint/human-resources-process/

[22] D. Choi, H. R'bigui, and C. Cho, "Robotic Process Automation Implementation Challenges," in International conference on smart computing and cyber security: strategic foresight, security challenges and innovation, 2020, pp. 297–304.

[23] C. Holder, V. Khurana, F. Harrison, and L. Jacobs, "Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)," Computer law & security review, vol. 32, no. 3, pp. 383–402, 2016.

[24] Convertio – File Converter.

[25] Online OCR. [Online]. Available: https://www.onlineocr.net/.

[26] Free Online OCR. [Online]. Available: https://www.newocr.com/.

[27] i2OCR. [Online]. Available: http://www.i2ocr.com.

[28] ABBYY FineReader. [Online]. Available: https://pdf.abbyy.com/.

Computerization of Local Language Characters

An Innovative Model for Language Maintenance in South Sulawesi, Indonesia

Yusring Sanusi Baso[1]
Center for Media
Learning Resources and E-Learning
Quality Assurance and Education Development Institutions
Hasanuddin University, Makassar, Indonesia 90245

Andi Agussalim[2]
Laboratory of Educational Technology
Department of Western Asian Languages
Faculty of Cultural Science, Hasanuddin University
Makassar, Indonesia 90245

***Abstract*—The objective of this study is to provide innovative model for the approach of language preservation. It is necessary to maintain indigenous languages in order to avoid language death. Script applications for indigenous languages are one of the solutions being pursued. This script program will facilitate communication through writing between speakers of indigenous languages. Additionally, the study illustrates the implementation of the Lontara script (Bugis-Makassar local language letters and characters). This script application is compatible with the Microsoft Windows operating system and the Hypertext Transfer Protocol (HTTP). This study employed the research and development (R&D) approach. Six stages are followed in this R & D study: 1) doing a requirements analysis to determine the viability of Bugis-Makassar indigenous languages in everyday life and also to determine ways to retain them 2) designing and constructing Lontara scripts with hypertext-based applications, 3) producing Lontara scripts with hypertext-based applications, and 4) validating the hypertext-based applications through one-to-one testing, small and large group testing. 5) Lontara application revision; and 6) Lontara application as a finished product. This product is designed to be used in conjunction with other interactive applications.**

***Keywords*—Innovative model; language maintenance; Lontara script; Makassarese; local language; hypertext-based application**

I. Introduction

Indonesia is a multi-ethnic and multilingual country. The Indonesian Central Bureau of Statistics [1] revealed that Indonesia consists of 1,128 ethnic groups. Besides, BPS also presents ethnographic data and records about 700 local languages owned by Indonesia's state. The diversity of ethnic groups in Indonesia is caused by various factors, including historical and natural isolation factors. The old nature isolation factor also influences Indonesia's condition, which various ethnic groups inhabit. Therefore, ethnics are diverse and have their characteristics in terms of language and culture, although they all belong to the same language family, the Austronesian. On the other hand, there are 7,117 languages in the world that are spoken today, and it has estimated in the year 2020 that 40-90% of them will disappear during this century [2]. However, 20-50% of this amount is no longer used by the younger generation of its native [3].

The picture of the vitality of local languages in Indonesia, in South Sulawesi, is still unclear. It is still challenging to obtain accurate data on local language vitality in South Sulawesi. However, this condition can be seen by a glance at native speakers of Makassarese, one of the local languages in South Sulawesi. The vitality of the Makassar language can be seen by the Extended Graded Intergenerational Disruption Scale (EGIDS) rubric that has been developed by Simons and Lewis [4] and [5]. Based on EGIDS, the Makassar language is at level 6b [6], [7].

The vitality of local Indonesian languages is more easily recognized by the EGIDS than other scales, for example, the Fisman scale [8]. Fishman's Graded Intergenerational Disruption Scale (GIDS) and UNESCO's five steps endangerment framework. EGIDS is a union of the Fishman scale and the UNESCO scale. The EGIDS scale can be seen as in the following Table I.

Based on EGIDS, we can assess the vitality of the Makassar language. However, before assessing its vitality, we should first understand the local language environment in Indonesia. Indonesia's local language environment is the pressure most like in China, both vertical and horizontal pressure [9]. The pressure on Indonesia's local languages generally comes from Indonesian and its dominant or more native languages. This pressure is due to prestige for Indonesian speakers, educational interests, and even economic factors. Another is the Indonesian language pressure, which comes from international languages, Arabic and English.

On the other hand, regional languages are under the position of Indonesian. Therefore, horizontal pressure can occur from other regional languages with a better sociolinguistic status [10]. This condition further weakens the vitality of local languages in Indonesia.

Ethologically, this condition can be understood that Indonesia's local languages are very worrying [4], [11], [12].

Based on the above Table II, it can be seen that almost half of Indonesia's languages are "in trouble or worse." The number of languages that are rated "Vigorous" (260) compared with those that are "In trouble" (272) can be concluded that the local languages in Indonesia are in trouble. We can also have an alternate assessment and visual representation to obtain local language vitality in Indonesia through UNESCO"s. "Interactive Atlas of the World's Languages in Danger" classifies 144 of Indonesia's languages as "Vulnerable" or worse. Ethnologue website has stated that:

"The number of individual languages listed for Indonesia is 719. Of these, 707 are living, and 12 are extinct. Of the living

languages, 701 are indigenous, and six are non-indigenous. Furthermore, 18 are institutional, 81 are developing, 260 are vigorous, 272 are in trouble, and 76 are dying [13].

Based on this fact, the research question is, "How to design an innovative model of local language maintenance?"

TABLE I. EXPANDED GRADED INTERGENERATIONAL DISRUPTION SCALE

Level	***Label***	***Description***	***UNESCO***
0	International	The language is used internationally for a broad range of functions.	Safe
1	National	The language is used in education, work, mass media, government at the national level.	Safe
2	Regional	The language is used for local and regional mass media and governmental services.	Safe
3	Trade	The language is used for local and regional work by both insiders and outsiders.	Safe
4	Educational	Literacy in the language is being transmitted through a system of public education.	Safe
5	Developing	The language is used orally by all generations and is effectively used in written form in parts of the community.	Safe
6a	Vigorous	The language is used orally by all generations and is being learned by children as their first language.	Safe
6b	Threatened	The language is used orally by all generations, but only some child-bearing generation transmit it to their children.	Vulnerable
7	Shifting	The child-bearing generation knows the language well enough to use it among themselves, but none are transmitting it to their children	Definitely endangered
8a	Moribund	The only remaining active speakers of the language are members of the grandparent generation.	Severely endangered
8b	Nearly Extinct	The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language.	Critically endangered
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency.	Extinct
10	Extinct	No one retains a sense of ethnic identity associated with the language, even for symbolic purposes.	Extinct

TABLE II. EGIDS OVERVIEW OF THE VITALITY OF LANGUAGES OF INDONESIA

Languages	Institutional (EGIDS 1-4)	Developing (EGIDS 5)	Vigorous (EGIDS 6a)	Threatened (EGIDS 6b-7)	Dying (EGIDS 8a-9)	Extinct (EGIDS 10)
719	18	81	260	272	76	12

II. RELATED WORK

Language maintenance programs, language shift, and endangered language are topics that have never-ending debate. The linguists are always looking for alternative plans to maintain minority languages that a language shift has caused. Language researchers often face people who leave their mother tongue due to political and economic pressures [14]–[16]. A language preservation program is carried out in legal recognition of the minority language [17], [18]. Other researchers propose a theory for heritage language, literacy, and identity processes to maintain minority languages [19]. Other language researchers make breakthroughs in maintaining minority languages by using a smartphone [20].

Other studies have shown that sometimes a language appears suddenly, spreads quickly, and quickly disappear. This case challenges linguists to document, describe, preserve, and revitalize languages [21], [22]. Another linguistic phenomenon that often affects decreasing local language speakers is the influence of strong language used on daily life, both informal and non-formal [10]. Preservation of minority languages can survive because it is a language preservation policy [23].

In his book on global paradox, John Naisbitt reiterated that globalization promotes a paradoxical tendency [24]. John Naisbitt's view is proven at present. Technological advances in transformation and the informatics revolution have led to human beings' tendency to a one-tier, modern, and global world. On the other hand, modern humans also long for past histories in ethnic romance, values, and primordial styles. The point of these tendencies can cause conflict, friction, and shock.

If John Naisbitt's view is compiled with ethnological data, modern world-class and ethnic romance and primordial values are inevitable. Our society is undoubtedly inseparable from cyberspace and connectivity that can no longer be encountered. At the same time, the use of local languages is decreasing. The shortcomings can be seen in the number of websites or sites that provide local language information.

The clash of these trends creates implications for the use of local languages. The local language users are more likely to use Indonesian as their daily language. This condition cannot be avoided because Indonesian is the national and official language, both oral and written. It is understandable if people use Indonesian to communicate daily, both at work and sharing information. The local language is only used at the oral or spoken level and rarely uses local language in written form.

The shift in language utilization from the local to the Indonesian language is caused by various factors, including economic factors, migration, and marriage. Economic factors caused Makassar speakers (for example) in Gowa Regency's highlands to use the Indonesian language. Javanese speakers and other tribes who trade in this area use the Indonesian language. The influence is quite crucial, which appears from the public began to get used to speaking Indonesian. On the other hand, the Indonesian language is considered prestigious and proud of society's expression when using the Indonesian language. When we as researchers met Ngawing, one of an elder in this village, he said in Makassarese that "*abbicara*

malayu tauwwa," which means "Wow, they speak Indonesian." This sentence is an expression of those who can speak Indonesian [25]. The Indonesian language is forced to be used as a communication language in the commerce sector. Although it unwittingly has shifted the attitude of the language of some highland communities of Gowa Regency, especially in the Tompobulu district, this condition is very reasonable.

Population migration to Makassar City is another factor that contributing to the shift of language use from Makassar to the Indonesian language. Some Gowa highland residents in Tompobulu Sub-district moved to Makassar city looking for work after harvesting in the dry season. Young people in their productive ages left home to earn a living. Some of them migrated to the island of Borneo and the country of Malaysia. When returning to the village or returning home, they generally use the Indonesian and Malay languages. Not infrequently among them "back home" to meet their relatives and back again to the island of Borneo and the State of Malaysia. They prefer to speak Indonesian or Malay rather than Makassar [25].

Inter-tribal marriage also became the cause of the shift of language use from Makassar to Indonesian. Some Gowa highlands residents have been married to other tribes like Bugis, Mandar, Toraja, and Java. Families who are married to other tribes use Indonesian in their daily conversation. Makassar language is used only if their relatives who use it visit them, but when other relatives visit them from mixed-marriage status, they use Indonesian.

Shifting the use of language from this local language to the Makassar language needs to be addressed. If the shift has taken place at the oral language level, it can be expected that the use of local languages in writing may be more severe. More severe in the sense that as majority speakers of local languages, in this case, speakers of Makassar, they no longer use the language of Makassar in writing. This condition is worsened by the digital era that does not provide characters or lontara letters required in various applications, for example, on the hypertext level or the website page.

Various attempts at writing Lontara script have been made, including typewriters, machine-set photos, and computers. The recording of Lontara script making was once presented by Barbara Friberg [26] at a Makassar Golden Hotel seminar. Barbara is evident in the exposition of the history of the digitization development of the Lontara script.

In 1985, the Consulate General of Japan in Makassar sponsored the manufacture of the Lontara typewriter. This device is good enough, and it can produce a good letter. However, typewriters have some weaknesses, including the letters' size is small and the same width. This size results in less than perfect spaces. Whatever the outcome, Taufik Sakuma's efforts should be rewarded for speakers of languages that use the Lontara script [26].

Five years following, in 1990, at Language Center in Ujung Pandang, Ms. Astuti Hendrato from Jakarta collaborated with Monotype Typography in England and worked on Lontara script using LASERCOMP the photo-set machine. A year later, in 1991, USI / IBM, for the guidance of the former Rector of Hasanuddin University and South Sulawesi Governor, Prof. Dr. Ahmad Amiruddin, and Makasar community leaders, provided a computer with a scanner to preserve the Lontara texts. As a result, the manuscripts can be stored on a computer that can then be published quickly. This effort means preserving the Lontara manuscripts in the file image form on the computer. Thus, the systems (font/font types, such as TrueType Font types compatible with Windows) used to write computer-assisted Lontara scripts did not exist yet [26].

Barbara Friberg has also done a person who works on writing Lontara script. She once applied to Monotype Typography in the United Kingdom to help the Makasar community acquire fonts-computers that can be used in various personal computers, not just on specific machines. In 1991, Barbara Friberg's colleague in Singapore was willing to help prepare the Lontara font used on similar IBM-PC and Microsoft Word programs. With the help of Drs. Djirong Basang, Barbara Friberg sent a picture for every Lontara letter used in the Makasar language. The effort was running, but eventually, Barbara Friber's colleague in Singapore could not complete making the Lontara font as expected. In 1994, Barbara Friberg then tried to build the Lontara script font. She continued to develop the program, which was not finished in Singapore. The program that Barbara Friberg used to build the font was the FONTMONGER program. With this program, she created a Lontar21 font with type, True-Type, later known as Lontar21.ttf. The file size is 28 kb. This font was presented in Makassar Golden Hotel in October 1995 [25].

In the same year, precisely in December 1995, Andi Mallarangeng and Jim Henry made Lontara font volume one with BugisA (also with True Type type). BugisA.ttf font file size is 16 kb. However, BugisA font is often experiencing constraints or unstable in Office Word (Windows). An example of a common obstacle is that the Lontara script's initial letters often turn into Latin or boxes in the first word in each paragraph's first sentence. The two types of fonts, both Lontar21 and BugisA fonts, do not add specific numbers to Lontara. Both types of fonts can only be run and used on Windows-based operating systems. Both of these fonts also cannot be used in hypertext-based platforms. In other words, both types of fonts cannot be used to write messages in emails (email) and social media Facebook. The two fonts are also not prepared a consonant marker (diacritic) to facilitate Lontara character speakers reading the Lontara script. Thus, this hypertext-based Lontara font is a development of the previous font.

The character development history triggers the addition of numeric characters and consonant markers to the Lontara script application. The researchers found it essential to add the character of numbers and consonant markers in the Lontara script after seeing the Hijaiyah script's development (Arabic). At first, the Arabic script did not recognize a point and was without a vowel. The example of letters that do not have dots are [ج] ,[ح], and [خ]. These three letters are the same; both have no dots in the form [ح]. It is similar to the other letters, for example, ,[ت] ,[ب] and [ث]. These three letters are initially in the same form, i.e., [ب], but without dots. In this era, not many people can read and write in Arabic script.

History records the development of Arabic script mastery through the punishment for prisoners of the Badr war. After the war, the Muslims captured the Quraysh army. Umar bin Khattab r.a asked that they should be beheaded. However, Abu Bakr r.a. disagreed and proposed that the ransom of literate prisoners teach ten Muslims' sons. As for those who cannot read and write, redeemed in the form of payment. Both ransoms are indispensable for the Muslims at that time.

At the time of Ali bin Abi Talib came the scholar Abu al-As'ad al-Du'ali [27], who added dots to the Arabic character. Furthermore, the addition of marking (harakat) or the line is done by other scholars, namely Ahmad al-Khalil [28]. The addition of this vowel made it easier for non-Arabic speakers to read the Qur'an and hadith correctly. However, for educated and able to read Arabic script without a vowel, the additional marking is no longer needed. Therefore, the Hijaiyah script's development is the idea that inspires writers to add Lontara characters, numbers, and consonant markers.

It should be noted that the Lontara script is syllable and has no consonants. This condition makes a word written in the Lontara script sometimes difficult to read, not only by foreign speakers of the Makassar language but also the native speakers. The word "*paja*" /paja/ can be read /paja/ which means butt and can also be read /pa'ja/ which means salty. This condition is like an Arabic character with no marking (harakat) and can be read according to the position of words in a sentence. This condition has inspired the research team to modify or additions to the characters in the Lontara script.

III. Methodology

This study uses research and development (R & D) methods [29]–[31]. According to Borg and Gall, "R & D method is a process used to develop and validate educational products." Therefore, this sentence can be interpreted that the R & D method is developing and validating educational products. Thus, R & D has a series of cyclical research and development steps. Additionally, each step to be performed should refer to the results of the previous step. Therefore, a new educational product will be obtained at the end of every stage or step applied.

Borg and Gall have presented a series of stages to be followed in this R & D approach: research and information collecting, planning, preliminary proof of form, preliminary field testing, primary product revision, main field testing, operational product revision, operational field testing, final product revision and dissemination and implementation.

R & D research stages, both discovered by Borg and Gall and by Dick and Careys, have ten steps. Of course, this stage of the R & D model always culminates with a product or output. However, in this research, R & D steps are modified in such a way as required. According to the end of the research stage, the researchers determine that other researchers can continue these steps. Whatever stage is chosen at the end of the research stage, the result remains a product.

Research development of Lontara script application with this hypertext-based ends at the sixth step called the final product. The following are the steps:

- Needs analysis. The researchers have conducted a needs analysis and literature review.
- Design of the lontara character application. In this step, the researchers designed a model of the lontara script application. Again, the researchers focused on software, hardware, and humanware characteristics.
- Development of the lontara script application. By this step, the researchers have developed true type font for windows and a lontara character application for HTML called Yusring Keyboard.
- Formative evaluation step. This formative evaluation stage consists of expert validation, one-to-one testing, small group testing, and large group testing. In addition, participants who participated in this stage filled out an instrument called the Technology Acceptance Model (TAM).
- Revision of the lontara character application. Based on the instrument from participants, researchers revised bugs that occur during the stage of formative evaluation.
- Final Product. Thus, this application (yusring keyboard) has been tested in the small and large groups test phase of Formative Evaluation or the fourth stage. The yusring keyboard is ready to use.

In brief, the stages of this method can be seen in the following Fig. 1:

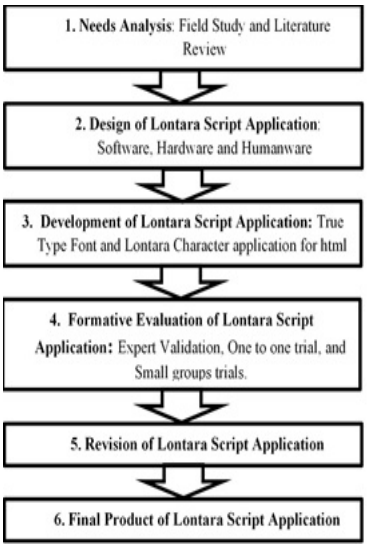


Fig. 1. R & D Steps.

IV. Result and Discussion

This research was conducted for three years, from 2017 to 2019. Application tests were carried out by speakers of Bugis and Makassar languages, two local languages in South Sulawesi, Indonesia. Since the beginning of 2020, this application has been free to download on the Hasanuddin University website, https://web.unhas.ac.id/arab/

Referring to previous research methods, in the first stage, the researchers have conducted a needs analysis and literature review since 2018. This R and D research continues previous

research that has produced a True Type Font (TTF) called Lontara Yusring. The Lontara Yusring is limited to use on Windows Operating Systems only. Along with the times, Lontara Yusring's user community needs a script application that can be used to communicate with popular media, including email, Facebook, and smartphones. The researcher then developed an application to meet the Lontara script user's needs based on this need.

The development of communication technology is the main reason for local language teachers in South Sulawesi to use the latest media to support the learning process. The research team conducted depth interviews in three groups. The first groups are the teacher group in the Tompobulu sub-district, Gowa Regency. Secondly is the Local Language Teachers Association in Parepare City. Furthermore, the third group is the teachers at the As'adiyah Islamic Boarding School Sengkang, Wajo Regency. Based on this dept interview, the researchers obtained information that they urgently need a local language application to be used in the learning and communication process.

In this in-depth interview, those groups explained that in the learning process, writing a lontara script as a local script was not enough; it had to be supported by technology that could be used to write the lontara script. For writing needs on a Windows-based laptop, Lontara Yusring has fulfilled these needs. However, for general communication purposes, another application is required. Based on this need analysis, the research team plans to develop the Lontara Yusring into an application that can be used on HTML and Android platforms.

To meet the needs mentioned in the first stage, the research team then designed the Lontara application. In this step, the researchers designed a lontara script application model focused on software, hardware, and humanware characteristics. This application model is designed to be integrated with the HTML and Android platforms. However, the Lontara characters' position on the keyboard is maintained according to QWERTY as in Yusring Lontara. Thus, the users of Lontara Yusring will remain familiar with the standard of the Lontara Yusring keyboard.

In the third stage, the lontara application design was developed. The research team agreed to give this application a name with Yusring Keyboard. This naming is intended to establish sustainability between the Lontara Yusring and the Yusring Keyboard Application. Therefore, the Lontara Yusring and Yusring keyboard will be combined in one term, namely the lontara application.

The Lontara Yusring, which is intended to be developed in this research, is as shown in the following Fig. 2.

Formative evaluation is the fourth step. This formative evaluation stage consists of expert validation, one-to-one testing, small group testing, and large group testing. Respondents who participated in this stage filled out an instrument called the Technology Acceptance Model (TAM). This TAM instrument is used to measure the acceptability of a technology product marketed to the public. The TAM instrument is also used to determine the ease with which technology is adopted and used by the community [32]–[36].

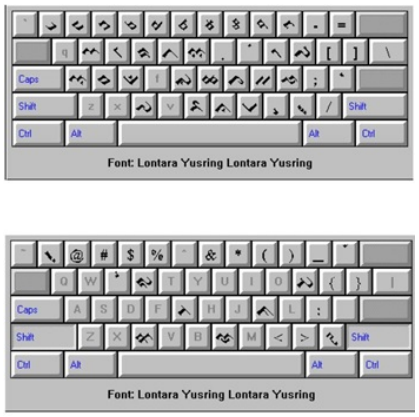


Fig. 2. The TTF Lontara Yusring.

The research team used three TAM instrument constructs: perceived ease of use, perceived usefulness, and attitude toward using. In addition, the team provided instruction that helped the respondent qualify how to interpret the scale (1 being terrible, ten being excellent).

- Perceived Ease of Use
  - How would you rate your experience using the lontara application on MS Office?
  - How would you rate your experience using the lontara application on the HTML Platform?
  - How would you rate your experience using the lontara application on WhatsApp or Telegram for Windows?
- Perceived Usefulness
  - This application is quickly used in writing Lontara characters
  - Users get the benefits of the Lontara application
  - This application is more effectively used in compiling local language learning materials
- Attitude Toward Using
  - The position of the QWERTY keyboard is easily recognized on a laptop
  - The Lontara Yusring is obtained quickly at the Theme Fonts of MS Office

Expert validation is done by giving this application to developers. The research team included additional requirements that the expert can speak either in Buginese or in Makassarese. This competency is needed because speakers of these two local languages will use the Lontara Yusring and Yusring keyboards. At this stage, two experts used the Lontara Yusring and Yusring Keyboard. One of them is the teaching staff at the Geophysics Study Program, Faculty of Mathematics and Natural Sciences at Hasanuddin University. Another expert is a lecturer at the faculty of computer science at the Indonesian Muslim University. The first expert can speak the Bugis language fluently. The second expert speaks fluently in Makassar.

The next stage is a one to one test. At this test, the lontara application is given to respondents to use on MS Office, email, Facebook, WhatsApp, and Telegram for Windows. In this stage, the application was addressed to three lecturers within the Faculty of Culture Science at Hasanuddin University and three students in the same environment. In selecting respondents, the research team required that they communicate either in Buginese or Makassarese languages. Also, the respondents had learned to write using the lontara script in elementary or middle school. These respondents were asked to use the lontara application then give a rating according to the TAM instrument.

The small group test is the third test in this fourth stage. Respondents involved in this test must also meet the requirements, namely speaking either Buginese or Makassarese. Also, they are teachers of Bugis or Makassarese languages. Another requirement is to use MS Office, send an email, and write on Facebook, WhatsApp, and Telegram.

In this test, two groups of language teachers were involved. The first group is the Makassar language teaching group. They are twenty teachers and live in the Tompobulu sub-district, Gowa regency. The second group is the Bugis language teaching group in Parepare, where twenty-four teachers participated in this application test. These two groups were chosen because the Lontara script is used by speakers of the Makassar language and Bugis language. However, the level of Makassar language in EGIDS is lower than that of Bugis.

The research team assisted the two groups in installing the Lontara application on their laptops. First, the research team explained how to use the Lontara application in MS Office, email, Facebook, and social media. Then they were asked to use the lontara application to write some local language vocabulary in MS Word. After that, they emailed the vocabulary to one of the teachers. Messages in emails must use the local language in Lontara characters. At the end of the meeting, the two groups were asked to rate the TAM instrument.

The fourth phase in the formative evaluation stage is the large group test. Respondents who participated in this phase had the same requirements as the previous small group test. The difference between small and large groups is only in the number of people involved, and they are not necessarily local language teachers.

The first group is the teachers of the As'adiyah Islamic boarding school in Sengkang, Wajo district. Forty-five teachers participated in this test. The second group is elementary school teachers in the Tompobulu sub-district, Gowa district. Eighty-one teachers were involved in this test.

The application test results in these two groups can be seen in the following Table III.

Revision of the lontara application. Based on the instrument from participants, researchers revised bugs that occur during the stage of formative evaluation. The respondents' suggestions are generally not related to this application, but rather on its socialization to users of this application. Another suggestion is that a guide to installing the Lontara application on the website that provides this application and a guide on using it on MS Office, HTML and Whatsapp for Windows has also been prepared.

Final Product. Thus, this application has been tested in the small and large groups test phase of Formative Evaluation or the fourth stage.

The need to present information in local languages must be well prepared, especially in local languages. The needs of these communities must be aligned with the development of information technology. The digital era has forced humans, including regional language speakers, to interact with other humans through cyberspace. The internet in various electronic mail and social media on Facebook is one of modern humans' most widely used media today.

Although Lontara literature is often marginalized with Latin characters in literacy, it does not mean that the Lontara script should be abandoned. However, some efforts to bring Lontara character to the young generation of the Bugis-Makassar community must be implemented through information technology in the form of Lontara-based hypertext applications. This Lontara application's readiness is undoubtedly expected to stimulate Lontara users' interest in the HTML platform.

The model of the Lontara application is still referring to the Lontara Yusring. However, this hypertext-based Lontara application has been adapted to the characteristics of the Lontara script itself. Also, the characteristics of software, hardware, and human ware remain to be considered.

TABLE III. LARGE GROUP TEST

TAM Instrument Constructs	Average scale of Gowa group	Average scale of Wajo group
a. Perceived Ease of Use		
How would you rate your experience using lontara application on MS Office?	8	8
How would you rate your experience using lontara application on HTML Platform?	10	9
How would you rate your experience using lontara application on WhatsApp or Telegram for Windows?	9	10
b. Perceived Usefulness		
This application is quickly used in writing Lontara characters	10	9
Users get the benefits of the Lontara application	8	9
This application is more effectively used in compiling local language learning materials	9	8
c. Attitude Toward Using		
The position of the QWERTY keyboard is easily recognized on a laptop	8	10
TTF Lontara Yusring is obtained quickly at the Theme Fonts of MS Office	10	9
AVERAGE	9	9

The Lontara Yusring is still paying attention to the QWERTY keyboard model. However, this Lontara Yusring can only run on a Windows-based operating system. In the browser application, some treatments on settings need to be done. The treatment is required after the hypertext-based Lontara Yusring is installed. This need should be done because this Lontara letter's application does not include the Windows operating system's default fonts. Thus, a personal computer or laptop that has not been installed with a Lontara application cannot recognize the character.

Via email, Bugis-Makassar speakers can send messages using this research product. Indeed, to read these lontara characters on personal computers and laptops, the user needs to install the lontara application since the Lontara Yusring is not the official default of the Windows operating system. Therefore, lontara script users must install this application. So, both the message's sender and the lontara script's message must first install the lontara application. If it is not installed, lontara characters will appear on the personal computer screen or laptop users in the form of boxes.

Users of the lontara application can communicate either when sending messages via email or Facebook or creating interactive applications. Of course, it is expected that the younger generation of Bugis-Makassar knows and utilizes this lontara application. Using the lontara application, both oral and written, can be maintained in the writer's view. Thus, on the other hand, there is a local language sustainability action in place in this country expressing their ideas using the local language. Sometimes past romanticism has been just right and appropriate if it is expressed in the local language. Of course, local characters such as lontara characters should be able to support the speakers' wishes. In other words, the local script's character should be integrated with the development of information technology.

The same procedure can be done when the lontara application will be used on Facebook or social media. For example, an essential step for Facebook users (Facebooker) is to change the keyboard system from a standard keyboard to a lontara keyboard. The trick was relatively easy, that is, by choosing the lontara keyboard. In other words, this stage is the same if the user of the lontara application will send a message via email. An example of the lontara script on Facebook social media can be seen in the following picture.

The users of the lontara application can also take advantage of creating interactive material questions. Various applications can make this type of interactive exercise or interactive questions, including Hot Potatoes [37]. Hot Potatoes [38] is one of the most popular closed-circuit maker applications. This app is used to create multiple choice questions, match, crosswords, short stuffing, and composing sentences—result or output of this application in HTML-based file. The language characters that can be used in this Hot Potatoes application are all characters that are the default Windows operating system. Thus, the lontara application must be installed first for a personal computer or laptop user lontara application.

Lontara Application can also be used on other social media, for example, Whatsapp and Telegram. In this case, the lontara application must be used on the keyboard of a laptop or personal computer (input) and not on a smartphone. However, Whatsapp or Telegram, where the lontara application is used, can still be seen (output). For example, the lontara application, which is inputted on Whatsapp and Telegram laptop but seen on mobile, can be seen in the following Fig. 3:

Fig. 3. Lontara Yusring on Whatsapp.

Although the Lontara application can be utilized in Microsoft Office, the HTML platform can be seen on android. It does not mean maintaining the local language ends at this stage. The central and local government policies are essential to maintain regional languages. Without this policy, the local language observers, including language researchers, will face many potential challenges. For example, government policy requires every elementary and secondary level to continue studying their local languages, which is very important. However, the policy is not enough. The government of society must create space and a stage to display regional languages and literature. Government policy can be implemented by preparing columns in local newspapers once or twice a week. Publication of language and literature and local wisdom written in the local language familiarize the speakers of local languages accustomed and aware of the region's local wealth and wisdom.

To sustain the local language is a joint effort of various nation components. The government has a role in making policies and preparing budgets to support the policy. On the other hand, educators should teach the important role of regional and literary languages in life, especially in primary and secondary schools. Do not miss the local language users to use the local language in everyday conversations, especially non-formal events. Communication in the local language is not intended to undermine the national language's function since Indonesian can be found in every segment of communication, both formal and informal. This condition makes the Indonesian nation, directly and indirectly, use the Indonesian language in both oral and written forms.

This condition is somewhat different from the existence of local languages. Besides, local language characters should be brought closer to the development of technology. Space and stage use must also be considered. The writer expects that the lontara application with hypertext-based can be utilized by both tribes Bugis and Makassar, in expressing their ideas using the local language. Of course, local characters such as the

lontara application should support the speakers' wishes. In other words, the local script's character should be integrated with the development of information technology.

With this lontara application, the researchers hope that Bugis-Makassar speakers and other Lontara script users can communicate using this application, either when sending messages via email or Facebook or creating interactive applications. Of course, it is expected that the younger generation of Bugis-Makassar knows and utilizes this Lontara application. Only by using the Lontara script, both oral and written can be maintained in the writers' view. Thus, on the other hand, there is a local language sustainability action in place in this country expressing their ideas using the local language.

V. Conclusion

The lontara application model has been created. This application has been tested on a one-to-one test, small groups test, and large groups test. The participants from groups were very enthusiastic and were happy with this lontara application. This application helps them in teaching the local language. Also, they find it easy to prepare lontara script learning materials through this application.

The next step is launching this application for lontara users. The researchers recommend that the government issue a policy to implement learning lontara manuscripts in elementary to high school officially. Policies in the form of regulations will be a guide for local language teachers. This policy can be called a real, maintaining the local wisdom of South Sulawesi. Optimizing the use of this application is expected to be one of the concrete actions to maintain the local language in South Sulawesi.

Acknowledgment

This study is supported by the Hasanuddin University research grant scheme of BMIS number 2018.

References

[1] BPS, “Kewarganegaraan, Suku Bangsa, Agama, dan Bahasa Sehari-hari Penduduk Indonesia; Hasil Sensus Penduduk 2010,” Badan Statistik Indonesia, 2010. [Online]. Available: https://www.bps.go.id/publication/2012/05/23/55eca38b7fe0830834605b35/kewarganegaraan-suku-bangsa-agama-dan-bahasa-sehari-hari-penduduk-indonesia.html. [Accessed: 28-Oct-2020].

[2] C. D. F. Eberhard, David M., Gary F. Simons, “Indonesia,” Dallas, Texas: SIL International, 2020. [Online]. Available: https://www.ethnologue.com/country/ID. [Accessed: 24-Dec-2020].

[3] A. A. Ghani, The Teaching of Indigenous Orang Asli Language in Peninsular Malaysia, vol. 208, no. Icllic 2014. Elsevier B.V., 2015.

[4] G. F. Simons, S. I. L. International, M. Paul, and L. Sil, “The world ’ s languages in crisis : A 20-year update A 20-year update,” 26th Linguist. Symp. Lang. Death, Endangerment, Doc. Revital., no. June, 2012.

[5] UNESCO, “Towards UNESCO guidelines on Language Policies: a Tool for Language Assessment and Planning,” UNESCO, 2012. [Online]. Available: http://www.unesco.org/new/en/communication-and-information/resources/news-and-in-focus-articles/allnews/news/towards_unesco_guidelines_on_language_policies-1/. [Accessed: 28-Oct-2020].

[6] M. David and G. F. Simons, “Makasar in the Language Cloud,” 2020. [Online]. Available: https://www.ethnologue.com/cloud/mak. [Accessed: 23-Dec-2020].

[7] B. Marshall, “Local Language Vitality in Indonesia; Assessing and Intervening using Makassarese as a case study,” 2018.

[8] K. Gao, “Assessing the Linguistic Vitality of Miqie: An Endangered Ngwi (Loloish) Language of Yunnan, China,” vol. 9, pp. 164–191, 2015.

[9] D. Gumina, “Language power and hierarchy: Multilingual education in China,” Int. Multiling. Res. J., 2018.

[10] G. M. H. Kim, “Practicing Multilingual Identities: Online Interactions in a Korean Dramas Forum,” Int. Multiling. Res. J., 2016.

[11] G. F. Simons and M. P. Lewis, “The world’s languages in crisis,” no. January 2013, pp. 3–20, 2013.

[12] G. F. Simons, “Two centuries of spreading language loss,” Proc. Linguist. Soc. Am., vol. 4, no. 1, p. 27, 2019.

[13] C. D. F. Eberhard, David M., Gary F. Simons, “How many languages are there in the world?,” Dallas, Texas: SIL International, 2020. [Online]. Available: https://www.ethnologue.com/guides/how-many-languages. [Accessed: 24-Dec-2020].

[14] M. G. Delavan, V. E. Valdez, and J. A. Freire, “Language as Whose Resource?: When Global Economics Usurp the Local Equity Potentials of Dual Language Education,” Int. Multiling. Res. J., 2017.

[15] M. Brenzinger, A. Yamamoto, and N. Aikawa, “Language vitality and endangerment,” /ich/doc/src/00120-en. pdf., 2003.

[16] M. Brenzinger, “Language Maintenance and Shift,” in Encyclopedia of Language & Linguistics, 2006.

[17] D. I. Odugu, “Antinomies of Ideologies and Situationality of Education Language Politics in Multilingual Contexts,” Int. Multiling. Res. J., 2015.

[18] I. C. Pérez, “Indigenous Languages, Identity and Legal Framework in Latin America: An Ecolinguistic Approach1,” Procedia - Soc. Behav. Sci., 2015.

[19] S. W. Y. Lo-Philip, “Towards a theoretical framework of heritage language literacy and identity processes,” Linguist. Educ., 2010.

[20] D. Cunliffe, “The market for Welsh language mobile applications – A developers’ perspective,” Telemat. Informatics, 2019.

[21] G. Valdés, “Latin@s and the Intergenerational Continuity of Spanish: The Challenges of Curricularizing Language,” Int. Multiling. Res. J., 2015.

[22] A. M. Nonaka, “Estimating size, scope, and membership of the speech/sign communities of undocumented indigenous/village sign languages: The Ban Khor case study,” Lang. Commun., 2009.

[23] W. K. Sisamouth and S. C. Lah, Attitudes towards Thai, Patani Malay, and English of Thai Undergraduates: A Case Study at Prince of Songkla University Pattani Campus, Thailand, vol. 208, no. Icllic 2014. Elsevier B.V., 2015.

[24] D. Naisbitt and J. Naisbitt, Mastering Megatrends: Understanding and Leveraging the Evolving New World. 2019.

[25] Y. S. Baso, “Model Aplikasi Aksara Lontara Berbasis HTML sebagai Salah Satu Solusi Pemertahanan Bahasa Daerah,” J. KATA, vol. 2, no. 1–12, 2018.

[26] Y. S. Baso, Model Inovatif Pemertahanan Bahasa Daerah, 1st ed. Makassar: Pusat Kajian Media, Sumber Belajar dan E-Learning LKPP UNHAS, 2017.

[27] Wikipedia, “Abu al-Aswad al-Du’ali,” 2020. [Online]. Available: https://en.wikipedia.org/wiki/Abu_al-Aswad_al-Du%27ali. [Accessed: 24-Dec-2020].

[28] H. Aliane, Z. Alimazighi, and M. A. Cherif, “Al-Khalil: The Arabic linguistic ontology project,” in Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 2010.

[29] N. Bennett, W. R. Borg, and M. D. Gall, “Educational Research: An Introduction,” Br. J. Educ. Stud., 1984.

[30] M. D. Gall, J. P. Gall, and W. R. Borg, “Educational Research: An Introduction, 8th Edition,” Educ. An Introd., 2006.

[31] W. Dick and L. Carey, The systematic design of instruction. 6th. 2015.

[32] G. Khorasani and L. Zeyun, “Implementation of Technology Acceptance Model (TAM) in Business Research on Web Based Learning System 113,” Int. J. Innov. Technol. Explor. Eng., 2014.

[33] D. Persico, S. Manca, and F. Pozzi, “Adapting the technology acceptance model to evaluate the innovative potential of e-learning systems,” Comput. Human Behav., 2014.

[34] V. Venkatesh and H. Bala, “Technology acceptance model 3 and a research agenda on interventions,” Decis. Sci., 2008.

[35] M. Chuttur, “Overview of the Technology Acceptance Model: Origins , Developments and Future Directions,” Sprouts Work. Pap. Inf. Syst., 2009.

[36] A. U. Jan and V. Contreras, “Technology acceptance model for the use of information technology in universities,” Comput. Human Behav., 2011.

[37] University of Victoria, “Hot Potatoes,” 2020. [Online]. Available: https://hotpot.uvic.ca/. [Accessed: 24-Dec-2020].

[38] Y. S. Baso, Aplikasi pembuat soal-soal interaktif pembelajaran bahasa, 1st ed. Makassar: Pusat Kajian Media dan Sumber Belajar LPMPP Universitas Hasanuddin, 2017.

Trend of Bootstrapping from 2009 to 2016

Paulin Boale Bomolo, Eugene Mbuyi Mukendi, Simon Ntumba Badibanga
Department of Computer Sciences and Mathematics
University of Kinshasa, Kinshasa, Democratic Republic of Congo

***Abstract*—The pedestal of fully homomorphic encryption is bootstrapping which allows unlimited processing on encrypted data. This technique is a bottleneck in the practicability of homomorphic encryption. From 2009 to 2016, the execution time of bootstrapping decreased from several hours to a few thousandths of a second for processing a logic gate on two encrypted bits. This paper makes a comparative study of the evolution of bootstrapping during the period. An implementation of multiplication on 16-bit integers on an Intel i7 architecture through three schemes whose libraries are respectively DGHV, FHEW and TFHE makes it possible to corroborate the trend that to date the best bootstrapping on bits is that of the TFHE which executes this processing in 29 seconds improving that of the FHEW 30 times despite the multiplication algorithm used.**

Keywords*—Bootstrapping; homomorphic encryption; binary multiplication; logic gates

I. Introduction

Encryption is said to be probabilistic if a plaintext message is encrypted in several ciphertexts. This feature is found by adding a random value during the encryption operation. It is said to be homomorphic if it allows performing processing on ciphertexts with corresponding results on plaintexts. If it performs only additions [1, 2], multiplications [3] or binary operations [4] it is called partial homomorphic otherwise if it performs additions and multiplications in a limited number then it is somewhat homomorphic [5, 6, 7]. On the other hand, if it performs processing on unlimited number it is said to be fully homomorphic [5, 6, 7, 8, 9, 10, 11].

Fully homomorphic encryption was a breakthrough made by Gentry [6] in his thesis in response to the conjecture state in [12]. This breakthrough is based on the bootstrapping technique that evaluates its own decryption circuit to refresh noise in the ciphertexts thus allowing an unlimited number of processing in the encrypted domain. Initially, it requires the squashing technique and an additional security assumption to reduce the complexity of the decryption circuit [9, 10, 11, 13, 14]. With the advent of fully homomorphic encryption schemes based on the difficult problem of LWE [15] which belongs a low-complexity decryption algorithm, squashing was eliminated in bootstrapping [8, 9,10, 11,13, 14].

The removal of squashing allows in [9,10] to use a second homomorphic encryption scheme in the homomorphic evaluation in the decryption algorithm. This consideration improved performance of the homomorphic processing of gate NAND on two-bits encrypted in less than a second [9]. This processing was improved to 13 milliseconds by [10, 11].

N-bits arithmetic operations such as multiplication or addition can be built from the universal gate NAND. While knowing that an addition or multiplication operation performed on encrypted bits can respectively multiple or raise to the power the noise by the number of operations. We seek to know in this paper whether the performance of bootstrapping on a multiplication on two encrypted integers of 16 bits through an implementation carried out with three libraries of the comparative schemas that each marked a period in this trend is in the same order of processing as on the encrypted bits.

Roadmap. Section II presents the literature review of the encryption scheme from gentry's breakthrough to the period under review. Section III establishes the criteria for comparison through bootstrapping with a focus on the concepts behind them and presents a comparative study based on the concepts between the relevant algorithms of each period. Section IV shows three multiplication algorithms on two 16-bit integers and a shifter. Finally, section V extends the bootstrapping processing of said algorithms to integers of 16 to perform homomorphic multiplication. Discussions close this comparison.

II. Literature Review

In 2009, Gentry published the first homomorphic encryption scheme based on ideals lattices. This scheme is characterized by too large parameters and additional security assumptions. Two schemes improved respectively the reduction of the size of encrypted keys and messages and the removal of the additional security assumptions [16, 17, 18, 19]. The majority of schemes in this category are unusable in everyday applications.

To facilitate an understanding of gentry blueprint, an integer-based homomorphic encryption scheme was published in 2010[7]. It is based on the difficult assumptions of Approximate Greatest Common Divisor [20]. Several integer-based schemes have been proposed to improve the efficiency of DGHV. These improvements could be achieved based on different variants of the AGCD security assumptions to reduce the size of the public key in security parameter and the expansion of this schema in [21, 22, 23,24 ,25].

The hardness of implementing homomorphic encryption schemes based on hard problems mentioned above have steered the research towards another security problem. Thus, the first complete homomorphic encryption scheme based its security on the assumptions of LWE that removes squashing was presented by Brakerski and Vaikuntanathan [13]. Said scheme is based on two procedures which are the relinearization and the reduction of the module or dimension. Relinearization is a technique that reduces the size of ciphertexts from $n^2$to $n + 1$. It starts from a quadratic function

in a secret key $s$to a linear function in a secret key $t$dependent on s. Reducing the module or dimension naturally reduces the complexity of the decryption function and also reduces the size of the digits. Then, Brakerski and Vaikuntanathan also proposed a version of their schema based on the assumptions RLWE [14,26,27]. Many homomorphic schemes by levels or complete was introduced [8, 27]. Each new scheme brings techniques aimed essentially at reducing the size of the parameters and increasing the multiplicative depth. But relinearization is still a bottleneck for the majority of these schemes.

Of all these schemes, [8] stood out. It relies on assumptions of approximate vectors to perform a homomorphic operation of gate NAND on encrypted messages. Encrypted messages are square matrices, addition and homomorphic multiplication is addition and multiplication matrix respectively. The author in [8] removes relinearization which is an expensive technique used in other LWE schemes in favor of the eigenvector approximation technique. The author in [9] uses a variant of the [8] to improve its bootstrapping based on gate NAND in less than a second. The author in [10, 11] introduces the external product in a variant of the [8] to reorganize bootstrapping of [9] and achieve 30 times better performance.

Our goal is to establish the criteria for comparing bootstrapping in three schemes which are the DGHV scheme [7], the Ducas Micciancio [9] scheme and the scheme in [10, 11]. In addition, use the libraries that implement them to corroborate the trend of bootstrapping in a 16-bit binary multiplication.

III. BOOTSTRAPPING

It is a technique that was introduced by Gentry [6] to solve the open problem stated in 1978 by [12] which consists in carrying out the processing on the encrypted data. Before Gentry's breakthrough, the noise increases with the circuit depth to be evaluated. The consequence is that decryption fails. The solution to this concern for inefficiency was through the encryption technique to reduce noise in the encrypted message and a homomorphic evaluation of the decryption algorithm.

A. The reencryption [6]

1) Definition of reencryption: reencryption is a noted function $\mathrm{Rencrypt}_\varepsilon$ that converts a message encrypted under a key $\mathrm{p_{k1}}$ into another message encrypted under an another key $\mathrm{p_{k2}}$ without revealing any information about the private key $\mathrm{s_{k1}}$ or the plain text m it is clear that $\mathrm{Rencrypt}_\varepsilon(\mathrm{c}) = \mathrm{Encrypt}_\varepsilon(\mathrm{p_{k2}}, \mathrm{m})$ where $\mathrm{c} = \mathrm{Encrypt}_\varepsilon(\mathrm{p_{k2}}, \mathrm{m})$ [2].

2) Reencryption algorithm: A reencryption can be evaluated in the following steps:

Generate a key pair (s_{ka}, p_{ka}) and (s_{kb}, p_{kb}) respectively belonging to A and B.

Evaluate A reencrypting key A between B and as follows $r_k = Encrypt_\varepsilon(p_{kb}, s_{ka})$.

Calculate a ciphertext $c = Encrypt_\varepsilon(p_{ka}, m)$ where m is plaintext.

Redefine the decryption function $Decrypt_\varepsilon$ as follows $f_c(s_{ka}) = Decrypt_\varepsilon(s_{ka}, c)$.

Evaluate the reencryption as follows $Evaluate_\varepsilon(p_{kb}, f_c, r_k) = Encrypt_\varepsilon\big(p_{kb}, f_c(s_{ka})\big) = Encrypt_\varepsilon(p_{kb}, m)$.

Reencryption allows to A to designate B by giving it the ability to encrypt the plaintext that it has encrypted with another key. B is called proxy.

3) One-way reencryption: Given a pair of keys (s_{k1}, p_{k1}) and (s_{k2}, p_{k2}). A one-way reencryption is a conversion from $Encrypt_\varepsilon(m, p_{k1})$ to $Encrypt_\varepsilon(m, p_{k2})$ by the evaluation of $Evaluate_\varepsilon(p_{k2}, f_c, r_k)$ where $r_k = Encrypt_\varepsilon(p_{k2}, s_{k1})$, not the inverse $Encrypt_\varepsilon(m, p_{k1}) = Evaluate_\varepsilon(p_{k1}, f_c, r_k)$ where $Encrypt_\varepsilon(m, p_{k2})$ and $r_k = Encrypt_\varepsilon(p_{k1}, s_{k2})$. The reciprocal of this assertion is false.

B. Hard Problems of Homomorphic Encryption

1) The problem of learning with error: The Problem of Learning With Error (LWE) was introduced by Regev in 2005 [15]. The Ring version of this problem called RingLWE, was introduced by Lyubashevsky, Peikert and Regev in 2010[28]. All variants are widely used nowadays in the construction of lattices-based homomorphic encryption schemes.

a) The Regev problem: For a security setting λ, Either $\mathrm{n} = \mathrm{n}(\lambda)$an integer dimension, an integer $\mathrm{q} = \mathrm{q}(\lambda) \geq 2$, and a distribution $\chi = \chi(\lambda)$ under $\mathbb{Z}$. The hard problem of $\mathrm{LWE}_{\mathrm{n,q},\lambda}$ the is to distinguish two following distributions:

In the first distribution, the sample $(\overrightarrow{a_i}, b_i)$ drawn uniformly from $\mathbb{Z}_q^{n+1}$;

In a second distribution, draw $\vec{s} \leftarrow \mathbb{Z}_q^{n+1}$and $(\overrightarrow{a_i}, b_i) \in \mathbb{Z}_q^n$ then a sample by drawing uniformly $\overrightarrow{a_i} \leftarrow \mathbb{Z}_q^n$ and $e_i \leftarrow \chi$ respectively, and initializing $b = \langle \overrightarrow{a_i}, \vec{s}_i \rangle + e_i$. The assumptions of $LWE_{n,q,\chi}$ are such that the problem of $LWE_{n,q,\chi}$ is hard.

b) The RLWE problem: For a secret $s \in R_q$, the RLWE distribution under $R_q \times R$ is drawn by respectively a uniform and random $a \in R_q$ and $e \leftarrow \chi$, and gives the output expression $(a, b = \langle s, a \rangle) + e \bmod q$.

RLWE is said to be decisional if given m independent samples $(a_i, b_i) \in R_q \times R_q$ where each sample is distributed either $A_{s,\chi}$ for random and uniform $s \in R_q$ (fixed for all samples) or the uniform distribution, distinguish which is the case (with a significant probability).

c) The General Problem of the LWE (GLWE) [26]: For a security parameter λ, that is $f(x) = x^d + 1$, where $d = d(\lambda)$ is a power of 2. Let be two integers respectively the modulus $q = q(\lambda)$ and the dimension n, let $R = \mathbb{Z}[x]/f(x)$ and $R_q = R/qR$. Let be $\chi = \chi(\lambda)$ a distribution on R. The problem with GAAE is to distinguish between the following two distributions:

The first distribution is a uniform sample $(a_i, b_i) \in R_q^{n+1}$;

In the second uniformly drawn distribution $a_i \leftarrow R_q^n$, $s \leftarrow R_q^n$ and $e_i \leftarrow \chi$, the second distribution is the sample $(a_i, b_i) \in R_q^{n+1}$ where $b_i = \langle a_i, s \rangle + e_i$. The assumption of GLWE is that the GLWE problem is hard.

If $d = 1$ then the LWE problem is that of the LWE problem. If n = 1 then the GLWE problem is that of the RLWE problem.

The author in [8] based on assumptions of LWE, it constructs its schema with a plaintext space $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, an encrypted message space $\mathbb{Z}_q$ with an error or noise $E < \frac{q}{16}$ where q is the modulus that determines the key space from which the secret key s is taken and n is the encrypted message dimension.

To encrypt a plaintext $m \in \mathbb{Z}_2 \subset \mathbb{Z}_4$, draw $a \leftarrow \mathbb{Z}_q^n$, $e \leftarrow \chi$ and output the ciphertext c as follows $LWE_s^{4/q}\left(m, \frac{q}{16}\right) = \left(a, a.s + \frac{2m}{4} + e\right) \in \mathbb{Z}_q^{n+1}$.

The authors in [10, 11] redefines the problem of LWE and RLWE on the real torus $T = \mathbb{R} \, mod \, 1$ and the torus of polynomials $T[X] = T[X] \, mod \, X^N + 1$ respectively. This redefinition produces three types of ciphertexts for this schema. It also generalized and improved the encryption scheme based on the [8] and several of its variants.

To encrypt a plaintext $m \in T$, pick a secret key $s \in \mathcal{B}^n = \mathbb{Z}_2^n$ and calculate $c = (a, b) \in T^{n+1}$ where $a \in T^n$ is a random mask, $b = a.s + \varphi$ and $\varphi = e + m$ where e is a parameter that is drawn in a Gaussian distribution.

To encrypt the plaintext $m \in T_N[X]$, draw a key $s \in \mathcal{B}_N[X]$ and calculate $c = (a, b) \in T_N[X]^2$ where a is a random mask and $b = s.a + e + m$ where $e \in T_N[X]$.

To encrypt the plaintext $m \in \mathbb{Z}_N[X]$, pick the secret key $s \in \mathcal{B}_N[X]$ as in the RLWE and calculate $c = Z + m.G_2 \in T_N[X]^{2l \times 2}$ where Z is a list of ciphertexts of type RLWE of 0 and G_2 is the matrix with $\begin{pmatrix} g & 0 \\ 0 & g \end{pmatrix} g^T = (2^{-1}, \dots\dots\dots\dots\dots\dots\dots, 2^{-l})$.

2) The problem of the Approximation of the Greatest Common Divisor (AGCD)[20, 29].

The AGCD's problem with the parameters (γ, η, ρ) is the problem of finding the secret integer p given several samples $x_i = pq_i + r_i$ of arbitrarily provided where:

The secret integer p has bits η;

The terms noises r_i are uniform samples from the interval $[-2^\rho + 1, 2^\rho - 1] \cap \mathbb{Z}$;

The terms q_i are uniform samples of $[0, 2^{\gamma-\eta}] \cap \mathbb{Z}$.

[7] is the first known scheme applying the AGCD problem in cryptography to produce a homomorphic encryption scheme. In its symmetric version, it encrypts the plaintext $m \in \{0, 1\}$, two random integers are drawn uniformly to evaluate the encrypted message as follows $c = pq + 2r + m$. In other words, a sample of AGCD is calculated by adding the even noise $2r$ to the product pq which is added to m.

C. Bootstrapping [6]

1) Fundamental properties: In Gentry construction, bootstrapping is based on three fundamental properties that belong a partial or somewhat homomorphic encryption scheme that make it fully homomorphic encryption. These properties are listed and noticed below:

The complexity of the decryption algorithm is greater than that of the circuits to be evaluated. Given d the maximum degree of the decryption algorithm $Decrypt_\mathcal{E}$ and p the maximum degree of the function or polynomial to be evaluated by scheme. If $d < p$ then the decryption algorithm $Decrypt_\varepsilon$ is been useful in homomorphic evaluations. If $d > p$ then the complexity of this algorithm is reduced to $Decrypt'_\varepsilon$ for homomorphic evaluations hence $f_c(s_k) = Decrypt'_\varepsilon(s_k, c)$ where $c = Encrypt_\varepsilon(p_k, m)$.

Bootstrappability is a critical property of an encryption scheme that allows you to homomorphically evaluate your own decryption algorithm under an encrypted decryption key. Given an encryption scheme $\mathcal{E}$, $\mathcal{E}$ is said to be bootstrappable if $Evaluate_\mathcal{E} = (p_{k1}, f_c, e_k) = Encrypt_\mathcal{E}(p_{k1}, m)$ where $f_c(s_k) = Decrypt_\mathcal{E}(c, s_k)$, $c = Encrypt_\mathcal{E}(m, p_k)$ and $e_k = Encrypt_\mathcal{E}(p_{k1}, s_k)$ it is obvious that $\mathcal{E}$ evaluates homomorphically its decryption algorithm.

Circular security is a property that an asymmetric (symmetric) encryption scheme has to encrypt one's private key securely (secretly) by its corresponding public (secret) key. A homomorphic encryption scheme $\mathcal{E}$ has the circular security property if for a couple of given keys, (s_k, p_k) the bootstrapping key is evaluated as follows $e_k = Encrypt_\mathcal{E}(s_k, p_k)$: it is obvious that the private key is securely encrypted by its public key.

2) Definition of bootstrapping: Bootstrapping is a technique for reducing noise in the ciphertext c and getting noise b' in a refreshed ciphertext c' such as $b' < b$ where $b' \supset Encrypt_\varepsilon(p_k, m) \leftarrow Evaluate_\varepsilon(p_k, f_c, e_k)$ and b is the original noise in the ciphertext c by the homomorphic evaluation its own decryption circuit $f_c(s_k) = Decrypt_\varepsilon(s_k, c)$ on a decryption key called bootstrapping key$e_k = Encrypt_\varepsilon(p_k, s_k)$.

3) Bootstrapping algorithm: Given two pairs of keys (p_{k1}, s_{k1}) and (p_{k2}, s_{k2}) generated by a homomorphic encryption scheme ε.

Let be two ciphertexts c_1 and c_2 evaluate as follows: $c_1 = Encrypt_\varepsilon(p_{k1}, m_1)$ and $c_2 = Encrypt_\varepsilon(p_{k1}, m_2)$ where $m_1 \, and \, m_2$ are plaintexts.

The bootstrapping key e_k is calculated as follows $e_k = Encrypt_\varepsilon(p_{k2}, s_{k1})$. And the decryption function $Decrypt_\varepsilon$ is redefined in the following way $f_{c_1,c_2}(s_k) = NONET\big(Decrypt_\varepsilon(s_k, c_1), Decrypt_\varepsilon(s_k, c_2)\big)$ where is the private key s_k.

A homomorphic evaluation of f_{c_1,c_2} on c_1 *and* c_2 is carried out as follows: $Evaluate_{\varepsilon}(p_{k2}, f_{c_1,c_2}, e_k) = Encrypt_{\varepsilon}\left(p_{k2}, NONET\left(Decrypt_{\varepsilon}(s_{k1}, c_1), Decrypt_{\varepsilon}(s_{k1}, c_2)\right)\right) = Encrypt_{\varepsilon}(p_{k2}, NONET(m_1, m_2)) = NONET(Encrypt_{\varepsilon}(p_{k2}, m_1), Encrypt_{\varepsilon}(pk_2, m_2)) = NONET(c_1', c_2')$ where c_1' *and* c_2' are refreshed ciphertexts of c_1 *and* $c_2$whose noise $b' \ll b$.

4) Squashing: Squashing is a procedure that consists of expressing the decryption algorithm $Decrypt_{\varepsilon}$ into a polynomial or function $p_c(s_k)$ whose variables are the ciphertext c and the secret key sk. $p_c(s_k)$ is equivalent to a shallow circuit.

In [3], the decryption algorithm is expressed by the function $c^d \bmod N$. The complexity of the operation of exponentiation does not make it possible to rewrite this function into an equivalent function of low degree.

In the [7], the decryption algorithm is expressed by the expression $c \bmod p \bmod 2$ (1) which is not a low complexity. To do this, it is transformed into a circuit of expression $[c]_2 \oplus \left[\left\lfloor c.\left({}^{1}/_{p}\right)\right\rfloor\right]_2$ (2). ${}^{1}/_{p}$ is replaced in the evaluation (2) by the expression $\sum_{i=1}^{\Theta} s_i z_i$ which represents the sum of the subsets where $s_i = {}^{u_i}/_{2^{\kappa}}$. Evaluation (1) becomes $\left[c - \sum_{i=1}^{\Theta} s_i z_i\right]_2$ (3). (3) is the equivalent function of (1). (3) is an expression that has a low complexity.

5) Concept of Bootstrapping from 2015 [8 9, 10, 11]: Bootstrapping of scheme based on the problem of assumptions of LWE and its variants removes squashing. The decryption algorithm has a complexity that allows it to be evaluated homomorphically in the reencryption. This reencryption is carried out by a homomorphic accumulator which makes it possible to refresh the encrypted message into an equivalent encrypted message containing a small noise.

A homomorphic accumulator is a quadruplet of algorithms $Encrypt_{\varepsilon}$, Init, Incr and msbExtract. $Encrypt_{\varepsilon}$ is an encryption scheme that uses a key and is different from the first. It is called an internal scheme.

Init is the algorithm that initializes the contents of the accumulator. More briefly, this operation is written as follows: $ACC \leftarrow Encrypt_{\varepsilon}, (v)\ pour\ ACC \leftarrow Init(Encrypt_{\varepsilon}, (v))$.

Incr is the algorithm that allows you to add a value to the contents of the accumulator. This operation is written as follows: $ACC \overset{+}{\leftarrow} Encrypt_{\varepsilon}$ for $, (v) Incr\left(ACC,\ Encrypt_{\varepsilon}, (v)\right)$.

$msbExtract$ calculates with high probability from the contents of the homomorphic accumulator to produce a valid number. This operation is summarized by the expression $c \leftarrow msbExtract(ACC)$ with $c \in LWE_s^{t/q}\left(msb(v), e(l)\right)$ where e is the noise.

6) Type of bootstrapping: There are two types of bootstrapping that bootstrapping by squashing or by homomorphic accumulator.

A bootstrapping is said by squashing if a new security assumption is added in the reduction of the complexity of the decryption algorithm to ensure optimal security in the encryption scheme during the refresh of the noisy message.

Refreshing the ciphertext c with the addition of the assumption of the sum of subsets to re-encrypt c using the encrypted secret key $\sum_{i=1}^{\Theta} s_i z_i$ of $\frac{1}{p}$ which is used to obtain the ciphertext $c^* = \left[c - \sum_{i=1}^{\Theta} s_i z_i\right]_2$ [7].

A bootstrapping is said by homomorphic accumulator if a homomorphic accumulator is used to refresh a ciphertext in the reencryption operation.

a) Homomorphic accumulator in [9]: In [9], the homomorphic accumulator is based on the encryption scheme [8] defined under the assumptions of the Ring LWE. Let be a message m and the key $z \in \mathbb{Z}$, $Encrypt_{\varepsilon_z}(m)$ encrypts as described below:

Pick Randomly and uniformly the vector $a \in \mathcal{R}_Q^{2dg}$ and $e \in \mathcal{R}_Q^{2dg}$ into a Gaussian distribution χ of parameter ζ where $\mathcal{R}_Q^{2dg} = \mathbb{Z}_Q^{2dg}, N = 2^k$;

Calculate $Encrypt_{\varepsilon,z}(m) = [a, a.z + e] + uY^m G\ de\ \mathcal{R}_Q^{2dg\times 2}$ where m is encoded as the root of the unit $Y^m \in \mathcal{R} = \frac{\mathbb{Z}_N}{X^N+1}$ of where $N = 2^k$.

To upload the accumulator with the ciphertext $v \in \mathbb{Z}_q$, the function $Init(ACC \leftarrow v)$ uploads the content of accumulator with v as follows $ACC := uY^v G\ de\ \mathcal{R}_Q^{2dg\times 2}$;

To add an ciphertext to the contents of the accumulator, a decomposition of $u^{-1}.ACC$ in the base B_{dg} is performed as follows: $u^{-1}.ACC = \sum_{i=1}^{d_g} B_g^{i-1} D_i$ where the $D_i \in \mathcal{R}^{2dg\times 2}$ with the coefficients $\left\{\frac{1-B_g}{2}, \ldots\ldots, \frac{B_g-1}{2}\right\}$ and then perform $Incr(ACC \overset{+}{\leftarrow} C)$ where $ACC, C \in \mathcal{R}_Q^{2dg\times 2}$ to output $ACC := \left[D_1 \ldots\ldots D_{dg}\right]$.

Finally, use the msbExtract function with two entries that are a switch key $\mathfrak{R}$, a test vector $t = -\sum_{i=0}^{{}^{q}/_{2}-1} \vec{Y}_i$ to find $c \in LWE_s^{4/q}\left(m, \frac{q}{16}\right)$.

In [10], bootstrapping by accumulator is performed on the one-bit encrypted message $m \in \mathcal{B}$, $(a, b) \in T^n \times T = LWE_q^s(m, e)$ where $\mathcal{B} = \{0, 1\}$ and $e < \frac{1}{4}$ for valid decryption. Said message is first rounded to $\left(\bar{a}, \bar{b}\right) \in \mathbb{Z}_{2N}^n \times \mathbb{Z}_{2N}$ where $\bar{b} = \lceil 2Nb \rfloor$ and $\overline{a_i} = \lceil 2Na_i \rfloor$.

Given a test vector $testv = (1 + X + \cdots\ldots\ldots + X^{N-1}).X^{N/2}.u'$ where $u' = \frac{m}{4} \in T$, the result of the expression $X^{\bar{b}}.(0, testv)$ is loaded into ACC: $ACC \leftarrow (0, X^{-b}, testv)$. The evaluation of the

expression $[h + (X^{-\bar{a}_i} - 1)] \odot ACC$ update the content of ACC: $ACC \leftarrow X^{\bar{b}-\bar{a}s}.testv$.

An extraction is performed with the function $SampleExtract$ that receives as input the contents of: $ACCX^{\bar{b}-\bar{a}s}.testv$. It extracts the terms of said polynomial in a sample $msg((a', b'))$ where $(a', b') = (coefs(a''(X)), b'') \in T^n \times T$ where $coefs(a''(X))$ is the coefficient of the vector $a'' \in T_N[X]$ and $b_0'' \in T$ is the constant term of the polynomial $b'' \in T_N[X]$.

Key switching allows you to find a sample $TLWE(a, b) \in T^n \times T$ of the message $\frac{m}{2} \in T$ under the secret key s. It receives as input the result of the expression $msg(u) = u' + msg(SampleExtract(ACC))$.

7) Processing bootstrapping [9, 10]: There are two types of processing bootstrapping which are logic gate bootstrapping and logic circuit bootstrapping.

It is said that a homomorphic encryption scheme supports logic gate processing bootstrapping if a refresh is performed after each logic gate it is obvious that $Evaluate_\varepsilon(p_k, f_c, e_k)$ where f_c is a logic gate of the type AND, OR, NOT, ...

In [8], the homomorphic NAND gate is defined by $HomNAND: LWE_s^{4/q}(m_0, {}^{q}/_{16}) \times LWE_s^{4/q}(m_1, {}^{q}/_{16}) \rightarrow LWE_s^{2/q}(m_0 \overline{\wedge} m_1)$ where $m_0 \overline{\wedge} m_1 = 1 - m_0 m_1$ and $c_i = LWE_s^{4/q}(m_i, {}^{q}/_{16})$ with $i \in \{0, 1\}$. The refresh is performed on the result as follows: $LWE_s^{2/q}(m, {}^{q}/_{4}) \longrightarrow LWE_s^{4/q}(m, {}^{q}/_{16})$.

It is said that a homomorphic encryption scheme supports circuit processing bootstrapping if a refresh is performed after each logic circuit it is obvious that $Evaluate_\varepsilon(p_k, f_c, e_k)$ where f_c is a circuit that includes more than one logic gate of the type AND, OR, NOT.

In [8], let be a circuit for calculating the retention in an n-bit adder of two numbers a and b and an incoming retention c_0= 0, the expression (2). $c_i = (a_i \oplus c_{i-1}).(b_i \oplus c_{i-1}) \oplus c_{i-1}$ where $\oplus$ is XOR logic gate. From the Table I which is table of truth below of this expression a bootstrapping by circuit can be performed from a function majority with three variables noted $Maj(m_1, m_2 m_3)$ gives a value equal to 1 if the majority of bits is 1 otherwise 0.

Specifically, given three encrypted messages; Expression (1) can evaluate these three ciphertexts and produce a resulting ciphertext. Being calculated modulo 4, this makes it possible to homomorphically process the majority function described above. $c_1, c_2\ et\ c_3$.

This addition modulo 4 of the encrypted messages makes it possible to find the encrypted $Encrypt_\varepsilon(m)$, $m \in \{2,3\}$ with if majority is equal to 1 or if $m \in \{0,1\}$ the majority is equal to 0. An affine transform of $\frac{9q}{8}$ is performed to find the majority function in $\mathbb{Z}_4$. The circuit retained out of three is illustrated with the majority function noted maj as follows:

$$Maj\left(LWE_s^{4/q}(m_0, {}^{q}/_{16}), LWE_s^{4/q}(m_1, {}^{q}/_{16}), LWE_s^{4/q}(m_2, {}^{q}/_{16})\right) \rightarrow LWE_s^{2/q}(m, {}^{q}/_{4}).$$

The refresh is carried out on the result of the circuit as follows: $LWE_s^{2/q}(m, {}^{q}/_{4}) \longrightarrow LWE_s^{4/q}(m, {}^{q}/_{16})$.

8) Bootstrapping: Analysis and comparison of algorithms: Table II shows that TFHE bootstrapping performs better than bootstrapping performed and executed in the other two schemes. This fact is due to the removal of the decomposition step in any basis of the vector a of assumptions LWE [10, 11].

TABLE I. TRUTH TABLE OF THE SELECTED FUNCTION OF THE THREE BITS

a_i	b_i	c_{i-1}	$a_i \oplus c_{i-1}$(1)	$b_i \oplus c_{i-1}$(2)	1 ⊕ 2(3)	3 ⊕ c_{i-1}	Maj
1	1	1	0	0	0	1	3
1	1	0	1	1	1	1	2
1	0	1	0	1	0	1	2
1	0	0	1	0	0	0	1
0	1	1	1	0	0	1	2
0	1	0	0	1	0	0	1
0	0	1	1	1	1	0	1
0	0	0	0	0	0	0	0

TABLE II. ANALYSIS AND COMPARISON OF ALGORITHMS

Encryption scheme	Type of homomorphy	hard problem	Type of bootstrapping	Homomorphic operations	Squashing	Security parameter size (bits)	Complexity of the decryption algorithm	Bootstrapping key size	Bootstrapping execution time(s)
DGHV	fully	AGCD	By squashing	$+\ et\ \times$	Yes	72	great	NA	660
FHEW	fully	LWE RLWE	By accumulator	NAND	No	88	low	2.4 GB – 1 GB	0.63
TFHE	fully Leveled	TLWE	By accumulator	NAND, AND, ...	No	110	low	24 MB	0.052 0.0013

IV. Applications: Binary Multiplication

The operation of multiplying two integers is described in Fig. 1, for any calculation basis (binary, decimal, etc.) by the following two steps:

The calculation of partial products;

The sum of the partial products obtained.

The product of two numbers of n digits can be given by a number of $2n$ digits. In the binary system, the gate AND is used to generate the partial products $a_i b_i$ between each bit of the two multiplicandes. A binary addition is performed on each column of partial products.

				a_3	a_2	a_1	a_0
				b_3	b_2	b_1	b_0
				a_3b_0	a_2b_0	a_1b_0	a_0b_0
			a_3b_1	a_2b_1	a_1b_1	a_0b_1	
		a_3b_2	a_2b_2	a_1b_2	a_0b_2		
	a_3b_3	a_2b_3	a_1b_3	a_0b_3			
p_7	p_6	p_5	p_4	p_3	p_2	p_1	p_0

Fig. 1. Example of Multiplying Two Numbers at 4 Bits.

A. The Classic Multiplication Algorithm[30]

Let a and b be two numbers of k bits, expressed as a basis:$\beta = 2$

$$a = (a_{n-1}a_{n-2} \dots \dots \dots \dots a_0) = \sum_{i=0}^{n-1} a_i \beta^i \quad (1)$$

$$b = (b_{n-1}b_{n-2} \dots \dots \dots \dots b_0) = \sum_{i=0}^{n-1} b_i \beta^i \quad (2)$$

Where the and a_i are b_i in the interval $[0, 1]$. The classical algorithm of multiplication of a and b consists in calculating partial products by multiplying the b_i of the multiplier by b the whole number a and then adding these partial products in order to obtain the final product p which is a number of $2n$ bits.

Note p_{ij} the pair (carry, Sum) obtained from the partial product $a_i b_j$. Fig. 1 illustrates the results p_{ij} of multiplication of a and b at 4 bits.

The last rank denotes the total sum of the partial products which is also the product $a\ by\ b$ represented by a number of $2k$ bits.

Algorithm 1: Classical Multiplication MC

Input: a, b
Output: p = ab
Initialize $p_i \coloneqq 0$ for $i = 0, 1, \dots \dots, 2n-1$
for $i = 0\ to\ n-1$
r
for $j \coloneqq 0\ to\ n-1$
$(r, s) = p_{i+j} + b_i a_j + r$
$p_{i+j} = s$
End For
$p_{i+n} \coloneqq r$
End for
Return $(p_{2n-1}p_{2n-2} \dots \dots \dots \dots p_0)$

This algorithm requires $O(n^2)$ bit-level operations to multiply two n bit encrypted numbers.

B. Horner's Algorithm

It was originally introduced to effectively evaluate the value of a polynomial $p(x) = \sum_{i=0}^{n} a_i x^i$ for a given value α. It is based on the following rewrite:

$$p(x) = a_0 + a_1 x + \cdots \dots + a_n x^n \quad (3)$$

$$= a_0 + x\left(a_1 + x\left(a_2 + \cdots \dots \dots \dots . . + x\left(a_{n-1} + x(a_n)\right) \dots \dots .\right)\right) \quad (4)$$

The expressions below evaluate a polynomial $p(x)$ at a given point α by performing n multiplications and n additions to calculate $p(\alpha)$.

$$ab = a.\sum_{i=0}^{n-1} b_i 2^i = ab_0 2\left(ab_1 + 2\left(ab_2 + \cdots . . + 2\left(ab_{n-2} + 2(ab_{n-1})\right) \dots \dots \dots \dots\right)\right) \quad (5)$$

The equation below can be written in the following recursive form:

$$p_0 = 0$$

$$p_i = 2p_{i-1} + b_{i-1}a \quad (6)$$

From these equations, Horner's algorithm (2) for multiplying binary integers is written as follows:

Input: $a_0, a_1, \dots \dots \dots, a_{n-1}$ and $b_0, b_1, \dots \dots \dots \dots \dots \dots, b_{n-1}$
Output : $p = ab$
$p_0 \coloneqq 0$
For $i = n - 1\ to\ 0$
Do
$p_i \coloneqq 2p_{i-1} + b_{i-1}a$
End do
End for
Return p

This algorithm has the same complexity as the classical multiplication algorithm is $O(n^2)$.

C. Karatsuba's Algorithm

The Karatsuba algorithm is a recursive algorithm introduced by the Russian mathematician Karatsuba in 1962. This algorithm requires $O\left(n^{\log_2 3}\right)$ to multiply two numbers of n bits. Its complexity is reduced by method of the divide-and-conqueror which uses fewer multiplications than the classical algorithm.

Let a and b be two integers of n bits and $l = \lceil n/2 \rceil$. Karatsuba initially breaks down $a\ and\ b$ into two equal parts:

$$a = 2^l a_1 + a_0\ , b = 2^l b_1 + b_0 \quad (7)$$

Such as a_1 is the l high-weight bits of a and a_0 is the l low-weight bits of a. Note that the 2^lvalue thus constitutes the basis of the representation β.

1) Naïve recursion method: The naïve recursion method reduces the multiplication of a and b multiplication of their components $a_1, a_0, b_1\ et\ b_0$ including the size of the initial integers as shown in the following equation:

$$p = a.b = (2^l a_1 + a_0)(2^l b_1 + b_0)$$

$$= 2^{2l}(a_1b_1) + 2^l(a_1b_0 + a_0b_1) + a_0b_0$$

$$= 2^{2l}p_2 + 2^lp_1 + p_0 \quad (8)$$

Said formulation reveals that the multiplication of two numbers of k bits require 4 multiplications of $l = \frac{k}{2}$ bits. Its complexity is not far from that of a classical algorithm.

2) Karatsuba algorithm: Its algorithm improves the performance of equations in (1). By reducing the number of multiplications to three but adding four additional additions. A rearrangement of the terms of the product $p = a.b$ makes it possible to obtain:

$$p_0 = a_0b_0 (9)$$

$$p_1 = (a_0 + a_1)(b_0 + b_1) - p_0 - p_2 \quad (10)$$

$$p_2 = a_1b_1 \quad (11)$$

Of these equations, a remark is made of the presence of three multiplications, two bits n and $n + 1$ one bit. The karatsuba algorithm requires $O(n^{1.59})$ operations to give the product of two numbers.

Algorithm 3: Karatsuba multiplication.MK

Input: a, b, k
Output : $p = a.b$
If (is small) then k
Return: Call the classic algorithm.$MC(a, b)$
Finsi
$l := k/2$
$a_0 := a/2^l$
$a_1 := a \ mod \ 2^l$
$b_0 := b/2^l$
$b_1 := b \ mod \ 2^l$
$p_0 := MC(a_0, b_0)$
$p_1 := MC(a_1, b_1)$
$temp := MC(a_0 + a_1, b_0 + b_1)$
$p_1 := temp - p_0 - p_2$
Return $2^{2l}p_2 + 2^lp_1 + p_0$

The version of the algorithm that has been implemented in this paper is iterative. It performs operations on 8-bit encrypted integers.

D. The Shifter

A shifter is formed of n + 1 inputs d_1, d_2, ..., d_n, c and n outputs s_1, s_2, ..., s_n and operates an offset of 1 bit on the inputs if c = 1, it is an offset to the right and if c = 0 then it is an offset to the left.

Algorithm 4: shifting to left or right.

```
Input: a: n-bit encrypted integer, right or left Boolean: offset direction
Positions: Number of offset positions
b: encrypted integer shifted by offset over n bits of positions.
cx1, cx2 two null encrypted integers of n bits
i: integer counter
flag: A Boolean integer that determines the offset direction.
if flag = 0 then
 right = 1;
 left = no(right)
otherwise
 right = 0;
 left = no(right)
finsi
for i of 1 to positions
do
 for k from 0 to n – 1
do
 if k > 0 and k < n – 1 then
 cx1k = and(ak-1, left);
 cx2k = and(ak+1, right);
 bk = or(cx1k , cxk2);
 otherwise
 if k = 0 then
 bk = and(ak+1, right)
 finsi
 if k == n then
 bk = and(ak-1, left);
 finsi
 finish
end do
return b
```

Algorithm 4 has a complexity of $O(p \times n)$ where p is the number of offset positions and n is the bit size of the number to be shifted.

V. Implementation and Interpretation of Results

A. Implementation

The implementations were tested on the Intel® core™ i7-5500 CPU @2.4 GHZ processor of a laptop with a cache memory of 4019 kilobytes, a clock clock clock of 1100 MHZ and a volatile memory of 8 Gigabytes that supports extensions of the following instruction sets: MMX, SSE, SSE2, SSE4_1, SSE4_2, FMA, AVX and AVX2.

The DGHV code was implemented in Python with Sage and GMP (GNU Multi Precision). These two libraries provide machine compiled mathematical libraries that are fast in their executions. We have not been optimal to work with these tools in the implementation of multiplication.

The FHEW library that is written in C/C++ language. An optimization to quickly perform convolution was achieved by an implementation of the Fourier transform FFTW3 to process bootstrapping. Functions useful for performing multiplication have been added to the FHEW.cpp source file [31].

The TFHE library is written in C/C++ language and an optimization has been implemented for the fast processing of bootstrapping with the data parallelism of fused-multiply add and as an Advanced Vector eXtensions assembler through a SQLIOS fast Fourier transform parameterized in either AVX or FMA. Useful functions have been added to the cloud file.c and alice.c [32].

Synthesis and comparison:

In Table III, the columns represent the circuit type used in the implementation of multiplication operations and the type of logic gates. As for the rows, they represent the implementation of different types of multiplication. The intersection between the row and the column gives the number of circuits or gates implemented to achieve each type of multiplication.

TABLE III. Circuit used in each Type of Multiplication

	Adder	Subtractor	Shifter	And	Multiply	Weighting
Horner	1	0	1	1	0	N
Classic	2	0	0	1	0	N^2
Karatsuba	4	4	4	1	3	1

The implementation of Karatsuba is less expensive in circuits and logic gates than the other two implementations are about three offsets respectively of 8 bits on 8 bits and 16 bits on 16 bits, two subtractors on 8 bits which represents the modulo 2^8, four additions respectively two on 8 bits and two on 16 bits and three multiplications on 8 bits. And on the other side, the classic implementation takes 512 complete additions on one bit and 256 multiplications with the door and on one bit. And in the same proportion as Horner's is 16 offsets of 1 bit by 16 bits, 256 multiplications on 1 bit with the door and 16 additions on 16 bits.

B. Interpretation of Results

In Table IV, the columns represent the implementation library and the rows represent the type of multiplication implemented. The intersection is the second execution time of a type of multiplication of two 16-bit numbers with one of the column libraries.

TABLE IV. PERFORMANCE TABLE OF MULTIPLICATION BY DGHV, FHEW OR TFHE

	DGHV	FHEW	TFHE
Horner	NA	671	41
Classic	NA	649	39
Karatsuba	NA	483	29

The library implemented for the DGHV did not provide results in a reasonable time to be taken into account in this paper. As for the FHEW and TFHE libraries, the theoretical results corroborated the theoretical hypotheses in memory and time complexity. It appears that the choice made in the design and implementation of the TFHE makes its bootstrapping more efficient.

VI. DISCUSSION

TFHE bootstrapping improves 15 times that of FHEW for this homomorphic multiplication on two 16-bit encrypted integers. This multiplication deteriorates the performance of the TFHE compared to the FHEW by halving the starting assumptions for a logic gate on ciphertexts bits. But in practice, this improvement is negligible if we consider that a binary multiplication of two 16-bit on plaintext numbers on the same architecture is carried out in less than 1 nanosecond. The ratio of improvement of the TFHE by adding the decryption time of the result is close to zero. This observation is also valid for the FHEW.

VII. CONCLUSION

Bootstrapping is the basis of unlimited homomorphic processing on encrypted data. This study compared bootstrapping through three patterns to identify its evolution from 2009 to 2016. It emerges from this comparison that the best design and implementation is that of the TFHE which is based respectively on the problem of the LWE on the real torus modulo 1, the bootstrapping by accumulator, on the fast Fourier transform coupled with the parallelism of FMA and AVX data. One avenue to explore is to study the performance of the implemented FHEW with a rapid transform based on the stockham algorithm, optimized throttle calculation and data parallelism.

REFERENCES

[1] Pascal Paillier, Public Key cryptosystem based on composite degree residuosity classes, In Stern 97, pages 223-238. 27, 29, 51, 53, 55.

[2] Taher El Gamal, A public key cryptosystem and a signature scheme based on discrete logarithms. In GR Blakey and David Chaum, editors, CRYPTO 1984, volume 196 of Lectures Notes in computer Sciences, pages 10-18, Springer 1984.

[3] R. L. Rivest, A. Shamir and L. Adleman, A method for obtaining digital signatures and public key cryptosystems. Common of the ACM, 21:120-126, 178.

[4] Shafi Goldwasser and Silvio Micali, probabilistic encryption, J. Computer. System. Sci, 28(2): 270-299, 1984. 6, 33.

[5] Dan Boneh, Eu-Jin Goh and Kobbi Nissim, Evaluating 2-DNF formulas on ciphertexts, In Joe Killian, editor, TCC 2005, Volume 3378 of Lectures Notes in Computer Science, Pages 325-341, Springer, 2002. 2, 31, 66, 67, 97, 98, 99.

[6] Craig Gentry. "A fully homomorphic encryption scheme". crypto.stanford. edu/craig. PhD thesis. Stanford University, 2009.

[7] Marten van Dijk, Craig Gentry, Shai Halevi, and Vinod Vaikuntanathan. "Fully Homomorphic Encryption over the Integers". In: EUROCRYPT 2010. Ed. By Henri Gilbert. Vol. 6110. LNCS. Springer, Heidelberg, May 2010, pp. 24–43.

[8] Craig Gentry, Amit Sahai, and Brent Waters. "Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based". In: CRYPTO 2013, Part I. Ed. by Ran Canetti and Juan A. Garay. Vol. 8042. LNCS. Springer, Heidelberg, Aug. 2013, pp. 75–92. doi: 10.1007/978-3-642-40041-4_5.

[9] Léo Ducas and Daniele Micciancio. "FHEW: Bootstrapping Homomorphic Encryption in Less Than a Second". In: EUROCRYPT 2015, Part I. Ed. by Elisabeth Oswald and Marc Fischlin. Vol. 9056. LNCS. Springer, Heidelberg, Apr. 2015, pp. 617–640. doi: 10.1007/978-3-662-46800-5_24.

[10] I. Chillotti, N. Gama, M. Georgieva, and M. Izabachène. TFHE: Fast Fully Homomorphic Encryption Library over the Torus. https: / / github . com /tfhe/tfhe. 2016.

[11] Ilaria Chillotti, Nicolas Gama, Mariya Georgieva, and Malika Izabachène. "Faster Fully Homomorphic Encryption: Bootstrapping in Less Than 0.1 Seconds". In: ASIACRYPT 2016, Part I. Ed. by Jung Hee Cheon and Tsuyoshi Takagi. Vol. 10031. LNCS. Springer, Heidelberg, Dec. 2016, pp. 3–33. doi: 10.1007/978-3-662-53887-6_1.

[12] R. L. Rivest, L. Adleman, and M. L. Dertouzos. "On Data Banks and Privacy Homomorphisms". In: Foundations of Secure Computation, Academia Press (1978), pp. 169–179.

[13] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. "(Leveled) fully homomorphic encryption without bootstrapping". In: ITCS 2012. Ed. by Shafi Goldwasser. ACM, Jan. 2012, pp. 309–325.

[14] Brakerski Z. and Vaikuntanathan V., Fully Homomorphic Encryption from RingLWE and security for key dependent messages, LNCS, vol. 6841, Springer Verlag, Proceedings of CRYPTO, pp. 505-524, 2011.

[15] Oded Regev. "On lattices, learning with errors, random linear codes, and cryptography". In: 37th ACM STOC. Ed. by Harold N. Gabow and Ronald Fagin. ACM Press, May 2005, pp. 84–93.

[16] Carlos Aguilar Melchor, Philippe Gaborit, and Javier Herranz. Additively homomorphic encryption with d-operand multiplications. In CRYPTO, pages 138–154, 2010.

[17] Nigel P. Smart and Frederik Vercauteren. Fully homomorphic encryption with relatively small key and ciphertext sizes. In Phong Q. Nguyen and David Pointcheval, editors, Public Key Cryptography, volume 6056 of Lecture Notes in Computer Science, pages 420–443. Springer, 2010.

[18] Craig Gentry and Shai Halevi. Implementing gentry's fully-homomorphic encryption scheme. In Kenneth G. Paterson, editor, EUROCRYPT, volume 6632 of Lecture Notes in Computer Science, pages 129–148. Springer, 2011.

[19] Craig Gentry and Shai Halevi. Fully homomorphic encryption without squashing using depth-3 arithmetic circuits. Cryptology ePrint Archive, Report 2011/279, 2011. http: //eprint.iacr.org/2011/279.

[20] N. Howgrave-Graham. Approximate integer common divisors. in J. Silverman (ed), Cryptography and Lattices, Springer LNCS 2146 (2001) 51−66.

[21] Jean-Sebastien Coron, Avradip Mandal, David Naccache, and Mehdi Tibouchi. Fully Homomorphic Encryption over the Integers with Shorter Public Keys. Cryptology ePrint Archive, Report 2011/441, 2011. http://eprint.iacr.org/.

[22] Jean-Sébastien Coron, Tancrède Lepoint, and Mehdi Tibouchi. Scale-invariant fully homomorphic encryption over the integers. 9 In Public-Key Cryptography–PKC 2014, pages 311–328. Springer, 2014.

[23] JungHee Cheon, Jean-Sébastien Coron, Jinsu Kim, MoonSung Lee, Tancrède Lepoint, Mehdi Tibouchi, and Aaram Yun. Batch Fully Homomorphic Encryption over the Integers. In Thomas Johansson and PhongQ. Nguyen, editors, Advances in Cryptology – EUROCRYPT 2013, volume 7881 of Lecture Notes in Computer Science, pages 315–335. Springer Berlin Heidelberg, 2013.

[24] Jean-Sébastien Coron, David Naccache, and Mehdi Tibouchi. Public Key Compression and Modulus Switching for Fully Homomorphic Encryption over the Integers. In David Pointcheval and Thomas Johansson, editors, EUROCRYPT, volume 7237 of Lecture Notes in Computer Science, pages 446–464. Springer, 2012.

[25] Jean-Sébastien Coron, Tancrède Lepoint, and Mehdi Tibouchi. Batch Fully Homomorphic Encryption over the Integers. Cryptology ePrint Archive, Report 2013/036, 2013. http://eprint. iacr.org/.

[26] Brakerski Z. and Vaikuntanathan V., Efficient Fully Homomorphic Encryption from standard LWE, Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, pp. 97-106, IEEE Computer Society, 2011.

[27] BLLN13. Joppe W. Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved Security for a Ring-Based Fully Homomorphic Encryption Scheme. In Martijn Stam, editor, IMA Int. Conf., volume 8308 of Lecture Notes in Computer Science, pages 45–64. Springer, 2013.

[28] V. Lyubashevsky, C. Peikert, and O. Regev. On ideal lattices and learning with errors overs rings. In In Proc. Of EUROCRYPT, Volume 6110 of LNCS, pages 1-23. Springer, 2010.

[29] Yuanmi Chen and Phong Q. Nguyen. Faster Algorithms for Approximate Common Divisors: Breaking Fully-Homomorphic Encryption Challenges over the Integers. In David Pointcheval and Thomas Johansson, editors, EUROCRYPT, volume 7237 of Lecture Notes in Computer Science, pages 502–519. Springer, 2012.

[30] Kassem Kalach, Implementation of the multiplication of large numbers by FFT in the contexts of cryptographic algorithms, August 2005, dissertation, Université de Montréal.

[31] P. Boale Bomolo, S. Ntumba Badibanga, E. Mbuyi Mukendi, Implementation of homomorphic arithmetic operations on integers, volume 5, number 2, May 2021, pp 125-137, IJISR.

[32] P. Boale Bomolo, S. Ntumba Badibanga, E. Mbuyi Mukendi, performance of Adder Architectures on encrypted integers, volume 10, issue-6, august 2021, IJEAT.

A Hybrid Similarity Measure for Dynamic Service Discovery and Composition based on Mobile Agents

Naoufal EL ALLALI, Mourad FARISS, Hakima ASAIDI, Mohamed BELLOUKI
Department of Computer Science, Mohammed First University
FPD Nador Laboratory MASI
Nador, Morocco

***Abstract*—With the ever-present competition among companies, the prevalence of web services (WSs) is increasing dramatically. This leads to the diversity of the similar services and their developed nature, which makes the discovery of a relevant service during the composition phase a complex task. Since most of the competition companies aim to discover high-quality services with minimum charges in order to increase the number of customers and their profit. The semantic WSs allow performing dynamic service discovery through the entities software and intelligent agents. However, the solutions provided to the discovery process are limited to their performance in terms of the quickness to respond to the request in real-time, without considering the constraints such as the accuracy in the discovery phase and the quality of the similarity mechanism evaluation. They usually are based on the similarity measure of distance between concepts in the ontology instead of taking into consideration the relationships semantically and the strength of the semantic relationship between concepts in the context. In this paper, we proposed a novel hybrid semantic similarity method to improve the service discovery process. The hybrid method is applied to an architecture based on mobile agents, where cooperative agents are integrated to facilitate and speed up the discovery process. In the first hybrid method, we defined the Latent Semantic Analysis (LSA) with a semantic relatedness measure to avoid the ambiguity of the terms and obtain a purely semantic relatedness at level of the service description. The second one is defined to analyze the relationships at the level of the I/O service based on the subsumption reasoning, called IO-MATCHING. Experimental results on a real data set demonstrate that our solution outperforms the state-of-the-art approaches in terms of precision, recall, F-measure, and consumed time of the service discovery.**

Keywords*—IO-MATCHING; latent semantic analysis; mobile agents; OWL-S; semantic web services; semantic similarity; semantic relatedness

I. INTRODUCTION

Over the last years has become widely popular as the number of Web services deployed in the world is rapidly increasing owing to their low-cost and cross-organizational construction of distributed applications in heterogeneous environments [1]. In another term, as the number of WSs increases, the discovery of web services needed by the user becomes more and more critical [2, 3]. However, the requested information and knowledge from the data remain difficult to obtain precisely. Since there are some conventional approaches based on WSDL [4] as the description of Web Service, it provides limited results due to lack of semantic service description. Contrary to other service descriptions such as OWL-S [5], WSMO [6] and SAWSDL [7], which are based on the semantic description of web services. Thus, The Semantic Web Services (SWS) concept is the result of integrating Web services and Semantic Web technologies [8].

The key point behind integrating Web Service and Web semantic is developing intelligent service-based applications and carrying out high-precision semantic discovery and automated service composition based on formal ontology-based service semantics representations [9, 10]. These service-based applications can reason based on such formal service semantics. This can support not only semantic interoperability between services, but also planning of their logic-based automated composition and more precision service discovery [11]. Thus, the process of service discovery and composition is generally based on service description, increasingly beyond syntactic descriptions to incorporate the semantics of the service to enable more accurate analysis.

With the advancement of semantic technology in web services has become more attractive to researchers in recent years due to the importance of existing web services on the Internet [12]. However, that does not mean there are no complex challenges confronting researchers to improve web service discovery in real-time. Since some solutions [13–15] aim to minimize the execution time of web service discovery but generally lead to low productivity with marginal performance, they do not target semantic analysis of the request to achieve an accurate solution, making it challenging to realize the semantic web discovery process. Most of these solutions are based on the distance between two concepts of the ontology to measure the degree of similarity rather than to consider the semantic relatedness existing between these two concepts in a contextual way.

The crucial issue in the discovery process is that consists on the way to measure the correspondence ratio between the request and the service concepts, and also the semantic correlation strength between both. So that the semantic similarity and the semantic relatedness are two different concepts, because the semantic relatedness includes the strength of relationship between two concepts in a context, while the concept of semantic similarity is more specific than the semantic relatedness [16]. The semantic similarity is done by evaluating two concepts in a taxonomy or ontology, which are constructed only by "is a" relations. For example, "book" is similar to "novel", but is also related to "author" and "publication". Thus, more the similarity between two concepts

is higher, more the relatedness is increased in the given context [17]. For this reason, there are some semantic discovery methods based on simple matching of the concepts annotated to services and requests, without considering the relationship of these concepts to the desired service context, rather than a simple semantic matching of terms that are related to I/O. This is considered to be insufficient to improve performance either in the semantic discovery process or during composition, producing results according to the similarity ratio between terms without taking into account the semantic relatedness of the terms to the desired service.

In this paper, we proposed a novel approach that addresses the service discovery problem based on a cooperative system by mobile agents developed in [18]. This approach aims to analyze the semantic services in a contextual way. The provided approach takes into account all the constraints discussed in the above paragraphs. In particular, the novelties of our proposal are:

- The use of the parallelism of the agent technology to make the service discovery process more efficient and dynamic.
- The cooperative agents enable the semantic analysis of the request autonomously to improve the accuracy rate.
- The integration of a semantic analysis agent in order to facilitate the retrieval of ontology relationships between concepts and to enhance the performance of the discovery process. This integration provided to the proposed module by [18], which allows reinforcing in a robust and more flexible in responding to any point of the execution, enables to achieve better performance and lower memory consumption to select the composition of the services dynamically.
- The support of a secondary database to improve the quickness and reliability of semantic analysis without reproducing the extraction of ontological relations between concepts.
- The semantic analysis agent targets to extract the semantic relatedness strength between the wanted keywords and the service description, in order to return the service in context for responding the desired request to be realized.
- The use of a hybrid similarity method proposed to maximize the matching process between the query and service.
- The first hybrid similarity measurement method is a classical tool to retrieve the description services similarities automatically; through dimensionality compression, it is known as Latent Semantic Analysis [19]. In addition, the semantic relatedness between the concepts and the desired service is performed to support the LSA.
- The second hybrid method is based on the relationship between the input/output (I/O) concepts in their OWL ontologies; it is known as IO-MATCHING [20].

The experimental results on a real dataset demonstrate that our solution outperforms the state-of-the-art approaches in terms of precision, recall, F-measure and consumed the time of the service discovery.

The remainder of this paper is organized as follows: Section 2 introduces the related works, Section 3 presents our proposal approach, Section 4 demonstrates the experiment results and discussion, and the final section concludes the paper and the future work.

II. Related Work

With the radical proliferation occurring in web services technology, it is becoming difficult to discover a service that is adequate to the user's requirements. For that reason, there are many solutions to reinforce the service discovery problem in terms of functionality and QoS. In this regard, we present only related works to achieve a better understanding of the advantages that can be obtained and put our contributions in context.

The technique suggested in [18] provided a method for discovering and composing SWSs in a distributed environment. This technique is based on a mobile agent, which has the characteristics of self-reliance, social capability, self-learning, and, most importantly, mobility. It is a technology suitable for autonomously exploiting SWSs to provide end-user applications. The mobile agent aims to discover the SWS desired from different locations and the generated graphs to perform the composition process. Despite a sufficient result provided by the discovery process, it may provide relevant services, but it does mean that there is no accurate measure to find a service that satisfies the user's requirement.

The authors [21] suggested an approach to automatically compose web services based on multi-agent systems and an algorithm to dynamically select an optimal solution as a service that responds to the customer's requirements. This composition is based on the quality and composition-capacity of the participating services. They aim to design, deploy and manage distributed systems more efficiently by combining, reorganizing, and adapting the services. Despite the efficiency feasibility provided by their proposed module, it does not cover some evaluation metrics and the performance of the similarity semantic method during composition.

The work of [22] proposed a new WSs discovery method based on semantic matching and service clustering for effective and practical web services discovery, which integrates functional similarity with process similarity. Their suggested approach is based on the knowledge available from the semantic description model, based on improving Lin's [23] semantic similarity measure to include opposition or degree of contrast as specified in [24]. Their vision is to develop a practical WS discovery approach based on pre-clustering that enable them to perform semantic and scalable WS discovery in a short period and thus minimize the search space. However, their method similarity proposed is not accurate to describe semantically due to the ignorance of the two concepts' antisense relationships.

A new semantic similarity method is proposed by the authors [25] that may be performed on both the textual

description and the interface of WSs. Their proposed semantic similarity method incorporates multi-conceptual relationships for service discovery. It is based on the relational semantic distance between concepts in WordNet and other ontologies. This method provides a more accurate estimation of the similarity between the terms, the web services and the query. Although the experimental results are promising in terms of precision, recall and f-measure, but it is limited to the semantic similarity of the terms based on the generic WordNet ontology.

In [26], a proposed method for discovering and selecting WSs that use OWL-S to represent web services, quality of service, and customer demand. This architecture is built on system-multi-agent approaches that make use of semantic web services. Their proposed technique discovers services similar to the consumer request based on functional and QoS parallels and reputation computing. Their model is based on four-layers: the web service and request description layer, the functional match layer, the QoS computing layer, and the reputation computing layer. Their Future work includes combining several Web services into an atomic service (service composite) and composes Web services based on customer preferences and QoS.

The authors in [27] suggested an automated approach to discovering semantic Web services. It is characterized by an ontology-based service preprocessor, a reasoning-based service filter, and a parameter-based matcher of the service. The first uses the ontology defined by services and requests to reduce the number of candidate services. The second one consists basically on a reasoning-based service filter to extract the concepts tagged to the input and output parameters of the selected services from the SAWSDL set of documents. Consequently, it logically deduces the concepts and filters out the services that are insufficient to satisfy the user's parameter needs. Finally, the third one is a parameter-based service matcher based on the measure of semantic similarity in the matching algorithm (PBSM_R). This semantic similarity measure is mainly based on the relationship between the concepts of the domain ontology. Lastly, it returns services adequate to the requirements of the user. Although the results achieved in performance of the runtime through the narrowing the search space, but it lacks on one side precision and recall.

The authors [28] proposed a novel service discovery scheme based on a combination of similarity methods using the WSDL specification and ontology to make the service discovery process more automated, discover the best match rapidly, and improve the Hungarian algorithm [29] is used. This method combination includes the structural similarity, the semantic similarity and the concept similarity based on bipartite matchmaking techniques used to discover web services. This suggested scheme includes two phases to discover the most suitable services to the request. In the first phase, measuring similarity between the requested service and a set of advertised services. In the second phase, a bipartite graph of nodes defined based on the ontology is used to describe semantic Web service matching. The obtained experimental results are better than other existing schemes using the Hungarian algorithm in terms of precision, recall and f-measure, but it is lacked parallelizing some steps in the discovery process.

III. The Proposed Approach

This section presents an approach for supporting the discovery process during web service composition using the cooperative agents, which targets the minimization of the discovery performance overhead without requiring the memory pre-loading of service registries. Moreover, the maximization of the matching algorithm by the hybrid method proposed. This method consists in retrieving the context of the terms in the service description, where the service context can provide more accurate information regarding the services that are relevant to the request. The main novelty of our proposed discovery system aims to be self-adapting to unexpected variations, particularly in a heterogeneous environment.

Service discovery is a crucial issue to accomplish at each major step of the composition generation process. However, the increasing number of services on the Internet, the dynamic and unstable nature of these entities makes the composition tasks more difficult. Therefore, it considers the process of discovering a service during composition most important to ensure the system has the ability to semantically parse, respond to a request quickly and accurately in real-time. In order to have an efficient system for identifying the best solutions, we adopted the architecture developed by the authors [18], but with the incorporation of our hybrid semantic similarity measurement method and an agent to analyze ontological relationships as illustrated in Fig. 1, it can lead to successful results as detailed in the next section. This architecture is based on system multi-agent (SMA) for the discovery of Web services. Our contribution to this architecture is to improve the semantic similarity method's efficiency in the service discovery process.

Fig. 1 represents an improvement of the author's architecture [18], based on a primary agent in the distributed environment called a Mobile agent. This agent is used to discover the desired SWS in different locations and graphs to accomplish the composition process. This architecture includes the main entities to support the discovery process through the exploitation of the ontology domain used and facilitate the search process, which is corresponding between the request and the service offers. In the following, we will explain more details what happens in each entity.

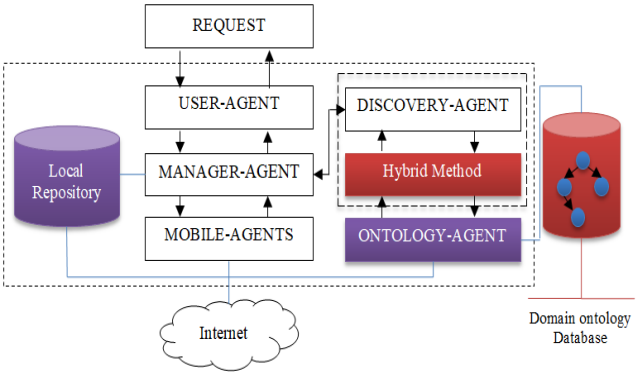


Fig. 1. Integration of the Proposed Solution for Mobile Agent Architecture to Discover and Compose SWS.

- USER-AGENT: is a point to interact between the request and the service discovery system. This agent is responsible for providing the user with an OWL-S standard semantic module [5] to express the request. The user's request is composed of inputs/outputs, a reference to the domain ontology to be used, and after the processing returns the desired results to the user.
- MANAGER-AGENT: Checks the availability of the desired service index in the local repository or whether the service can be composed of the local repository services. In the absence of the desired service, it is up to a set of mobile agents to search in different locations on the web to find the desired service.
- MOBILE-AGENTS: are responsible for retrieving the semantic web services from different websites instead of using a crawler due to their speed performance and low network overhead. It satisfies the needs of "MANAGER-MOBILE" to reinforce these services during the composition.
- ONTOLOGY-AGENT: is designed to facilitate the semantic analysis of the I/O of the service required by the discoverer agent (as illustrated in Fig. 2). It is considered as a cooperative agent. It analyzes the ontology domain that corresponds to the request of the discovery agent. Thus, it extracts the classes and their links to deduce the generalization relations between the concepts, which means a concept is more general than another in the arborescence (as shown in Fig. 3). These domain ontologies are stored in the domain ontology database.

This agent can exploit the relations already deducted in advance that are stored in the secondary database. This database is developed as a memory cache to avoid spending more interaction and extract these relations quickly without reproducing the operation of processing analysis.

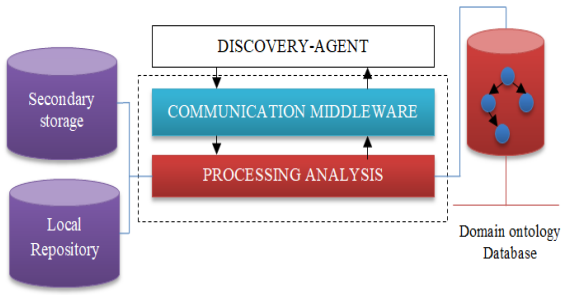


Fig. 2. Components of "ONTOLOGY-AGENT".

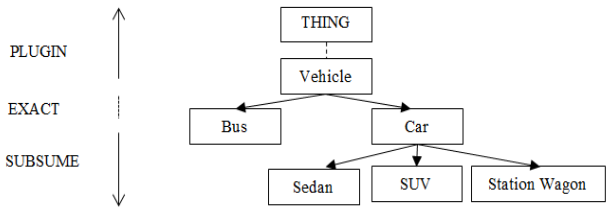


Fig. 3. A Vehicle Ontology Fragment [20].

- DISCOVERY-AGENT: allows discovering semantic web services that fulfills the requirements of the manager agent. The discovery process is based on the mobilization of the hybrid measurement similarity method and the "ontology-agent". The hybrid method integrated into the discovery agent are the following: The first method, based on the LSA, aims to investigate the relationships between a set of service descriptions and the terms embedded, by producing a set of concepts related to the service descriptions and the terms. In addition, to analyze the semantic relatedness between the concepts and the desired service. The second method, based on IO-MATCHING, intends to describe the degree of matching between two I/O concepts using the ontology agent.

To maximize the contextual similarity of service discovery, we propose a hybrid method which focuses on the semantic similarity between the input/output concepts of services, and to find the semantic relatedness between the service description terms in a contextual way. This is intended to facilitate the performance of "ontology-agents" to cooperate in a more intelligent and explicit sense with other agents. In the following, we will describe in more detail the different definitions to clarify the mechanism of similarity provided.

Definition 1 (Request): the request of the user is defined as $R = \langle R_{in} | R_{out} | R_{des} \rangle$, where R_{in} denotes the set of required input parameters, R_{out} denotes the set of required output parameters , and R_{des} denotes the required service description.

Definition 2 (Web Service): A web service is described by the OWL-S ontology. The service defined as a 3-tuple: $S = \langle S_{in} | S_{out} | S_{des} \rangle$, where S_{in} and S_{out} are the input and output concepts respectively, S_{des} is the description of the service.

Definition 3 (LSA): Latent Semantic Analysis is used to discover the hidden and subjacent (latent) semantics of words in a corpus of documents by constructing "concepts" related to documents and terms. It is a standard technique to extract automatically similarities between documents, by reducing the dimensionality. A word-document matrix is packed with weights according to the extent of the word in the specific document and is then reduced by singular value decomposition to a reduced dimensional space called conceptual space. The LSA process includes four steps illustrated in Fig. 4 as follows.

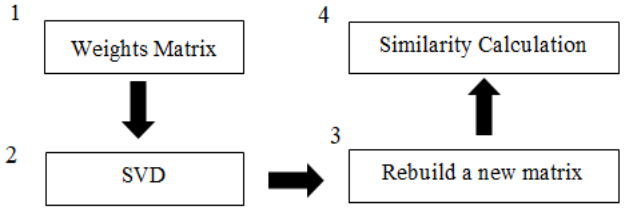


Fig. 4. The LSA Algorithm Processing Steps.

Step 1: Before building a weights matrix, the short text of the service description must be treated in the pre-processing phase as normalization of the service description to reduce the information ambiguity. The next phase is the tokenization task,

which consists of transforming the short text existing in S_{des} to a set of separate terms as a set of tokens. After finishing the tokenization task, it will be the stemming task to convert different forms of terms into a similar canonical form. Before finishing the pre-processing task, the terms must be sorted alphabetically. This step can be completed by building a weigh matrix as illustrated in equation (1) bellow.

$$A=\begin{pmatrix} & t_1 & \cdots & t_j \\ s_1 & w_{1,1} & \cdots & w_{1,j} \\ \vdots & \vdots & \ddots & \vdots \\ s_i & w_{i,1} & \cdots & w_{i,j} \end{pmatrix} \quad (1)$$

where $W_{i,j}$ represents the weight of the term i in the service j. The Term Frequency-Inverse Document Frequency (TF-IDF) can be calculated as:

$$W_{ij} = TF_{i,j}(T,S) \times IDF(T) = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log\left(1+\frac{N}{df_i}\right) \quad (2)$$

where $n_{i,j}$ is the number of occurrences of the term t in the service, N is the total number of services in the corpus, and df_i is the number of services where the term T_i occurs.

Step 2: After the generation of weight matrix, it follows the step of decomposition of matrix A by SVD as illustrated in equation (2),

$$A=U\sum V^T \quad (3)$$

where,

U Term matrix.

$\sum$ Descriptions service matrix.

V^T Singular value matrix.

Step 3: Once the matrix A is decomposed by SVD, we have to reduce the vector space to an approximation with a rank of $k=4$, it becomes to find a service description closest to the request. A request is represented in the k-dimensional vector space as a service. A request (R) can be represented as follows:

$$R_{des} = R_{des}{}^T U_k \sum\nolimits_K^{-1} \quad (4)$$

Step 4: Then, we need to measure the cosine similarity to evaluate the similarity between the query description and the service description. The cosine similarity measure is defined as follows.

$$sim(R_{des}, S_{des}) = \frac{R_{des} \cdot S_{des}}{\| R_{des} \| \| S_{des} \|} \in [0,1] \quad (5)$$

Definition 4 (Description Similarity): A service description S_{des} is a short text which describes the typical properties of a service. The service descriptions include rich information to be evaluated semantically. To calculate the description similarity (DS) will be evaluated by the similarity of the hidden topics using the LSA method as mentioned in definition 3, additionally evaluating the correlation rate to the I/O concepts with the service description. These concepts are considered essential keywords to improve the precision rate. The description semantic similarity is defined as follows.

$$sim_{DS}(R_{des}, S_{des}) = \delta \times sim(R_{des}, S_{des}) + (1-\delta) \times Relatedness(R_{des}, S_{des}) \quad (6)$$

where $\delta \in [0,1]$ is weight factor of the LSA similarity and $Relatedness(R_{des}, S_{des})$ is the semantic relatedness.

To infer the semantic relatedness between the wanted keywords and the service description, our ontology method mentioned in definition five correlates the I/O concepts to get the diversification of the service description related to these concepts in a semantically precise way. We define the semantic relatedness method as follows.

$$Relatedness(R_{des}, S_{des}) = \begin{cases} e^{-((1-\frac{\ln(1+n_k)}{1+\ln(n)})\times(\frac{T(T_R,T_S)}{\max(W_{T_R},W_{T_S})}))} & \text{if } T(T_R,T_S) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $Relatedness(R_{des}, S_{des})$ is normalized in the range [0,1], n_k is the number of occurrences that a combination of terms appears in the service descriptions, and n is the number of services in the corpus. $T_R = R_{in} \cup R_{out}$ is a set of I/O concepts of the request, and $T_s = \{t_1, t_2, ..., t_n\}: t_1, t_2, ..., t_n \in S_{des}$ is a set of terms of the web service description that are semantically related to concepts of the T_R request. Also, w_{T_R} and w_{T_S} are the term weights of the wanted keywords T_R and T_s respectively.

To analyze the semantic compatibility between the wanted keywords T_R and the concepts of T_s, the semantic matchmaking method is used, which is in charge to assess the degree of compatibility between the concepts included in the keywords with the concepts of the S_{des}. This method uses semantic reasoning (subsumption reasoning) to analyze the relationship between concepts. These ontology relationships allow extracting the concepts compatible to the T_s concepts, and this allows improving the performance of the LSA method semantically. So, to retrieve the web service which is related to the wanted keywords T_R, we determine the subsumption relationship as follows.

$$T(T_R, T_S) = \begin{cases} 1 & if\ T_R = T_s \\ \frac{1}{2} \times \left(\frac{\sum_{c \in T_R} max_{t \in T_S}\{Match(t,c)\}}{|q_{T_R}|} + \frac{\sum_{t \in T_S} max_{c \in T_R}\{Match(c,t)\}}{|q_{T_s}|}\right) & otherwise \end{cases} \in [0,1] \quad (8)$$

where c and t are any concepts of wanted keywords T_R and the service description T_s respectively. Moreover, $|q_{T_R}|$ and $|q_{T_S}|$ are the total number of the ontology relationships that have the maximum value and greater than zero.

As illustrated in the formula (8), if the all concepts of T_R appear in T_s, at this point the value of $T(T_R, T_S)$ is 1. In otherwise, the concepts of T_R are adjusted by other related

concepts using the ontology-agent, these concepts are different than the initial concepts T_R, in order to identify of the different keywords that are closer to the desired service description. It leads to avoid the ambiguity of the terms frequency and to enrich the terms which have a purely semantic relatedness with the service description. Thus, more the value of $T(T_R,T_S)$ is higher, more the concepts of keywords T_R are suitable to the concepts of T_s, that's means the value of semantic relatedness will be maximized.

Definition 5 (Interface Similarity): Interface similarity (IS) is determined by the semantic compatibility between S_{in} and S_{out}. This compatibility is evaluated by the degree of semantic matching between concepts, which is called IO-MATCHING. The relations of these concepts are deduced by the "ontology-agent" analyzer. This degree of semantic matching uses 4 types of matching score: Exact, Plugin, Subsume, and Fail to measure the matching between two S_{in} / S_{out} concepts as follows:

$$Match(C_i,C_j)=\begin{cases}\text{EXACT } if\ C_i \equiv C_j \\ PLUGIN\ if\ C_i \sqsubseteq C_j \\ SUBSUME\ if\ C_i \sqsupseteq C_j \\ FAIL\ otherwise\end{cases} \in [0,1] \quad (9)$$

where C_i and C_j are the request and service concepts respectively. The interface similarity between the request *R* and the service *S* is calculated as shown in the equation below.

$$sim_{IS}(R,S)=\frac{\sum Match(C_i,C_j)}{max\{Card(S_{in}\cup S_{out}),Card(R_{in}\cup R_{out})\}} \in [0,1] \quad (10)$$

In the literature [20], the different degrees of matching that are often considered are as follows:

- EXACT ($C_i \equiv C_j$): if the concepts C_i and C_j belong to the same ontology class.
- PLUGIN ($C_i \sqsubseteq C_j$) : where the C_j concept in the ontology is a sub-class of C_i, the concept C_i is more specific than the desired concept C_j
- SUBSUME ($C_i \sqsupseteq C_j$) : if the class of C_i is more general than the class of C_j, it indicates that the class of C_j is a sub-class of C_i.
- FAIL ($C_i \perp C_j$): when there is no subsumption relationship in ontology between C_i and C_j.

Definition 6 (Functionality semantic similarity): The functionality semantic similarity measure (FSM) includes two main components: description similarity and interface similarity. Functionality semantic similarity is defined as follows.

$$sim_{FS}(R,S)=\alpha\times sim_{IS}(R,S)+\beta\times sim_{DS}(R,S) \in [0,1] \quad (11)$$

where α and β are the interface similarity weight and the description similarity weight, respectively.

Table I represents the best-desired services to fulfill the request. As the desired service which should return a book price. It demonstrated the semantic relatedness/ similarity performance, which reinforces the LSA and IO-MATCHING similarity method to identify the hidden relationships in the service description rather than to focus the semantic analysis of the I/O. Although the similarity at the input/output level is similar, it provides different functionalities than expected. For example, the similarity at the interface level in the service "Cheapest Book Service" provides different request requirements. On the contrary, "BookPrice" and "BookPriceService" services respond to the users' same needs. As a result, it is crucial to measure similarity at the level of service description to extract hidden semantic relations and increase accuracy. This experiment is done by the weight of the interface similarity $\alpha=0.5$ and the description similarity $\beta=0.5$.

TABLE I. AN ILLUSTRATION OF THE RESULTS OBTAINED FROM THE SIMILARITIES BETWEEN THE SERVICES AND THE REQUEST

Service name	Inputs	Outputs	Text description	FSM
Cheapest Book Service	#_BOOK	#_PRICE	A Service that searchest the cheapest Price for a book	0.94
BookPriceService	#_BOOK	#_PRICE	Return price of a book	0.98
Bamzon RecommendedPriceService	#_BOOK	#_RECOMMENDERPRICEINDOLLAR	Bamzon is a popular service to return recommended price of a book	0.87
BookPrice	#_BOOK	#_PRICE	Uses the ISBN to return price of a book	0.96
BDe RecommendedPriceService	#_MONOGRAPH	#_RECOMMENDERPRICEINEURO	BDe is a competitor web service to return recommended price of a monograph in Euro	0.70
BookPriceTaxedPriceService	#_BOOK	#_TAXEDPRICE,#_PRICE	This service informs the taxed price of a book	0.88

IV. Experiment Results

In this section, we present our analysis that includes two main parts: In the first part, an overview of the experimental settings. The second part discusses the experimental results obtained by comparing the performance provided to another work [27].

A. Experimental Setup

To improve the performance mentioned in the last Section, we have been implemented our proposed approach in JADE Platform and OpenNLP Framework [30], which are based on the java language using an Intel® Core (TM) i7-4770 processor with 8 GB of main memory running Windows 10. Our experimental data is from the OWLS-TC version 3.0 dataset, which contains 1007 indexed OWL-S services, most of which were collected from public IBM UDDI registries semi-automatically transformed from WSDL to OWL-S. Table II below summarizes the features of the experimental environment.

To analyze the correctness and performance as discussed in our contribution, we carried out two experiments in different weights to prioritize each aspect of similarity (interface similarity and description similarity) as shown in the Table III. These parameter weights are scaled according to the importance of the similarity parameter in two different scenarios, these two scenarios will be experimented in order to understand the value added in our solution will be illuminated in the next Sub-section. Furthermore, the value of the weight factor δ=0.3 , which indicates a high importance of relatedness semantic than the LSA similarity, to reinforce the relatedness to cover the limitations of the LSA similarity.

TABLE II. The Experimental Environment

Environment	Description
Operating System	Windows 10
CPU	Core (TM) i7-4770
RAM	8 GB
Software Framework	JADE
NLP Toolkit	OpenNLP
Programming Language	JAVA
Dataset	OWL-S TC Version 3.0

TABLE III. The Weights of different Methods (Hybrid and IO-MATCHING Method)

Experiments	Method	Weight name	Parameter	Value
1	IO-MATCHING	Interface Similarity	α	0.50
	Hybrid Method LSA-IO	Description Similarity	β	0.50
2	IO-MATCHING	Interface Similarity	α	1
	Hybrid Method LSA-IO	Description Similarity	β	0

B. Results and Discussion

In order to carry out a standard and comparable analysis, we selected a set of 29 test queries (OWLS-TC3) associated with pertinence sets to lead performance evaluation experiments. These experiments are analyzed in more detail by comparing the precision, recall, and F-measure of the services retrieved by the two experiments, as illustrated in Table III. Then, the processing time is evaluated in function of the rising number of services, which are varied in each test (from 50 to 1007 services). It allows us to measure scalability according to the average speed to fulfill the query's requirements. For more analysis, we compared our solution with another method [27] to evaluate the system's performance in terms of scalability to clarify our system's success in dealing with all these constraints, as mentioned previously.

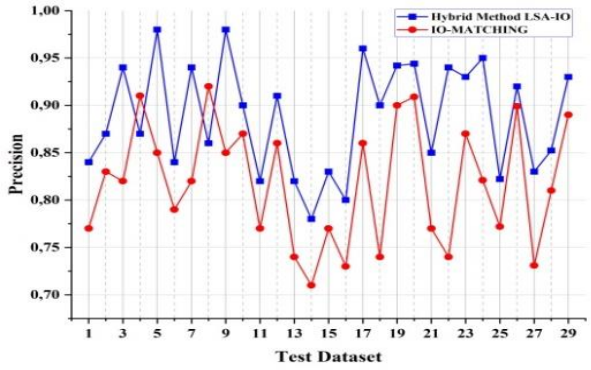
1) Evaluation metrics: As mentioned earlier, the experimental results should be analyzed in advance regarding the precision, recall and f-measure of the services retrieved by the hybrid and IO-MATCHING method. Precision is the ability to retrieve the most precise services. Higher precision means better relevance and more precise results but may imply fewer results returned. Recall means the ability to retrieve as many services as possible that match or are related to a query. F-Measure evaluates a weighted harmonic mean of precision and recall. As we used it for the evaluation process, it is then defined as follows.

$$precision = \frac{A_{Relevant} \cap B_{Retrieved}}{B_{Retrieved}} \quad (12)$$

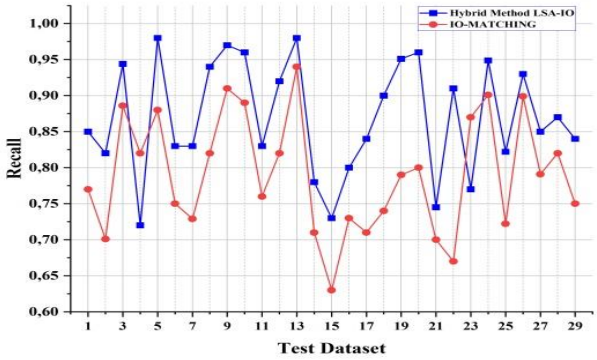
$$recall = \frac{A_{Relevant} \cap B_{Retrieved}}{A_{Relevant}} \quad (13)$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (14)$$

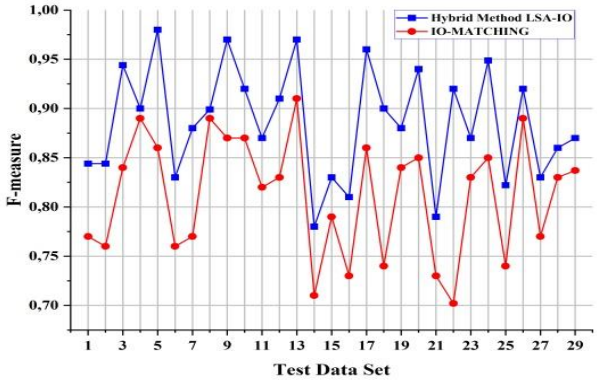
Where $A_{Relevant}$ is the set of relevant services, and $B_{Retrieved}$ is the number of relevant services retrieved. As indicated in the experimental results below, we run 29 test queries (OWLS TC-3) simultaneously to measure precision, recall, and F-measure in each experiment in Table III. The Fig. 5 demonstrates the efficiency of the interface similarity over description similarity as well as the performance provided by just the similarity measure at the I/O interface level. This proves that the hybrid method has a high value of precision, recall, and F-measure in all the query tests as compared to the traditional method (IO-MATCHING), which is purely based on the ontological relationships at the I/O level. With the exception of the query test 4 and 8, which record higher precision rate in concerning the IO MATCHING method. Relative to the query test 4 and 23, it also shows higher value of recall than the hybrid method.



(a) Precision.



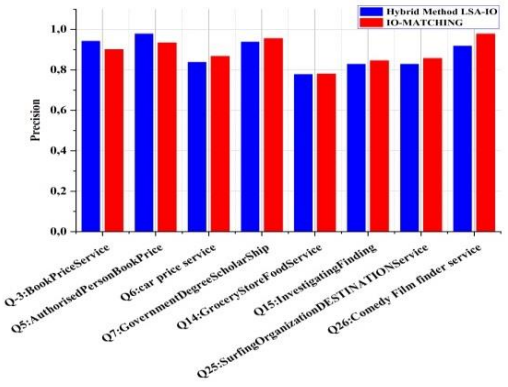
(b) Recall.



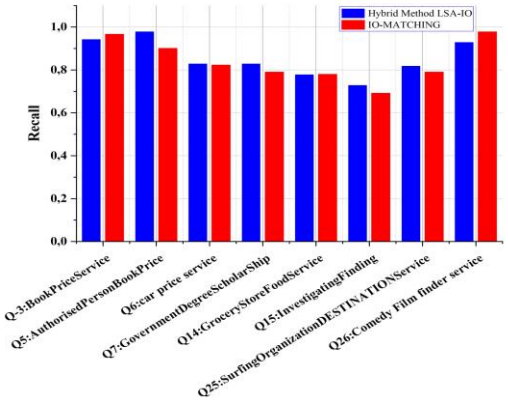
(c) F-measure.

Fig. 5. The Performance Measures for each Query in the Testing Dataset OWLS TC-3 for Both IO-MATCHING and Hybrid Method.

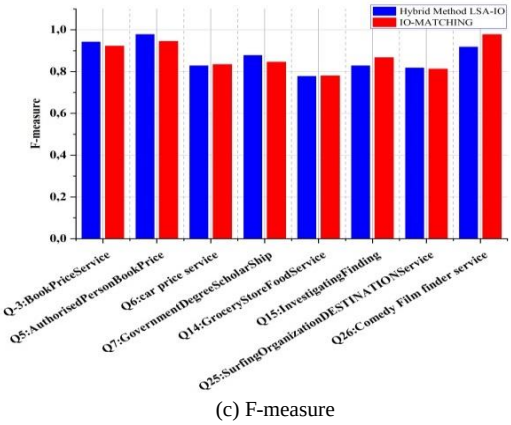
For more explicit the results obtained above, the Fig. 6 represents more in details by the selected test queries (OWLS TC-3), to clarify well the powerful hybrid method proposed in terms of the different evaluation metrics. It is the same criteria that we mentioned previously concerning the query tests, which are the identical to those shown in the Axis-x Fig. 6. These selected query tests indicate the challenges chosen to understand our hybrid method dominated by the IO-MATCHING method regarding precision, recall and F-measure. It was analyzed in Table III, the " Q3:BookPriceService" query can retrieve services similar but not as meaningful to the query in the case of experience 2. We deduced that it is not sufficient to just rely on the I/O interface instead of relating the concepts contained in the inputs/outputs with the description. But with Experimental 1, the hybrid method leads to relevant services to the query, due to the reinforcement provided to the LSA method by the semantic relatedness, which allows to improving the correlation ratio between the terms and the desired service instead of related to the frequency of terms. For example, the query "Q7:Government Degree ScholarShip " which requires the information on scholarships offered by a given government according to degree and government. Thus, the concept "#degree" can be related semantically to "#Academic-Degree" or "#Award", while the concept "#Government" can be related to the concept "#GovernmentOrganization". This is why the hybrid method is robust in terms of correlating these concepts with the service requested in the context.



(a) Precision



(b)Recall



(c) F-measure

Fig. 6. Evaluation of different Queries of OWL-S Dataset in Terms of Precision, Recall and F-measure.

Fig. 7 indicates the number of services retrieved via the hybrid or IO-MATCHING method for each query test (OWL-TC 3). These query tests vary in the number of inputs/outputs and related functionality desired to be achieved, in order to make a real challenge to the discovery systems concerned. As illustrated below, the significant number of services retrieved to satisfy the request set by our proposed system. Moreover, due to the performance provided by cooperative agents such as mobile agent and ontology-agent, that is proved in terms of different evaluation measures, to find a relevant service.

In addition, the comparison illustrated in Fig. 8 indicates the average precision-recall curve between the hybrid method and the IO-MATCHING method in order to expand the evaluation measures of the performance system proposed in retrieving a relevant service. In this experiment, we run 29 test queries in order to compute the average precision and recall. This demonstrates in the average precision-recall curve that the hybrid method excels considerably in the retrieval accuracy for relevant services based on the semantic correlations between the query concepts and the desired service. Consequently, the hybrid method has a higher precision value in service retrieval than the IO-MATCHING method.

2) Runtime performance comparison: To validate and evaluate the speed up performance for the proposed architecture, we compared our system with other works in the same scope [27] . To perform this comparison, we computed the average runtime according to each service set for different test queries. These sets vary in number of services to provide the scalability with a growing set of services (from 50 to 1007 services) and the response of the system in real time. This is demonstrated in Fig. 9, the results obtained in comparison with another work. The work [27] suggested to be compared, based on Ontology filtering and parameter matching relied on the discovery of function-oriented Web services to reduce the space of matching preprocessor and filter. While, we focused on the semantic analysis and the inter-relatedness between the concepts and the desired service, combined with the privileges provided by the agent ontology, which allows to exploit the pre-existing solutions in the second database, as discussed in Section 3, that facilitate the discovery of semantic relations and to reduce the consumption of reproducing the analysis of the operations required. This makes the discovery more flexible and rapid.

According to the results shown in Fig. 9, both approaches do not spend more time to the first sets of services (between 50 ms and 110 ms), while the OFPM method spends more run-time than the hybrid method when the number of services is scaled up. This is due to the cooperation of agents in the environment in order to make the discovery task and to respond in a real-time. This provides for a more flexible process of composition to accomplish its tasks.

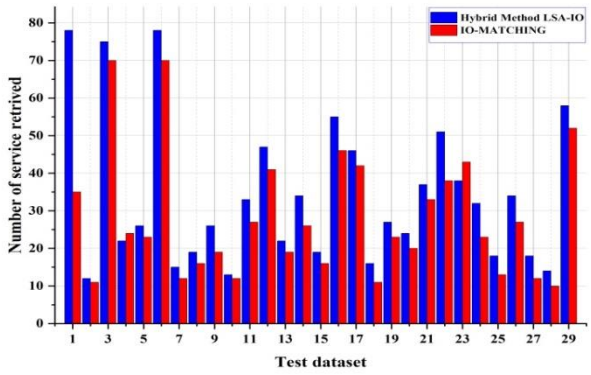


Fig. 7. The Number of Services Retrieved by the Hybrid Similarity Method and IO-MATCHING.

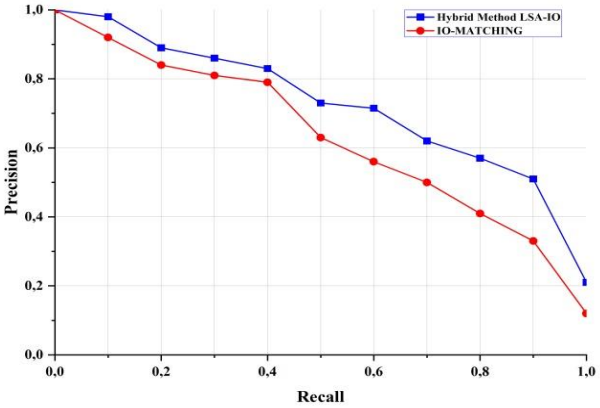


Fig. 8. Average 11-Points Precision-Recall Curve Across 29 Test Queries for the Hybrid Method and IO-MATCHING.

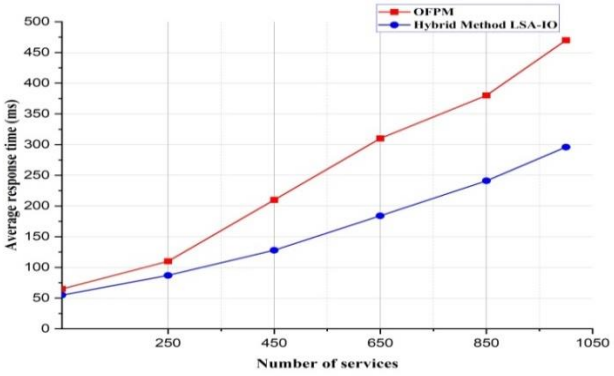


Fig. 9. Runtime Performance Comparison for Both Hybrid Method and OFPM [27].

V. Conclusion

This paper demonstrates the performance dynamicity provided to the architecture proposed. This allows handling the service composition more flexibly and quickly, with autonomous to find services accurately. Thus, it is proved in terms of the scalability and flexibility to respond in a real-time. This is due to the integration of the proposed hybrid method and the ontology analysis agent, which makes the architecture to be more dynamic in terms of autonomy, reliability and robustness. In addition, the proposed hybrid method makes the

system to meet the requirements of the query in the context of the semantic relatedness between the requested concepts and services. This leads to the high retrieval precision, recall and F-measure discovery process.

As future work, we will focus on integrating Micro-services and multi-agent systems (MAS) to reduce the time and complexity of composite semantic web services. Furthermore, we intend to enrich our system with the semantic descriptions of other functional aspects such as pre-conditions/post-conditions.

REFERENCES

[1] N. Niknejad, W. Ismail, I. Ghani, B. Nazari, M. Bahari, and A. R. B. C. Hussin, "Understanding Service-Oriented Architecture (SOA): A systematic literature review and directions for further investigation," Inf. Syst., vol. 91, p. 101491, Jul. 2020, doi: 10.1016/j.is.2020.101491.

[2] T. Aditya Sai Srinivas, S. Ramasubbareddy, and K. Govinda, "Discovery of Web Services Using Mobile Agents in Cloud Environment," in Innovations in Computer Science and Engineering, Springer, 2019, pp. 465–471. doi: 10.1007/978-981-13-7082-3_53.

[3] I. Ghani, W. M.N., and A. Mustafa, "Web Service Testing Techniques: A Systematic Literature Review," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, 2019, doi: 10.14569/IJACSA.2019.0100858.

[4] D. Booth and C. K. Liu, "Web services description language (WSDL) version 2.0 part 0: Primer," W3C Recomm., vol. 26, pp. 39–41, 2007.

[5] D. Martin et al., "OWL-S: Semantic markup for web services," W3C Memb. Submiss., vol. 22, no. 4, 2004.

[6] J. De Bruijn et al., "The Web Service Modeling Ontology," in Modeling Semantic Web Services, vol. 5, no. 1, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 23–28. doi: 10.1007/978-3-540-68172-4_3.

[7] J. Kopecký, T. Vitvar, C. Bournez, and J. Farrell, "Sawsdl: Semantic annotations for wsdl and xml schema," IEEE Internet Comput., vol. 11, no. 6, pp. 60–67, 2007.

[8] R. Studer, S. Grimm, and A. Abecker, Semantic web services. Springer, 2007.

[9] M. Klusch, P. Kapahnke, S. Schulte, F. Lecue, and A. Bernstein, "Semantic Web Service Search: A Brief Survey," KI - Künstliche Intelligenz, vol. 30, no. 2, pp. 139–147, Jun. 2016, doi: 10.1007/s13218-015-0415-7.

[10] F. Fakhar, "Semantic Constraints Satisfaction Based Improved Quality of Ontology Alignment," Bull. Electr. Eng. Informatics, vol. 2, no. 3, pp. 182–189, 2013, doi: https://doi.org/10.11591/eei.v2i3.202.

[11] M. Fariss, N. El Allali, H. Asaidi, and M. Bellouki, "Review of Ontology Based Approaches for Web Service Discovery," vol. 66, F. Khoukhi, M. Bahaj, and M. Ezziyyani, Eds. Cham: Springer International Publishing, 2019, pp. 78–87. doi: 10.1007/978-3-030-11914-0_8.

[12] R. Hammami, H. Bellaaj, and A. H. Kacem, "Semantic Web Services Discovery: A Survey and Research Challenges," Int. J. Semant. Web Inf. Syst., vol. 14, no. 4, pp. 57–72, Oct. 2018, doi: 10.4018/IJSWIS.2018100103.

[13] Z. Cong, A. Fernandez, H. Billhardt, and M. Lujak, "Service discovery acceleration with hierarchical clustering," Inf. Syst. Front., vol. 17, no. 4, pp. 799–808, Aug. 2015, doi: 10.1007/s10796-014-9525-2.

[14] A. Bukhari and X. Liu, "A Web service search engine for large-scale Web service discovery based on the probabilistic topic modeling and clustering," Serv. Oriented Comput. Appl., vol. 12, no. 2, pp. 169–182, Jun. 2018, doi: 10.1007/s11761-018-0232-6.

[15] K. Belmabrouk, F. Bendella, and M. Bouzid, "Multi-Agent Based Model for Web Service Composition," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 3, 2016, doi: 10.14569/IJACSA.2016.070320.

[16] A. Gupta, A. Kumar, and J. Gautam, "A survey on semantic similarity measures," Int. J. Innov. Res. Sci. Technol., vol. 3, no. 12, pp. 243–247, 2017, [Online]. Available: http://www.ijirst.org/articles/IJIRSTV3I12083.pdf

[17] Xiaojie Zheng, Jie Hu, and Lin Zhang, "Measuring semantic relatedness based on ontology," in International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), 2012, pp. 1335–1338. doi: 10.1049/cp.2012.1226.

[18] Latrache, "A MOBILE AGENT BASED APPROACH FOR AUTOMATING 'DISCOVER-COMPOSE' PROCESS OF SEMANTIC WEB SERVICES," J. Comput. Sci., vol. 10, no. 9, pp. 1628–1641, Sep. 2014, doi: 10.3844/jcssp.2014.1628.1641.

[19] S. T. Dumais, "Latent semantic analysis," Annu. Rev. Inf. Sci. Technol., vol. 38, no. 1, pp. 188–230, 2004.

[20] M. Paolucci, T. Kawamura, T. R. Payne, and K. Sycara, "Semantic Matching of Web Services Capabilities," in International Semantic Web Conference, Springer, 2002, pp. 333–347. doi: 10.1007/3-540-48005-6_26.

[21] G. Vadivelou, E. Ilavarasan, and S. Prasanna, "Algorithm for web service composition using multi-agents," Int. J. Comput. Appl., vol. 13, no. 8, 2011.

[22] F. Chen, M. Li, H. Wu, and L. Xie, "Web service discovery among large service pools utilising semantic similarity and clustering," Enterp. Inf. Syst., vol. 11, no. 3, pp. 452–469, Mar. 2017, doi: 10.1080/17517575.2015.1081987.

[23] D. Lin, "An information-theoretic definition of similarity.," in Icml, 1998, vol. 98, no. 1998, pp. 296–304.

[24] S. M. Mohammed, B. J. Dorr, G. Hirst, and P. D. Turney, "Measuring degrees of semantic opposition," Tech. Rep., 2011, doi: http://dx.doi.org/10.4224/19040608.

[25] F. Chen, C. Lu, H. Wu, and M. Li, "A semantic similarity measure integrating multiple conceptual relationships for web service discovery," Expert Syst. Appl., vol. 67, pp. 19–31, Jan. 2017, doi: 10.1016/j.eswa.2016.09.028.

[26] R. Benaboud, R. Maamri, and Z. Sahnoun, "Agents and OWL-S Based Semantic Web Service Discovery With User Preference Support," Int. J. Web Semant. Technol., vol. 4, no. 2, pp. 57–75, Apr. 2013, doi: 10.5121/ijwest.2013.4206.

[27] M. Fang, D. Wang, Z. Mi, and M. S. Obaidat, "Web service discovery utilizing logical reasoning and semantic similarity," Int. J. Commun. Syst., vol. 31, no. 10, p. e3561, Jul. 2018, doi: 10.1002/dac.3561.

[28] S. A. Khanam and H. Y. Youn, "A Web Service Discovery Scheme Based on Structural and Semantic Similarity.," J. Inf. Sci. Eng., vol. 32, no. 1, pp. 153–176, 2016.

[29] R. Jonker and T. Volgenant, "Improving the Hungarian assignment algorithm," Oper. Res. Lett., vol. 5, no. 4, pp. 171–175, Oct. 1986, doi: 10.1016/0167-6377(86)90073-8.

[30] A. OpenNLP, "a Machine Learning Based Toolkit for the Processing of Natural Language Text," URL http//opennlp. apache. org (Last accessed 2013-06-18).

Linear Mixed Effect Modelling for Analyzing Prosodic Parameters for Marathi Language Emotions

Trupti Harhare, Milind Shah
Dept. of Electronics and Telecommunications
Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India

***Abstract*—Along with linguistic messages, prosody is an essential paralinguistic component of emotional speech. Prosodic parameters such as intensity, fundamental frequency (F0), and duration were studied worldwide to understand the relationship between emotions and corresponding prosody features for various languages. For evaluating prosodic aspects of emotional Marathi speech, the Marathi language has received less attention. This study aims to see how different emotions affect suprasegmental properties such as pitch, duration, and intensity in Marathi's emotional speech. This study investigates the changes in prosodic features based on emotions, gender, speakers, utterances, and other aspects using a database with 440 utterances in happiness, fear, anger, and neutral emotions recorded by eleven Marathi professional artists in a recording studio. The acoustic analysis of the prosodic features was employed using PRAAT, a speech analysis framework. A statistical study using a two-way Analysis of Variance (two-way ANOVA) explores emotion, gender, and their interaction for mean pitch, mean intensity, and sentence utterance time. In addition, three distinct linear mixed-effect models (LMM), one for each prosody characteristic designed comprising emotion and gender factors as fixed effect variables, whereas speakers and sentences as random effect variables. The relevance of the fixed effect and random effect on each prosodic variable was verified using likelihood ratio tests that assess the goodness of fit. Based on Marathi's emotional speech, the R programming language examined linear mixed modeling for mean pitch, mean intensity, and sentence duration.**

Keywords*—Prosodic parameters; a marathi language prosody model; a two-way analysis of variance; linear mixed-effect models; r programming language

I. Introduction

The COVID-19 pandemic has drastically altered people's lifestyles in many parts of the world. The lockdowns and social distancing norms eventually increased human-machine interaction applications. If computers can recognise emotions, they can communicate in a human-like manner. The prosodic features employed for emotion recognition play an essential role in the quality of the human-computer interaction that replicates human speech emotions. Supra-segmental features or the prosody features such as intensity, pitch, duration, etc., contribute additional information to speech known as paralinguistic information [1-4] and characterize the emotional speech. Developing a prosodic model for emotional utterances for less-studied languages is very challenging. It entails a lot of work, such as creating a database, processing it for analysis, investigating the fluctuation of prosodic elements about emotions using acoustic analysis, and establishing the relevance of these aspects using statistical analysis. In India's Maharashtra and Goa states, the Marathi language is spoken by over 73 million people. In comparison, in the Marathi language, there is less research on prosody aspects for emotional speech. Few of them includes, the syudy of the effect of focus shift in Subject-Object-Verb type Marathi sentences on prosodic features such as F0, duration, and intensity variations[5]. Authors analyzed that the speakers consistently provide acoustic cues with increased duration, higher mean F0, and higher intensity, differentiating focus location. The authors of [6] used broadcast radio transmission Marathi news that are available to the general public to investigate the significant prosodic aspects of the Marathi news-reading style. The authors observed prominence and boundary as the important prosody cues for Marathi's news reading style. Acoustically, the boundaries showed pre-boundary lengthening and pitch contour slope on the final syllable, and the prominence correlated with maximum F0 and maximum intensity and lesser duration. The authors analyzed MFCC features and energy ratios in [7] to investigate the Marathi emotion recognition for anger and happiness. The authors observed the anger emotion recognition rate higher than happiness and neutral emotions. However, the authors suggested generating more emotional speech databases from skilled Marathi speakers.

Considering comparatively less work towards prosody features of Marathi language and lack of emotional database from trained speakers, the paper focuses on emotion analysis for the Marathi language. We constructed a Marathi emotion database from professional speakers with a theatre background in a recording studio expressing anger, fear, happiness, and neutral emotions. The detailed study of the relationship between acoustic features such as mean intensity, mean pitch, sentence duration, and the emotions such as anger, happiness, fear, and neutral for Marathi's emotional speech showed various prosodic cues based on the emotions. Also, a comprehensive statistical analysis was conducted to construct a practical framework for assessing emotional speech data. A two-way ANOVA test for emotion, gender, and their interaction for mean pitch, mean intensity, and sentence time, as well as a linear mixed-effect analysis, were used in the statistical study. The LMM analysis is used to examine the relationship between the emotions and prosodic features data while considering the impacts of fixed and random effect variables and their connection. The two-way ANOVA analysis and a linear mixed model (LMM) analysis contributed while selecting the optimal prosodic features for constructing a

prosody model.There is no comparable effort for the Marathi language that we are aware of.

This prosody model for the Marathi language using the acoustic and statistical investigation will help develop a human-machine interaction application such as emotion recognition from speech can help interpret students' answers and fit pupils with various learning abilities, Text-to-Speech systems (TTS) for Marathi storytelling, speaker recognition, speech recognition, online education etc. among other things.

The remainder of the paper is structured as follows: Section 2 contains a literature review, Section 3 explains the methodology and implementation for creating a Marathi database for various emotions and calculating prosodic features using the PRAAT speech analysis framework, Section 4 focuses on the results and discussion based on acoustic and statistical analysis to prepare a Marathi prosody model, and Section 5 summarises all of the discussions.

II. Literature Review

Speech is an important channel for the communication of emotion, yet studied little in the context of emotion. Speech conveys linguistic messages and includes a major paralinguistic part, prosody. The prosody of speech is defined in the linguistic literature as the suprasegmental properties of speech and include the pitch/F0, loudness/intensity, and rhythm/duration aspects (Brown 2005). Analyzing prosody features based on emotional speech is central to a few emotions. Although emotion classifications, in reality, are much larger, the majority of emotional speech statistics comprise four or eight emotions. Variations in prosodic elements concerning emotions, on the other hand, differ among languages and are dependent on culture and speaking style. As a result, it is vital to investigate the prosodic features for the emotional expressions specific to the language and culture. Fundamental frequency (F0), intensity, and duration are the essential acoustic characteristics influencing prosody. Fundamental frequency (F0) or pitch is the number of vibrations per second produced by the vocal cords, and the relative highness or lowness of a tone perceived by the ear determines pitch in speaking. The length of time a sentence, word, or syllable exists is called its duration. The intensity of a sound measures the energy contained in a given waveform signal. It is essential to analyze the prosodic features of emotion expression specific to the language and culture as emotions differ according to the cultural backgrounds, several international and national languages; the researchers are looking for acoustic correlates of prosody. Table I compares various prosody features studied in different languages based on distinct emotions and the corresponding dominant emotional signaling in the respective language.

Researchers often analyzed the database statistically after acoustic analysis to validate the acoustic analysis results and then select the best prosodic features to construct a prosody model [14-19]. Analysis of variance (ANOVA) findings investigate statistical discriminations of prosodic properties between various emotion classes. The success of ANOVA in identifying the best prosodic qualities to model the emotion recognition system has significantly reduced signal evaluation time. Hence, we have carried out a a two-way ANOVA analysis and linear mixed model (LMM) statistical analysis to design a prosody model for various emotions for Marathi. The LMM refers to using both fixed and random effects on the variables in the same analysis [20-23]. Due to the differences in prosodic variation patterns based on emotions, we examined three separate LMM models, one per prosody feature.

TABLE I. Comparison of Acoustic Feature Variations based on Various Emotions in different Languages

Reference	Language Studied	Emotions	Prosody Features	Emotional Signaling
Bansal S., Agrawal, S., Kumar, A., 2019[8]	Hindi	neutral, fear, anger, surprise, sadness, and happiness	pitch, intensity and duration	The most intense emotion is anger, followed by neutral, happy, surprise, sadness, and fear. For all emotions, the pitch fluctuates in accordance with the intensity feature of speech.
J. Kaur, K. Juglan, V. Sharma, 2018.[9]	Punjabi	happiness, anger, fear. Sad, neutral	Mean Pitch, Intensity and formants	Mean pitch highest for happiness and lowest for sad, Intensity is highest for anger and lowest for fear.
Swain, M.; Routray, A., 2016.[10]	Odia	anger, fear, happiness, disgust, sadness, surprise.	pitch, energy, duration, and formant	In both males and females, the feeling "happy" has the greatest mean pitch value, followed by "surprise" in a close second. All other emotions have significantly lower energy levels than disgust and fear. Female respondents showed no discernible differences in the amount of energy levels for distinct emotions.
Hellbernd, N.; & Sammler, D., 2014.[11]	German	Criticism, naming, suggestion, doubt, warning, wish	Mean duration, mean intensity, mean F0, Pitch rise, harmonic-to-noise ratio	The loudest and most arching pitch contour were seen in warning stimuli. Naming stimuli having a low mean pitch, flat pitch contour, and low intensity.
Rao, K.; Koolagudi, S., 2013.[12]	Telugu	Anger, Disgust, Fear, Compassion, Neutral, Happiness, Sarcasm, Surprise	Mean duration, mean pitch, mean energy	Anger emotion with the highest energy Anger, happiness and neutral have high pitch values
Liu, P., Pell, M.D. 2012.[13]	Mandarin	anger, sadness, happiness, disgust, fear, pleasant surprise, neutrality	mean fundamental frequency, amplitude variations, speech rate (in syllables per second). mean harmonics-to-noise ratio, HNR variation	Anger and pleasant surprise had comparatively high mean f0 values and significant f0 and amplitude variations, but sadness, disgust, fear, and neutrality had relatively low mean f0 values and minor amplitude variations, while pleasure had a moderate mean f0 value and f0 variation.

III. Methodology and Implementation

Because prosody varies by language and speaking style, studying the relationship between emotions and the accompanying prosody variants is vital for all languages. This study aims to see how the prosodic aspects of Marathi's speech change with emotions, and four sub-questions are investigated concerning it as below.

1) Do Marathi speakers employ changes in prosody elements to help them communicate the emotion they want to convey in their speech?

2) If so, what precise variations in an utterance's prosody are used by speakers to differentiate one emotion from another?

3) Is it possible to create a predictive statistical model of prosody variation based on emotions in Marathi that can be utilized as a prosody model for a variety of applications such as emotion recognition, speaker recognition, speech recognition, text to speech synthesis systems, etc.?

4) Is it possible to consider neutral emotion as a baseline and analyze variations of prosodic features concerning neutral emotion and be used for emotion conversion applications?

The workflow for conducting out the research is depicted in the steps below.

1) Collection of sentences.

2) Selection of trained speakers.

3) Recording the sentences in anger, happiness, fear, and neutral read-out style emotions.

4) Collecting the database in .wav format.

5) Processing the .wav files with segmentation, annotation, and creating corresponding text grid files in the PRAAT speech processing toolbox.

6) Calculating the mean pitch, mean intensity, and sentence duration for all the .wav files.

7) Calculating and analyzing acoustic behavior of the above prosodic features based on the emotions, gender, speakers, and sentences.

8) Statistial analyzation of these prosodic features using two-way ANOVA and LMM analysis.

The corpus was constructed by identifying the sentences for the recordings, finding expert Marathi speakers, practicing, and recording their acted utterances in a recording studio. Each line was deliberately crafted to avoid provoking any emotion. There were three to nine words in each sentence. Eleven Marathi professional artists, four females and seven males with experience in drama and television, aged 18 to 40, participated in the experiment. The research objective was conveyed to the speakers, and two practice sessions were arranged to get acquainted with the sentences. The participants were paid incentives for this work. The selected ten sentences from different Marathi storybooks listed in Table II along with their English translations.

TABLE II. The English Translation of Ten Marathi Sentences used for Recording

	Marathi Sentences	**English Translation**
1.	*आम्ही पण दहा बाय दहाच्या खोलीत राहतो .*	We stay in 10 by 10 room.
2.	*लेकीला सांगा तिचा बाबा आलाय.*	Tell daughter that her father has come.
3.	*मन मोठं असलं की सारं काही सामावून घेता येतं.*	Big heart accommodates everything.
4.	*अन्न वाया घालवू नये, त्याची किंमत कमवायला लागल्यावर कळेल.*	You will value food when you start to earn.
5.	*प्रत्येक दगड हा देव होतोच असं नाही.*	Every stone does not become God.
6.	*अंथरुण पाहून पाय पसरावे.*	Spend as per your earning.
7.	*गरीब माणसाची गम्मत करू नये.*	Do not make fun of poor people.
8.	*भाकरीची किंमत घाम गाळल्याशीवाय कळत नाही.*	You never understand value till you won't work for it.
9.	*डोकं शांत असेल तर निर्णय चुकत नाहीत.*	A calm mind takes always the right decision.
10.	*तुमचं बरोबर आहे.*	You are right.

Each speaker repeated the given sentences with different emotions such as anger, happiness, fear, and neutral. The speakers recorded the utterances in a recording studio with a condenser microphone and a digital audio tape (DAT) recorder using a lossless 44kHz, 16bit audio format and saved at a sampling rate of 16kHz. Each speaker initially recorded ten sentences in a single emotion during the recording. Between the two sentences, the speakers left a reasonable pause. After recording all ten sentences in one emotion, the speakers took a short rest before recording all ten sentences in another emotion. The recording procedure took over three months to complete. Each speech file was an a.wav file with 2-4 seconds duration. The entire database of 440 sentences (eleven speakers, ten sentences, and four emotions) was available for further study. Each line was listened to by fifteen people (twelve Marathi native speakers and three non-Marathi speakers). They were able to identify emotions such as anger, happiness, fear, and neutrality in each recorded voice recording. The perceptually verified sentences were segmented in a PRAAT Text Grid. The .wav file of all the sentences is annotated manually in a sentence and word level for better accuracy. We observed variations in pitch contour, intensity contour, and duration for the same sentence comprising four emotions uttered by every speaker. It showed that there is some relationship that exists between the emotional utterances and corresponding prosodic features even for the Marathi language. The mean values of mean pitch, mean intensity and sentence duration of 422 sentences were calculated using the PRAAT speech analysis framework. The mean pitch was calculated by 'getting the Pitch' command and the mean intensity was calculated by 'getting Intensity (dB)' by selecting the sound interval in the PRAAT editor window. Sentence duration is calculated by selecting the portion of the utterance and reading the duration of the selection (in seconds) from the duration bar from the PRAAT editor window.

IV. Result and Discussion

A. Acoustic Analysis

The mean and standard deviations of prosodic parameters such as mean pitch, mean intensity, and sentence duration were determined for all 422 utterances to check for variation in mean pitch, mean intensity, and sentence duration for distinct emotions. Table III shows the overall descriptive statistics for the three prosodic variables for all four emotional utterances.

From Table III, we can see that the amount of variation or standard deviations (SD) are high for the mean pitch with 25.58% and the sentence duration with 27.62%, while mean intensity appears fairly consistent with the amount of variation (SD) of 6.6%. The standard deviation provides some insight into the patterns of variation occurring within the data. We analysed the variations of means and SDs of all three prosodic variables independently of anger, happiness, fear, and neutral emotions to acquire a clear picture of the prosodic variations based on emotional utterances, as shown in Table IV.

Table IV shows that the mean and standard variation values of all the three prosodic features for anger and happy emotions, and fear and neutral emotions, are nearly identical. To understand variations of prosody features for emotions, other factors such as gender, speakers and sentences are also important. Fig. 1 gives the analysis of variations of mean pitch, mean intensity and duration for gender, speakers and sentences.

Fig. 1a demonstrated substantial differences in mean pitch values by gender, with males having lower values than females. Fig. 1b showed that males have variability in mean intensity than females, and Fig. 1c showed that both genders with similar observations for utterance duration. Fig. 1d, 1e, and 1f showed the variations of mean pitch, mean intensity, and sentence duration among the multiple speakers of the same gender. Fig. 1g, 1h, and 1i show the variations of the prosodic features concerning the ten different sentences. Fig. 1 shows that in the Marathi language, prosodic features vary for emotion change as well as variations in gender, speaker, and sentence. In Fig. 2, gives variation of prosodic features based on emotion and gender.

TABLE III. Descriptive Statistics for the Three Prosodic Variables

Prosodic Variables	Mean	Std. Deviation	Percentage
Mean pitch	217.70 Hz	55.69 Hz	25.58%
Mean intensity	69.74 dB	4.615 dB	6.6%
Sentence duration	2.69 sec.	0.743 sec.	27.62%

TABLE IV. Summary of Means and SDs of Each Prosodic Variable by Emotions

Emotion	Mean pitch (Hz)		Mean Intensity (dB)		Duration (sec.)	
	Mean	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Anger	254.7	46.6	73.69	2.58	2.27	0.57
Happiness	241.4	51.5	71.02	2.86	2.64	0.63
Fear	193.9	43.3	67.40	4.15	2.86	0.73
Neutral	179.5	42.2	66.77	4.71	3.01	0.82

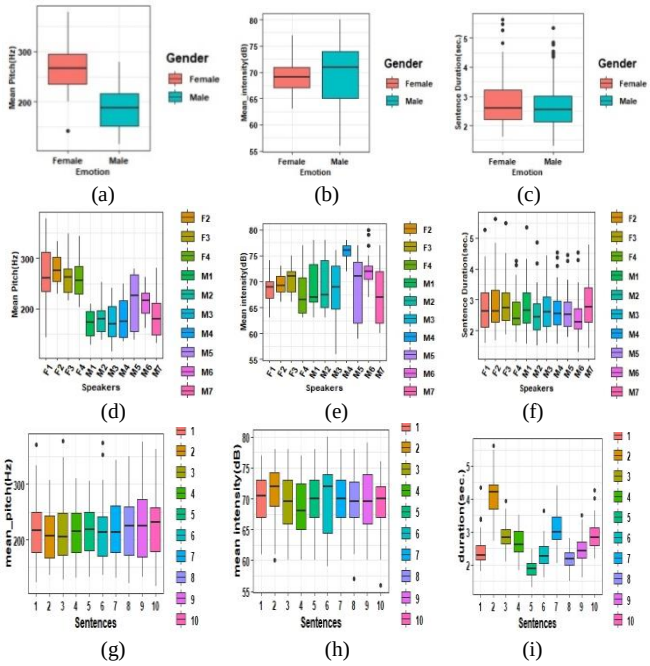


Fig. 1. Box Plots Showing Variations in mean Pitch mean Intensity and Sentence Duration (Figure 1a, 1b, and 1c) for Gender, Speakers (Fig. 1d, 1e, 1f) and Sentence Duration Due to Multiple Speakers (Fig. 1g, 1h, 1i).

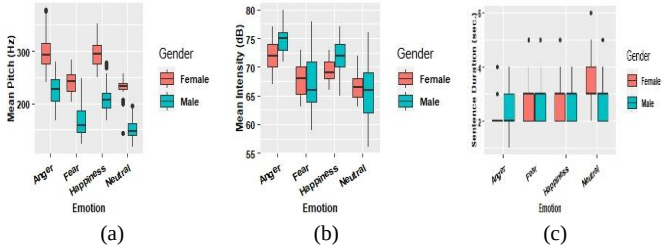


Fig. 2. Variations in (a). Mean Pitch, (b). Mean Intensity, (c). Sentence Duration for Emotions and Gender.

Fig. 2a and 2b showed variations in mean pitch and mean intensity for gender for all the emotions. There was little change in mean intensity levels for the fear emotion between male and female speakers. Each gender takes the same amount of time to speak the lines in fear and happiness observed in Fig. 2c. Female speakers took less time to express anger than male speakers, whereas female speakers' utterance duration was higher for neutral speaking style sentences. As a result, acoustic analysis of mean pitch, intensity, and duration revealed that prosodic features behave differently for emotions and gender.

For Hindi language, acoustic correlation of emotions were analysed for prosodic parameters such as pitch, intensity and duration in [24]. Authors have generated Hindi speech database with 10 speakers in all six emotions such as anger, Fear, Happy, Neutral, Sad and Surprise. Comparison between the prosodic parameters related to the emotions for Marathi and Hindi language given in Table V.

From Table V, comparing the prosodic variation patters for anger, happiness, fear and neutral emotions the pitch, intensity

and duration features observed with the following order by taking neutral emotion as reference.

Pitch : Neutral > Anger > Happiness > Fear (for Hindi).

Pitch : Anger > Happiness > Fear > Neutral (for Marathi).

Intensity : Happiness >Anger > Neutral > Fear (for Hindi).

Intensity : Anger > Happiness > Fear > Neutral (for Marathi).

Duration : Fear > Happiness > Neutral >Anger (for Hindi).

Duration : Neutral > Fear > Happiness > Anger (for Marathi).

When the behavior of variations of prosodic elements dependent on emotions is compared for the two Devnagari languages of India, Hindi, and Marathi, the relevance of studying each language separately for prosodic patterns based on the emotions becomes clear.

B. ANOVA Analysis

A two-way ANOVA analysis explored the impact of emotions and gender and their interaction on the mean pitch, mean intensity, and duration. Both emotion ($F = 205.7$, $p < 0.0001$) and gender ($F=872.1$, $p < 0.0001$) were significant for mean pitch, suggesting that gender is more responsible for mean pitch variations. The p-value for the interaction between emotion and gender was non-significant ($F =1.468$, $p=0.223$) indicated that the relationships between gender and mean pitch was independent of the emotion. For mean intensity, emotion ($F = 104.2$, $p< 0.0001$) and gender ($F= 9.528$, $p <0.001$) were statistically significant with emotion as the most significant factor variable. The p-value for the interaction between emotion and gender for mean intensity observed to be significant ($F =7.274$, $p <0.0001$) indicated that the relationships between gender and mean intensity depends on emotion. For sentence duration, emotion factor observed to be statistically significant($F = 16.142$, $p< 0.0001$) but gender as non-significant ($F= 0.923$, $p =0.337$). The p-value for the interaction between emotion and gender non-significant ($F =0.709$, $p = 0.547$) indicated that the relationships between gender and sentence duration depend on emotion.

TABLE V. THE PROSDIC FEATURE VALUES FOR MARATHI AND HINDI LANGUAGE FOR ANGER, HAPPINESS, FEAR AND NEUTRAL EMOTIONS

Prosodic Parameters	Emotions	Marathi	Hindi
Mean pitch (Hz)	Anger	254.7	303
	Happiness	241.4	300
	Fear	193.9	295.5
	Neutral	179.5	304.4
Mean Intensity (dB)	Anger	73.69	84
	Happiness	71.02	84.5
	Fear	67.40	81
	Neutral	66.77	83
Duration (sec.)	Anger	2.27	1.39
	Happiness	2.64	1.67
	Fear	2.86	2.8
	Neutral	3.01	1.66

C. LMM Analysis

Differences in prosodic features in Marathi are attributable to emotion fluctuations and gender, speaker, and sentence variations. Also, even if the independent variables such as emotions and gender have a somewhat consistent impact on prosodic feature variations, it can vary amongst speakers of the same gender or even between the different sentences. Linear mixed models are a type of regression model that takes into account variation explained by the independent variables of interest, known as fixed effects, and variation not explained by the independent variables, known as random effects [22]. The model is mixed since it combines both fixed and random effects. Thus, to calculate variations in prosodic features, emotion and gender factors are of primary interest and added as fixed effect variables. The emotion factor with four factors: anger, happiness, fear, and neutral emotions and gender factors included males and females: the speaker and the sentence considered random effect variables. Each prosodic feature was then verified for the model fit considering these factors. The goodness of fit of prosodic features for fixed-effect and random-effect variables, as shown in equation (1).

$$\text{Prosodic feature} = \text{emotion} + \text{gender} + (1/\text{speaker}) + (1/\text{sentence}) \quad (1)$$

Equation 1 shows the variations in the prosodic feature for the variations in emotion and gender as fixed effect variables and speakers and sentences as random effect variables with 1/speaker and 1/sentence as random intercept different for each speaker and each sentence individually.

The likelihood ratio tests to assess the goodness of fit to verify the significance of the fixed effect and random effect for each prosodic variable. The goodness of fit test confirmed the relevance of the fixed and random effect variables for prosodic feature variations.

1) Modeling mean pitch: The impact of fixed-effect variables on the mean pitch calculated by comparing the null effect models with fixed effect = 1 and two models with emotion as a fixed effect factor and the other model with emotion and gender as two fixed-effect elements shown in equation (2), (3) and (4) respectively.

$$\text{Mean pitch} = 1 + (1/\text{speaker}) \quad (2)$$

$$\text{Mean pitch} = \text{Emotion} + (1/\text{speaker}) \quad (3)$$

$$\text{Mean pitch} = \text{Emotion} + \text{Gender} + (1/\text{speaker}) \quad (4)$$

Chi-square difference tests showed the significant p-value of emotion with $\chi2(1) = 434$, $p < 0.001$ and of gender with $\chi2(1) = 23$, $p < 0.001$. Both emotion and gender observed to be significant with $p<0.001$ and considered fixed effect variables for mean pitch modeling.

In addition, likelihood ratio tests examined the goodness of fit and confirmed the relevance of the random effect variables' influence on the mean pitch. We compared the two null effect models, with fixed effect = 1 and one without sentence intercept, and sentence intercept with the following equation (5) and (6) respectively.

$$\text{Mean pitch} = 1 + (1/\text{speaker}) \quad (5)$$

Mean pitch = 1 + (1/speaker) + (1/sentence) (6)

Comparing the models with Chi-square difference tests, resulted $\chi 2(1) = 0$, $p > 0.01$. It showed that the inclusion of a sentence is not significant for mean pitch calculation since it does not improve the model fit.

The final design of the model fit to calculate the mean pitch model for Marathi emotional speech calculated as shown in equation (7) as below,

Mean pitch = Emotion + Gender + (1/speaker) (7)

As in equation (7), emotion and gender factors are fixed effect variables, and the speaker is a random effect variable for calculating the mean pitch model for Marathi's emotional speech. Table VI shows the impacts of fixed effect variables such as emotions (angry, happiness, fear, and neutral) and gender (male and female) on computing mean pitch values.

TABLE VI. FIXED EFFECT SUMMARY OF MEAN PITCH

Emotions	Estimate	Std. Error	t- value
Intercept (Neutral)	228.19	7.257	31.45***
Anger	75.51	3.19	23.64***
Fear	15.47	3.2	4.84***
Happiness	61.49	3.19	19.30***
Male	-77.40	8.756	-8.840***

*** =p< 0.001

The neutral emotion is used as an emotion baseline, while the female gender is a gender baseline. The estimate of the intercept value of 228.19 indicates that the mean pitch value for neutral emotion and female gender is 228.19Hz. The mean pitch values for other emotions were calculated from Table V based on the neutral emotion mean pitch value estimates. The estimate for the mean pitch value of anger emotion is 228.19+ 75.51= 303.7Hz, which is significantly higher than for neutral emotion ($t= 31.45$, $p<0.001$). Similarly, the estimate for the mean pitch value of fears emotion is 228.19 + 15.47= 243.66Hz, which is significantly higher than for neutral emotion ($t=4.98$, $p<0.001$). Similarly, the estimate of the mean pitch value for happy emotion is 228.19 + 61.49= 289.68 Hz, and this is significantly higher than for neutral emotion ($t= 19.30$, $p<0.001$). Also, the estimated value of the mean pitch of males of -77.40 based on a baseline of female gender means the pitch of males is lower than that for females by 77.40Hz. With this we can calculate mean pitch values for male gender for each of the emotion as; anger = 303.7 – 77.4 = 226.3Hz, happiness = 289.68 – 77.4 = 212.28Hz and fear =243.66 – 77.4 = 166.26Hz.

The Fixed Effects table, similar to most methods such as ANOVA, MANOVA, multiple regression analyses only focuses on group differences in changes in mean pitch values for emotions and gender. Understanding the mean pitch change at both the group and individual levels will be helpful to capture a complete overview of developmental changes in mean pitch values. Table VII summarizes the random effect of individual speakers on the mean pitch model design below.

TABLE VII. RANDOM EFFECT ANALYSIS FOR MEAN PITCH

Groups Name	Variance	Std. Dev.
Speaker	181.3	13.46
Residual	534.5	23.12

The variance due to the speaker is 181.3 and hence the standard deviation of 13.46Hz. This means that there can be variations in the fixed effect values due to variability between the individual speakers. The residuals are the random deviations from the predicted values that are due to some factors outside of the purview of the experiment. The estimate of the residual variance, with a standard deviation equal to 23.12Hz, represents the variability in individual emotion pitch values due to unknown factors.

2) Modeling mean intensity: The relevance of fixed effects on mean intensity was established by comparing null effect models with fixed effect = 1 to two models, one with emotion as the fixed effect factor and the other with emotion and gender as fixed effect factors as shown in equation (8), (9) and (10).

Mean intensity = 1 + (1/speaker) (8)

Mean intensity = Emotion + (1/speaker) (9)

Mean intensity= Emotion + Gender + (1/speaker) (10)

Chi-square difference tests showed the significant p-value of Emotion with $\chi 2(1) = 274.63$, $p <0.001$ and gender with $\chi 2(1) = 0.48$, $p > 0.1$. This means, emotion factor is significant for variations in mean intensity, but gender is non-significant.

A chi-square difference test calculated the inclusion of a random effect structure with random intercepts for speakers and sentences as shown in equations (11) and (12), respectively.

Mean intensity = 1 + (1/speaker) (11)

Mean intensity = 1 + (1/speaker) + (1/sentence) (12)

Comparing the models with Chi-square difference tests, we conclude that sentence inclusion is not significant for mean intensity calculation since it does not improve model fit, $\chi 2(1) = 0$, $p > 0.01$.

The final design of the model fit to calculate the mean intensity model for Marathi emotional speech calculated as shown in equation (13) as below.

Mean intensity = Emotion + (1/speaker) (13)

Table VIII gives the summary of fixed effect variables for calculating the mean intensity.

TABLE VIII. FIXED EFFECTS SUMMARY FOR ANALYSIS OF MEAN INTENSITY

	Estimate	Std. Error	t value
Intercept (Neutral)	66.76	0.73	91.39***
Anger	7.02	0.41	17.31 ***
Fear	0.68	0.41	0.096
Happiness	4.22	0.40	10.44 ***

*** =p< 0.001

The estimate of the intercept value of 66.76 indicates that the mean pitch value for neutral emotion 66.76dB. The estimates of mean intensity values for other emotions are calculated based on the estimates of intercept, i.e., neutral emotion mean intensity. The estimate for mean intensity of anger emotion is 66.76 + 7.02= 73.78 dB and this is significantly higher than for neutral emotion (t= 17.31, p<0.001). The estimate for mean intensity for fear emotion is 66.76 + 0.68 = 67.44 dB and this is not showing any significance with neutral emotion (t=0.096, p > 0.1). Similarly, the estimate for mean intensity for happiness emotion is 66.76 + 4.22 = 70.98 dB and this is significantly higher than for neutral emotion (t= 10.44, p<0.001).

Table IX summarizes the random effect of individual speakers on the mean intensity model design below.

TABLE IX. RANDOM EFFECT ANALYSIS FOR MEAN INTENSITY

Groups Name	Variance	Std. Dev.
Speaker	4.956	2.226
Residual	8.619	2.936

The variance due to speaker is 4.956, indicating the standard deviation in mean intensity is 2.23dB in the fixed effect values due to variability between the speakers. The residuals are the random deviations from the predicted values, with a standard deviation equal to 2.96dB representing the variability in intensity apart from speakers.

3) Modeling duration: The significance of fixed effects on the sentence duration was determined by comparing null effect models, where fixed effect = 1 and the two models one with fixed effect factor as emotion and the other with fixed-effect factors as emotion and gender as shown in equation (14), (15) and (16) respectively.

$$\text{duration} = 1 + (1/\text{speaker}) \quad (14)$$

$$\text{duration} = \text{Emotion} + (1/\text{speaker}) \quad (15)$$

$$\text{duration} = \text{Emotion} + \text{Gender} + (1/\text{speaker}) \quad (16)$$

Chi-square difference tests for duration showed the significant p-value for emotion with $\chi2(1) = 640.8$, $p < 0.001$ but non-significant for gender with $\chi2(1) = 2$, $p = 0.15$. This suggests that the variation in sentence duration is due to emotion rather than gender.

Also, the two null effect models, one without speaker intercept and the other with a sentence and speaker intercept, compared to determine the duration model fit for random effect variables as shown in equation (17) and (18) respectively.

$$\text{duration} = 1 + (1/\text{speaker}) \quad (17)$$

$$\text{duration} = 1 + (1/\text{speaker}) + (1/\text{sentence}) \quad (18)$$

The Chi-square difference tests showed that inclusion of sentence as one of the random effects is significant for the mean duration and it improved model fit, $\chi2(1) = 410.73$, $p < 0.001$.

The final design of the model fit to calculate the sentence duration model for Marathi emotional speech calculated as shown in equation (19) as below.

$$\text{duration} = \text{Emotion} + (1/\text{speaker}) + (1/\text{sentence}) \quad (19)$$

Table X gives the summary of fixed effect variables for calculating the duration given by the final design to calculate the duration for emotions equation.

TABLE X. FIXED EFFECTS SUMMARY FOR ANALYSIS OF DURATION

	Estimate	Std. Error	t value
Intercept (Neutral)	2.99	0.19	15.52***
Anger	-0.73	0.04	-16.42 ***
Fear	-0.14	0.04	0.002**
Happiness	-0.36	0.04	-8.104 ***

*** =p< 0.001, ** =p< 0.01

The estimate of the intercept value of 2.99 indicates that the mean sentence duration for neutral emotion is 2.99sec. The estimates of mean sentence duration values for other emotions are calculated based on the estimates of intercept, i.e., the mean sentence duration for neutral emotion. The estimate for the sentence duration anger emotion is 2.99 - 0.73= 2.26 sec and this is significantly lower than for neutral emotion duration (t= -16.42, p<0.001). The estimate for the sentence duration for fear emotion is 2.99 -0.14 = 2.85 sec and this is not showing any significance with neutral emotion (t=0.002, p = 0.01). Similarly, the estimate for the mean sentence duration of happy emotion is 2.99 -0.36 = 2.63 sec and this is significantly higher than for neutral emotion (t = -8.104, p<0.001).

Table XI, giving the summary of the random effect of individual speaker and individual sentence on the sentence duration model design as below.

TABLE XI. RANDOM EFFECT ANALYSIS FOR DURATION

Groups Name	Variance	Std. Dev.
Speaker	0.119	0.014
Sentence	0.59	0.35
Residual	0.32	0.10

The variance due to speaker is 0.014, and hence standard deviation of 0.119 sec. in the fixed effect values can be due to variability between the speakers. The variance due to the sentence is 0.35 and hence standard deviation of 0.59 sec. in the fixed effect values can be due to variability between the sentences. The residuals are the random deviations from the predicted values, with a standard deviation equal to 0.32 sec. represents the variation in duration values apart from speaker and sentence.

V. CONCLUSION

This work explains how to investigate acoustic clues for Marathi emotions, including anger, happiness, fear, and neutral. Eleven Marathi professional artists created a database of 440 words in anger, happiness, fear, and neutral emotions in a recording studio. According to an acoustic experiment, the features of mean intensity, mean pitch, and sentence duration

vary depending on the emotions. A two-way ANOVA and a linear mixed-effect analysis provided a valuable framework for studying emotional speech data and, as a result, best practices for generating an emotional speech corpus.

The following is the prosodic model for emotions in the Marathi language, with emotion and gender as fixed effect variables and speaker and sentences as random effect variables.

Mean pitch = Emotion + Gender + (1/speaker).

Mean intensity = Emotion + (1/speaker).

duration = Emotion + (1/speaker) + (1/sentence).

A detailed analysis of Marathi's emotional speech will help develop a prosody model. This model will help select appropriate input features for machine learning algorithms used in emotion classification applications. In the future, it will be beneficial to examine some more prosodic aspects for the Marathi language emotions. Sadness, surprise, and sarcasm are examples of other basic emotions that may be investigated for Marathi speech. Children, young adults, and the elderly can all be studied separately in a similar way. When these meticulously generated prosodic elements are fed into a machine learning model, they can aid in emotion recognition, text-to-speech synthesis, and other human-machine interaction applications in the future.

REFERENCES

[1] G. Zhang, S. Qiu, Y. Qin and T. Lee, "Estimating Mutual Information in Prosody Representation for Emotional Prosody Transfer in Speech Synthesis," 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021, pp. 1-5.

[2] T. Wani, T. Gunwan, S. Qadri, M. Kartiwi, "A Comprehensive Review of Speech Emotion Recognition Systems", IEEEAccess, vol. 9, pp. 47795–47814, April 2021.

[3] S. Bharadwaj and P. B. Acharjee, "Analysis of Prosodic features for the degree of emotions of an Assamese Emotional Speech," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1441-1452.

[4] V.Raju, H.Vydana, S.Gangashetty and A.Vuppala, "Importance of non-uniform prosody modification for speech recognition in emotion conditions," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 573-576.

[5] P. Rao, N. Sanghvi, H, Mixdorff, K Sabu., "Acoustic correlates of focus in Marathi: Production and perception", Journal of Phonetics, 2017, vol. 65, pp. 110-125.

[6] S. Barhate, S. Kshirsagar, N. Sanghvi, K.Sabu, P. Rao, and N. Bondale, "Prosodic features of Marathi news-reading style", IEEE Region 10 Conference (TENCON), 2016, pp.2215-2218.

[7] V. Degaonkar, Apte S., "Emotion modeling from speech signal based on wavelet packet transform", Int J Speech Technol, 2013, vol. 16, pp. 1–5.

[8] S. Bansal, S. Agrawal, A. Kumar, "Acoustic analysis and perception of emotions in Hindi speech using words and sentences", Int. j. inf. Technology, 2019, vol. 11, 807-812.

[9] K. Jasdeep, S. Vishal, " Role of Acoustic Cues in Conveying Emotion in Speech", Journal of Forensic Sci & Criminal Inves., 2018, vol. 11(1), pp. 555803.

[10] M.Swain, A. Routra, P. Kabisatpathy, J.Kundu,"Study of prosodic feature extraction for multidialectal Odia speech emotion recognition", IEEE Region 10 Conference (TENCON), 2016, pp. 1644-1649.

[11] N. Hellbernd & D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception", Cognitive Processing, vol. 15, 2014, S46-S46.

[12] K. Rao, S. Koolagudi, R.Vempada, "Emotion recognition from speech using global and local prosodic features", Int J Speech Technol, vol. 16, 2013, pp.143–160.

[13] P. Liu, D. Pell, "Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli", Behav Res, vol. 44, 2012, pp.1042–1051.

[14] H. Chang, S. Young, K .Yuen, "Effects of the Acoustic Characteristics on the Emotional Tones of Voice of Mandarin Tones", Proceedings of 20th International Congress on Acoustics, Sydney, Australia, 2010.

[15] A. Jacob, P.Mythili, "Upgrading the Performance of Speech Emotion Recognition at the Segmental Level", IOSR Journal of Computer Engineering (IOSR-JCE) Volume 15, Issue 3, pp. 48-52, 2013.

[16] T. Iliou,C.Anagnostopoulos, "Classification on Speech Emotion Recognition - A Comparative Study", International Journal on Advances in Life Sciences, vol 2 no 1 & 2, 2010.

[17] S. Ali, M. Andleeb, D. Rehman, "A Study of the Effect of Emotions and Software on Prosodic Features on Spoken Utterances in Urdu Language", I.J. Image, Graphics and Signal Processing, vol. 4, pp.46-53,2016.

[18] M. Yusnita A, Paulraj, S. Yaacobb, N. Fadzilah, Shahriman A., "Acoustic Analysis of Formants across Genders and Ethnical Accents in Malaysian English using ANOVA", International Conference On Design and Manufacturing, vol.64, pp. 385–394, 2013.

[19] A.Meftah, Y. Alotaibi, A. Selouani,"Evaluation of an Arabic Speech Corpus of Emotions: A Perceptual and Statistical Analysis",. IEEE Access, PP. 1-1, 2018. M. Ayadi, M. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, 2011, vol.44, pp. 572-587.

[20] H. Singmann, & D. Kellen, "An Introduction to Mixed Models for Experimental Psychology"'. In D. H. Spieler & E. Schumacher (Eds.), New Methods in Cognitive Psychology, 2019, pp. 4–31.

[21] Sherr-Ziarko, Ethan., PhD thesis, University of Oxford, 2017.

[22] B. Winter, "Linear models and linear mixed-effects models in R with linguistic applications", Cognitive and Information Sciences, University of California, Merced, 2013.

[23] M. Rouch, Undergraduate Honors Theses, Williamsburg VA, 2019.

[24] S.Bansal, S. Agrawal, A. Kumar, "Acoustic analysis and perception of emotions in hindi speech using words and sentences, Int. j. inf. tecnol. 11, pp. 807–812, 2019.

Low Time Complexity Model for Email Spam Detection using Logistic Regression

Zubeda K. Mrisho[1], Anael Elkana Sam[3]
School of Computational and Communication Science and Engineering, The Nelson Mandela Institution of Science and Technology, Arusha, Tanzania

Jema David Ndibwile[2]
College of Engineering
Carnegie Mellon University Africa
Kigali, Rwanda

***Abstract*—Spam emails have recently become a concern on the Internet. Machine learning techniques such as Neural Networks, Naïve Bayes, and Decision Trees have frequently been used to combat these spam emails. Despite their efficiency, time complexity in high-dimensional datasets remains a significant challenge. Due to a large number of features in high-dimensional datasets, the intricacy of this problem grows exponentially. The existing approaches suffer from a computational burden when thousands of features are used (high-time complexity). To reduce time complexity and improve accuracy in high-dimensional datasets, extra steps of feature selection and parameter tuning are necessary. This work recommends the use of a hybrid logistic regression model with a feature selection approach and parameter tuning that could effectively handle a big dimensional dataset. The model employs the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method to mitigate the drawbacks of Term Frequency (TF) to obtain an equal feature weight. Using publicly available datasets (Enron and Lingspam), we compared the model's performance to that of other contemporary models. The proposed model achieved a low level of time complexity while maintaining a high level of spam detection rate of 99.1%.**

Keywords*—Machine learning; feature selection; feature extraction; parameter tuning

I. Introduction

Email is an online application that enables the exchange of data using electronic devices [1]. Email communication is quick, inexpensive, easy to duplicate, and widely available. Email can extremely be beneficial to businesses and organizations because it allows for the efficient, productive, and effective transmission of all types of electronic data [2]. Email communication began in the 1960s with the restricted functionality of sending information to users within the same computing environment only [3]. Recently, email has become the most common way of communication [4], serving users across computing platform environments. The average number of emails exchanged per day reached 293 billion in 2019 and is forecasted to reach 347 billion by the end of 2023 [5].

Despite its importance, email has become a vehicle for a variety of malicious programs [6]. It is estimated that 50% of all emails are spam [1]. Email spam, also known as junk mail, refers to any form of undesired, uninvited digital communication sent in large quantities [7]. Spam is usually sent via email [8] but can also be delivered via text messages, phone calls, or other social media platforms. Spam has been a big challenge, disturbing users and consuming their time. Spam also leads to phishing attacks, storage space misuse, decreased internet speed, and theft of critical information [5]. The financial losses caused by email spam are estimated to reach a total of USD 257 billion between 2012 and mid-2020 [9]. As a result, substantial negative impacts on the global economy, such as lower productivity have been identified. These factors hinder the development of the communication sector that can benefit governments, individuals, and business companies [10].

To combat the problems, various scientific research studies have been conducted, including the application of machine learning [11]. Previous scientific studies were categorized into three approaches, single-based machine learning, hybrid, and feature engineering [12]. In the first classification, a specific single machine learning algorithm was used to build a spam detection method [12]. Some popular classifications of machine learning algorithms include Naïve Bayes, Random Forest, Support Vector Machines (SVMs), and K-nearest neighbor (KNN) [5].

Support Vector Machines are supervised learning models, which are mostly used to analyze data for regression analysis and classification [13]. Every data item is plotted as a point in n-dimensional space where n is the number of features present with the value of each feature being that of a certain coordinate in the SVM algorithm. The classification is accomplished by finding the hyper-plane that best differentiates the two classes. Support Vector Machines achieves great accuracy on small, clear datasets but performs poorly on larger, noisy datasets with overlapping classes [14].

Naive Bayes is a machine learning classification algorithm commonly used for binary and multi-class classification problems. This algorithm is based on the Bayes Theorem, which states that given the known independent probability of each event and the reverse conditional probability of the pair of events, one may compute an unknown conditional probability of the pair of events [15]. The disadvantage of this method is that it makes assumptions that all attributes are independent, which is incorrect. In fact, by recognizing that some attributes are related, one can create patterns or common attributes from related attributes to minimize the number of features, hence reducing storage consumption.

Random Forest is a classifier that uses the number of decision trees on separate subsets of a dataset and averages their results to enhance its predicted accuracy [16]. Instead of

relying on one decision tree, Random Forest collects the forecasts from every tree and calculates the final output based on the majority vote of predictions. The technique is well-suited to classification problems with small datasets because a large number of trees may make it slow for real-time prediction.

K-nearest neighbor, also called Lazy Learner is another learning algorithm that works well in simple classifications [17]. When an email is classified, KNN tries to find the K-nearest neighbors by calculating the distance in each prediction. In high dimensional datasets, it becomes challenging for the KNN algorithm to compute the distance in each dimension resulting in poor performance.

A combined machine-learning (hybrid) algorithm generates a new line of spam detection methods. The approach combines a specific machine learning algorithm and other methodologies [12]. Wijaya [18] proposed a hybrid decision tree with logistic regression with a focus on reducing noisy data. Another researcher Dedeturk [5] introduced a model which uses logistic regression combined with an artificial bee algorithm. However, this model faces high computation costs.

The feature engineering classification focuses on offering a new set of features. Farisa [19] proposed an intelligent spam detection method and recognizes the relevant features by categorizing spam features into three categories. These are payload, head features, attachment features. Payload features are those that involve the email body, readability, and lexical features [19], while attachment features are the files that are combined within an email. Despite its benefits, this methodology cannot be used when there is an imbalanced dataset [19].

As reviewed, we identify that machine learning is an efficient method for detecting email spam. However, most of the existing models failed to consider the number of features in high-dimensional datasets, leading to high time complexities. Nevertheless, the finding by Majeed [20] shows that time complexity is an important factor to be considered in model development since it reduces the training speed and decreases the importance of the model to be used in online spam filtering [11]. Time complexity depends on the number of features required in a given model as well as whether the proposed method is linear or nonlinear [21]. Xia [22] proposed an approach based on reducing time complexity in rule-based filtering. Nonetheless, this is not currently a recommended approach due to inefficient results that require every time to change the rule.

High-dimensional datasets are datasets with many features. It is the excess number of features that leads to a high time complexity and sometimes a low detection rate (meaning low accuracy) [23], as illustrated in (1) - (8).

Recall formula for finding accuracy of the model [24].

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Key: TN=True negative, TP=True positive, FN=False negative, and FP=False positive.

Example: Let us take 10,000 features for a high dimensional dataset and 2605 features for a low dimensional dataset with a test size of 0.34. For a high dimensional dataset, the number of correct predictions is described by the confusion matrix Table I.

Then from Table I, TP=1521, TN=1700, FP=136, FN=43.

$$Accuracy = \frac{1521+1700}{1521+1700+136+43} \quad (3)$$

From (3); $Accuracy = \frac{3221}{3400}$ (4)

Therefore, the accuracy of a high-dimensional dataset obtained from (4) = 0.947 (5)

For a low-dimensional dataset, the number of correct predictions is described in Table II.

Then from Table II, TP=98, TN=768, FP=0, FN=10

$$Accuracy = \frac{98+768}{768+0+41+98} \quad (6)$$

From (3); $Accuracy = \frac{866}{907}$ (7)

Therefore, the accuracy for a low-dimensional dataset obtained from (7) = 0.955 (8)

So, from (5) and (8), the accuracy of a high-dimensional dataset seems to be low compared to a low-dimensional dataset.

Therefore, this paper proposes an efficient hybrid model [25] of logistic regression, with the consideration of the time complexity in a high dimensional dataset. Our methodology combines feature extraction, feature selection, and parameter tuning methods. This approach will reduce the time complexity on high-dimensional datasets. It will equally reduce equal feature weight, overfitting, increase training speed and boost performance. The model uses the Big O notation to find the time complexity of different existing models with accuracy starting from 90%. The evaluation involves a calculation of time complexity in terms of the steps required to operate an input.

TABLE I. CONFUSION MATRIX FOR A HIGH DIMENSIONAL DATASET

Actual	Predicted	
	Non-spam	*Spam*
Non-spam	1700	136
Spam	43	1521

TABLE II. CONFUSION MATRIX FOR A LOW DIMENSIONAL DATASET

Actual	Predicted	
	Non-spam	*Spam*
Non-spam	768	0
Spam	41	98

The rest of the paper is structured as follows: the materials and methodology are presented in Section II while the results are discussed in Section III. Finally, the conclusion and future research direction are presented in Section IV.

II. Materials and Methods

A. Experimental Setup

The model was developed using Python (v3.7.1) in the Google Colab (GCC 7.5) environment on a 64-bit Windows operating system, equipped with 8GB of computer Random Access Memory (RAM).

B. Dataset

The experiments were carried out using two datasets derived from a public repository. This helped to validate the accuracy of the model for spam detection. The first dataset was obtained from the Kaggle repository, which was the Enron dataset with 10,000 samples, half of which were spam and half legitimate emails. The second dataset was the Lingspam with 2605 samples, out of which 433 were spam and 2172 legitimate emails. We analyzed the dataset in relation to their balance ratio which is computed by dividing the total number of genuine emails by the total number of spam emails. The balance ratios of Enron and Lingspam were 1 and 5 respectively. The dataset was then split into two, 67% for training and 34% for testing as described in Table III.

C. Pre-Processing

This step involved cleaning the data by removing missing values; transforming the data into a direct format that could be used by machine learning and splitting them for training and testing. Data transformation is a data mining approach that involves changing raw data into a usable format. This is because real-world data is usually inconsistent, inadequate, lacking in specific behaviors or patterns, and rife with mistakes [26]. Data preparation is a tried-and-true approach for overcoming such difficulties. Building a high-performing model needs a careful evaluation of the input data quality. Therefore, the dataset was pre-processed for the suggested model to perform intelligent diagnosis by extracting suitable characteristics from the data. The preprocessing involved several steps such as importation of the data and libraries, cleaning the data by removing missing values; converting the data into a direct format that could be used by machine learning, and splitting them for training and testing. The process of removing missing values and stop words is very important because of their non-informative in the email spam detection process. Apart from removing stop words, characters must also be converted to lowercase before tokenization. In our datasets, no missing values were found, and tokenization was done through the Sklearn library. The splitting test size was 0.34, meaning that 3400 samples of emails for the Enron dataset were used for testing and 6600 for training. For the Lingspam dataset 886 were used for testing and 1719 for training as shown in Table III.

D. Feature Extraction

This step involved converting email messages into a format that could be processed by a machine learning algorithm. Email spam features are obtained from three different methods, namely, the Heuristic approach, Term frequency (TF) analysis, and behavior approach [27]. In the first approach, emails are mined to discover and generate patterns and rules, while in the TF analysis; every word in an e-mail is specified as a feature. The behavior approach builds features based on knowledge about spammers' behavior. This is often gathered via header, attachment, and email flows between groups of e-mail users.

In this study, the Term Frequency Inverse Document Frequency (TF-IDF) method was employed as a feature extraction method. It is a combination of TF and IDF [28]. According to Kadhim [29], this helps to capture features that are more important within the body of an email. The importance of this method is that it reduces the limitation of equal feature weight obtained when TF is used. Term frequency is how many times a term appears in an email and IDF is how many times a term appears in all emails. Suppose an email contains 50 terms, where the term "none" occurs 10 times. Term frequency is obtained as shown in (9) and (10):

$$TF(t) = \frac{\text{Total no.of times a term occur in an email}}{\text{total number of terms in an email}} \quad (9)$$

$$TF(t) = \frac{10}{50} = 0.2 \quad (10)$$

Now let's say we have 5000 emails, and the term "none" occurs 50 times in all emails. Then TF-IDF is obtained as shown in (11) and (12):

$$IDF(t) = log\frac{500}{50} = 2 \quad (11)$$

$$TF - IDF(t) = 0.2 \times 2 = 0.4 \quad (12)$$

Therefore from (12) our TF-IDF (t) is 0.4

E. Feature Selection

Due to the presence of many features in a high-dimensional dataset, feature selection is an important step. This step involves picking up items that are more important to be used in model development [5]. Feature selection leads to less time complexity that increases the potential application in online spam filtering. Training an algorithm using all the features requires a large amount of memory and high time complexity [30]. Hence, reducing the number of features is very important, since it permits the machine learning algorithm to train faster due to the reduction of the number of steps taken to train the model. Additionally, reducing the number of features also eliminates overfitting [31]. This happens when the model fits more data than it needs and starts catching noisy and inaccurate data. Hence, the efficiency and accuracy of the model decrease.

To reduce the time complexity problem, our research used the Sklearn library, which implements the SelectKBest feature selection method. This method only selects the highest scoring features. It is a wrapper method that uses the score function of Chi-square to obtain the features. By using Chi-square only 450 features were selected. Chi-square is a mathematical formula used to determine if there is a relationship between the features and select those with the highest score only as indicated in (13).

$$\text{Chi-square formula} = \sum \frac{(O_{I-E_i})^2}{E_I} \quad (13)$$

TABLE III. DATASET SAMPLE, TRAINING AND TESTING SIZE

Dataset type	Sample size	Training	Test
Enron	10,000	6600	3400
Lingspam	2605	1719	886

Whereby O_i is the number of the class observed and E_i is the number of expected classes when there is no relationship between the feature and the target [32].

F. Proposed Model

Logistic regression is among the most commonly used algorithms for classification [33]. It is an efficiency model with low time complexity [25]. It is used to determine discrete data from a set of variables [34]. In logistic regression, instead of applying a line, we apply an "S" shape that determines the two largest values [35]. The "S" shape is called the logistic function [36] as shown in Fig. 1. It is used to convert every real value between 0 and 1 into another value [37]. The function uses the threshold value, which determines the likelihood of either 0 or 1. A value beyond 0.5 is 1 and below 0.5 is 0. A logistic regression formula can be formed from the linear equation as indicated in (14) – (16).

However, in logistic regression, y can be 0 - 1, so we divide (14) by 1-y. When y=0 we get 0, and when y=1 we get infinity. To match the equation, we need to transform (15) into a logarithm.

$$y = c + m_1x_1 + m_2x_2 + \cdots + m_nx_n \quad (14)$$

$$\frac{y}{1-y} = c + m_1x_1 + m_2x_2 + \cdots + m_nx_n \quad (15)$$

$$log[\frac{y}{1-y}] = c + m_1x_1 + m_2x_2 + \cdots + m_nx_n \quad (16)$$

After transformation, the formula obtained in (16) can be used for logistic regression.

In this research, the logistic regression was trained with 450 features obtained from the SelectKBest as identified in Fig. 2. The results were then optimized using random search with several parameters as shown in Table IV.

G. Parameter Tuning

When training the models in this study, the hyper-parameters were searched to find the ones with the best performance. A random search was used with the parameters as described in subsections 1-3 and values are presented in Table IV.

1) Penalty: This parameter has two options; ridge (L2) or lasso(L1). Both parameters are used in a regression method to reduce the time complexity of the model. However, while ridge is better for a high-dimensional dataset, lasso is better for a low-dimensional dataset.

2) Solver: This parameter has five solvers which are lbfgs, liblinear, sag, saga, and newton-cg. Liblinear is a decent choice for small datasets, while sag and saga are quicker for big ones [38]. Only lbfgs, sag, newton-cg, and saga can handle multinomial loss in multiclass issues.

3) The Inverse of Regularization Strength (c): It is a logistic regression trade-off parameter that affects the intensity of regularization. The larger values of c correlate to less regularization (where we can specify the regularization function).

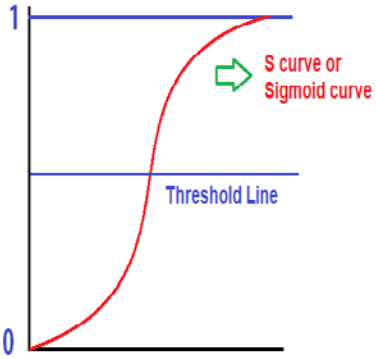


Fig. 1. Logistic Regression Graph [39].

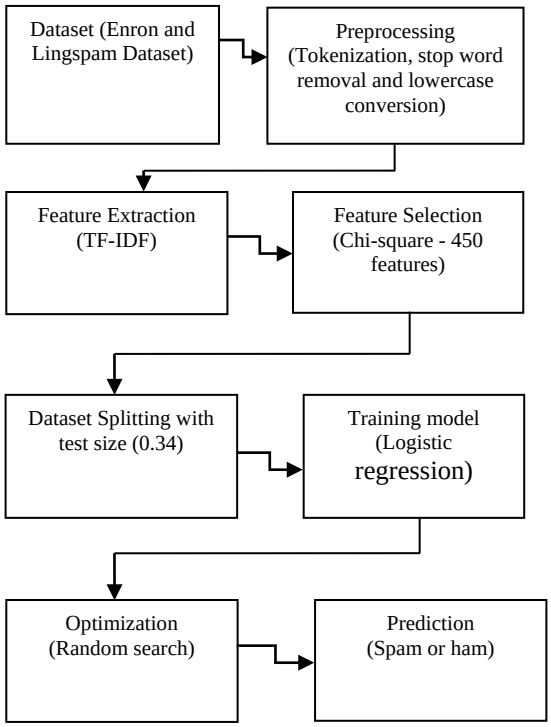


Fig. 2. A Proposed Logistic Regression Model Diagram.

TABLE IV. HYPERPARAMETER USED

Parameter	**Value**
Penalty	L1 and L2
Solver	Saga, sag ,lbfgs, newton-cg, liblinear
C	0.001,0.01,0.1,10,1000

H. Evaluation of the Classifiers

The evaluation of the classifier was evaluated by analyzing the model performance and time complexity during the training and testing procedures. In this study, evaluation was carried out by employing the confusion matrix and Big O notation. The evaluation measures utilized were accuracy, precision, recall, and time complexity each of which is described in subsections 1 – 4 below.

1) Confusion matrix: This is a table that defines the model's overall performance and displays the proper and wrong classifications for each class. Confusion matrix plots are used to show the trained model's ability in guessing the classes of data included in the set of test data. The test set evaluates a model's expected future performance. Table V shows the structure of the confusion matrix using our proposed model. Where TP = True Positive: the number of emails with spam and grouped as having spam, TN = True Negative: the number of emails without spam and grouped as not having spam, FP = False Positive: the number of emails with no spam and grouped as having spam, and FN = False Negative: the number of emails with spam and grouped as not having spam.

a) Accuracy: It computes the frequency with which predictions and labels are equivalent. The accuracy is calculated as shown in (17).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

b) Precision: The number of correctly identified results divided by the total number of positive outcomes results. Equation (18) describes how precision is obtained.

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

c) Recall: The number of accurately recognized positive findings by the total number of samples that should have been positive. A recall is obtained as indicated in (19).

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

2) Big O notation: This is a mathematical study used to describe the complexity of different algorithms [40]. Time complexity is divided into two categories: the number of inputs an algorithm takes to operate and if the algorithm is linear or nonlinear [21]. Each algorithm in machine learning has its formula for finding the time complexity in terms of the steps used to operate an input. This research uses a logistic regression formula to find the time complexity of the proposed model due to its low time complexity compared to others. Table VI presents the formula for finding the time complexity of the different algorithms in machine learning [41]. Given:

O=growth rate of a model.

C=number of the class label for spam is 2spam or ham).

d=number of input/features.

k=number of neighbors, number of support vector.

e=number of epochs.

n=number of neurons.

III. Result and Discussion

This section discusses the results of the proposed model as obtained from the experiments carried out in this research. The results are divided into two parts, the first part shows the time complexity obtained using the Big O notation method as identified in Table VII. The second part shows the performance of the logistic regression when combined with TF-IDF and feature selection method in terms of precision, accuracy, F1-score, and recall as presented in Tables VIII and IX for Enron and Lingspam, respectively.

A. Complexity Result

In machine learning, the time complexity of the model is measured by two things; the type of algorithm used and the number of inputs an algorithm takes to operate. In our model, we used logistic regression which is linear. The advantage of a linear algorithm is its low time complexity relative to non-linear algorithms. Furthermore, when considering the number of inputs, the study used the Big O notation to describe the time complexity of the model as described in Table VI. The result shows that the proposed model attained low time complexity compared to other conventional models as shown in Table VII.

B. Performance Result Analysis

The classifier was evaluated by analyzing the model performance during the training and testing the outcomes. In this study, evaluation was carried out using the confusion matrix as shown in Fig. 3 and 4 whereby 0 represents non-spam and 1 represents spam. The evaluation measures utilized were accuracy, precision, F1 score, and recall. The results showed that saga and L2 parameters are very resourceful parameters in a high-dimensional dataset compared to other solvers because of their performance. Nevertheless, the results show that the feature selection method is an important part to be considered in model development since it reduces the computation time while the optimization process increases model accuracy. The proposed model was compared to other conventional methods and the results showed that the performance of the proposed model was higher than other models as indicated in Table X.

TABLE V. Confusion Matrix Structure

Actual	Predicted	
	Non-spam	*Spam*
Non-spam	TN	FP
Spam	FN	TP

TABLE VI. Formula for Time Complexity

Algorithm	Formula
K-nearest neighbor	O (knd)
Logistic regression	O (nd)
SVMs	O (n^3)
Decision tree	O (n*log (n)*d)
Naïve Bayes	O (n*d)
Deep learning	O(c*d*e*n)

TABLE VII. RESULTS AND COMPARISON ANALYSIS FOR TIME COMPLEXITY

Author	Algorithm and Accuracy obtained	Number of features selected	Time complexity in the training phase(step used)
[5]	Logistic regression- 98.4%	500	O(cd)=2*500 1000 steps
[42]	Deep learning-96.43%	3000	O(c*d*e*n)=2*3000*2*2 24000 steps required
[43]	Naive Bayes-96.87 %	1319	O(c*d)=2*1319 2628 steps required
[44]	Naïve Bayes-96.63%	1000	O(c*d)=2*1000 2000 steps required
[12]	Neural network-96.8%	140	O(c*d*e*n)=2*140*600 168,000 steps required
Proposed approach	Logistic regression- 99.1% & 98.3%	450	O(cd)=2*450

TABLE VIII. RESULTS OF MODEL PERFORMANCE FOR ENRON

Label	Accuracy	Precision	Recall	F1-score
Non-spam(0)	0.98	0.99	0.97	0.98
Spam(1)		0.97	0.99	0.98

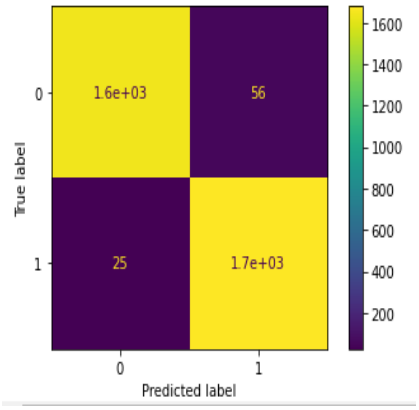


Fig. 3. A Confusion Matrix of Enron Dataset.

TABLE IX. RESULTS OF MODEL PERFORMANCE FOR LINGSPAM

Label	Accuracy	Precision	Recall	F1-score
Non-spam(0)	0.99	0.98	1.00	0.99
Spam(1)		1.00	0.91	0.95

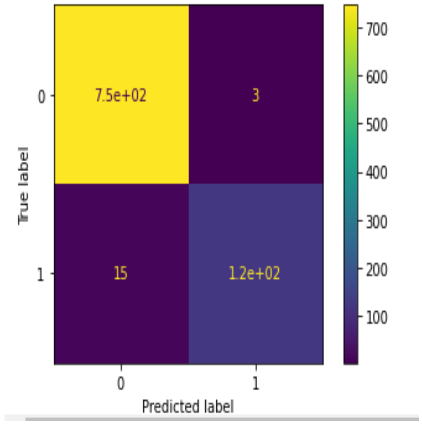


Fig. 4. A Confusion Matrix of Lingspam Dataset.

TABLE X. PERFORMANCE COMPARISON FOR DIFFERENT ALGORITHM

Model	Accuracy
[5]	98.4%
[44]	96.63%
[11]	96.7%
Proposed model	99.1% and 98.3 respectively

IV. CONCLUSION

In this paper, a hybrid logistic regression model was proposed to reduce the time complexity in a high-dimensional dataset that will increase the potential of the model in online spam detection. The model performs three different tasks that are feature extraction, feature selection, and parameter tuning. TF-IDF was used during feature extraction to replace the drawbacks of equal feature weight obtained when TF is used. To increase the training speed in the high-dimensional dataset the model uses Chi-square that helps to select the feature which is related to each other with the highest score only. A random search was used to optimize the model performance. The performed task help to reduce time complexity by decreasing the number of features in a high-dimensional dataset. The model also uses the TF-IDF feature extraction method to reduce the disadvantage of equal feature weight obtained when TF is used. The experiment shows that a better performance of 99.1% is achieved when feature selection is combined with parameter tuning. Overall, it can be concluded that feature selection is an important part of a high-dimensional dataset that helps to reduce an excessive number of features. Nevertheless, for future work, more research is needed in other feature selection and parameter tuning methods.

ACKNOWLEDGMENT

I would like to sincerely thank the Government of the United Republic of Tanzania for funding this research through the Ministry of Education, Science, and Technology (MoEST).

Furthermore, I would like to acknowledge my supervisors, Dr. Jema David Ndibwile and Dr. Anael Sam for their excellent guidance and supervision of this research. Also, I would like to thank my colleague Stephano Amoni for his constant support in this research.

REFERENCES

[1] A. Qashqari, D. Alhbshi, F. Alzahrani, H. Ghwati, and A. Aljahdali, "Electronic Mail Security," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, 2020.

[2] P. A. Gloor, A. F. Colladon, and F. Grippa, "The digital footprint of innovators: Using email to detect the most creative people in your organization," *Journal of Business Research*, vol. 114, pp. 254-264, 2020.

[3] M. Kekane, "New technology in business communication " 2020.

[4] C. Dürscheid and C. Frehner, "2. Email communication," in *Pragmatics of computer-mediated communication*, ed: De Gruyter Mouton, 2013, pp. 35-54.

[5] B. K. Dedeturk and B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Applied Soft Computing*, vol. 91, p. 106229, 2020.

[6] T. A. Kemp, M. C. Depaolis, W. R. Gemza, and R. J. Whalen, "Electronic mail security system," ed: Google Patents, 2020.

[7] R. K. Kumar, G. Poonkuzhali, and P. Sudhakar, "Comparative study on email spam classifier using data mining techniques," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2012, pp. 14-16.

[8] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *Journal of Machine Learning Research*, vol. 7, 2006.

[9] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168261-168295, 2019.

[10] R. Broadhurst and M. Alazab, "Spam and crime," *Regulatory Theory*, p. 517, 2017.

[11] A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal, and A. U. Akram, "A feature-centric spam email detection model using diverse supervised machine learning algorithms," *The Electronic Library*, 2020.

[12] H. Faris, A.-Z. Ala'M, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah, *et al.*, "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks," *Information Fusion*, vol. 48, pp. 67-83, 2019.

[13] N. Kumar and S. Sonowal, "Email Spam Detection Using Machine Learning Algorithms," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 108-113.

[14] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. javad Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 international conference on computer, communications, and control technology (I4CT)*, 2014, pp. 63-65.

[15] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," *Encyclopedia of machine learning*, vol. 15, pp. 713-714, 2010.

[16] D. DeBarr and H. Wechsler, "Spam detection using random boost," *Pattern Recognition Letters*, vol. 33, pp. 1237-1244, 2012.

[17] L. Firte, C. Lemnaru, and R. Potolea, "Spam detection filter using KNN algorithm and resampling," in *Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing*, 2010, pp. 27-33.

[18] A. Wijaya and A. Bisri, "Hybrid decision tree and logistic regression classifier for email spam detection," in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2016, pp. 1-4.

[19] Farisa, "An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks," *Information Fusion 48 (2019) 67–83*, vol. 48 2019.

[20] L. Chwif, M. R. P. Barretto, and R. J. Paul, "On simulation model complexity," in *2000 winter simulation conference proceedings (Cat. No. 00CH37165)*, 2000, pp. 449-455.

[21] A. Majeed, "Improving time complexity and accuracy of the machine learning algorithms through selection of highly weighted top k features from complex datasets," *Annals of Data Science*, vol. 6, pp. 599-621, 2019.

[22] T. Xia, "A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems," *IEEE Access*, vol. 8, pp. 82653-82661, 2020.

[23] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, pp. 293-314, 2014.

[24] V. M. Patro and M. R. Patra, "Augmenting weighted average with confusion matrix to enhance classification accuracy," *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, pp. 77-91, 2014.

[25] Y. Han, M. Yang, H. Qi, X. He, and S. Li, "The Improved Logistic Regression Models for Spam Filtering," in *2009 International Conference on Asian Language Processing*, 2009, pp. 314-317.

[26] K. A. Kaufman and R. S. Michalski, "Learning from inconsistent and noisy data: the AQ18 approach," in *International Symposium on Methodologies for Intelligent Systems*, 1999, pp. 411-419.

[27] B. Al-Shboul, H. Hakh, H. Faris, I. Aljarah, and H. Alsawalqah, "Voting-based Classification for E-mail Spam Detection," *Journal of ICT Research & Applications*, vol. 10, 2016.

[28] M. A. Hassan and N. Mtetwa, "Feature extraction and classification of spam emails," in *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2018, pp. 93-98.

[29] A. I. Kadhim, "Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 2019, pp. 124-128.

[30] M. Diale, T. Celik, and C. Van Der Walt, "Unsupervised feature learning for spam email filtering," *Computers & Electrical Engineering*, vol. 74, pp. 89-104, 2019.

[31] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, pp. 16-28, 2014.

[32] D. S. Moore, "A chi-square statistic with random cell boundaries," *The Annals of Mathematical Statistics*, pp. 147-156, 1971.

[33] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, pp. 352-359, 2002.

[34] R. E. Wright, "Logistic regression," 1995.

[35] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, pp. 3-14, 2002.

[36] A. DeMaris, "A tutorial in logistic regression," *Journal of Marriage and the Family*, pp. 956-968, 1995.

[37] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*: Wiley New York, 2000.

[38] Y. Tao, J. Jiang, Y. Liu, Z. Xu, and S. Qin, "Understanding performance concerns in the API documentation of data science libraries," in *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2020, pp. 895-906.

[39] P. Das. (2021). *Logistics Regression in python*. Available: https://www.codespeedy.com/logistics-regression-in-python/.

[40] M. J. Kearns, The computational complexity of machine learning: MIT press, 1990.

[41] A. Abdiansah and R. Wardoyo, "Time complexity analysis of support vector machines (SVM) in LibSVM," *International journal computer and application*, vol. 128, pp. 28-34, 2015.

[42] A. Tyagi, "Content based spam classification-a deep learning approach," Graduate Studies, 2016.

[43] M. Esmaeili, A. Arjomandzadeh, R. Shams, and M. Zahedi, "An anti-spam system using naive Bayes method and feature selection methods," *International Journal of Computer Applications*, vol. 165, pp. 1-5, 2017.

[44] S. Douzi, F. A. AlShahwan, M. Lemoudden, and B. El Ouahidi, "Hybrid Email Spam Detection Model Using Artificial Intelligence," *International Journal of Machine Learning and Computing*, vol. 10, 2020.

Securing Images through Cipher Design for Cryptographic Applications

Punya Prabha V[1]
Electronics and Communication Engineering, Ramaiah Institute of Technology Bangalore, INDIA

Dr. M D Nandeesh[2]
Electronics and Instrumentation Engineering, Ramaiah Institute of Technology, Bangalore, INDIA

Tejaswini S[3]
Medical Electronics and Engineering Ramaiah Institute of Technology Bangalore, INDIA

***Abstract*—The emphasis of this work is image encoding based on permutation as well as changes that utilize Latin cube as well as Latin square image cipher meant for both color and gray images. Generally, multimedia data are transmitted in the network as well websites, numerous methods have been established for securing the information without any negotiation. Security of information in all the areas is required to ensure that the information sustains privacy, is presentable for recovery as well as governance purposes. These data can be secured by taking CIA (confidentiality, Integrity, and Availability) to realize information like confidentiality about the data that can be reserved as undisclosed from an illegal user of source, the integrity of the data is maintained unaffected for unauthorized font, availability of resources for official personal to retrieve data for access the information. Authentication of a person is by identification, conserving information and its validation of data. Implementing this authentication will store the data in the required format that is either exchanged or transmitted for the internet application. By breaching the misuse of data is can be protected for confidential as well as sensitive data. To achieve security and maintain confidentiality cryptographic methods are implemented.**

***Keywords*—Decryption; encryption; Latin square generator; sequence generator**

I. INTRODUCTION

In the current scenario security of data is the main issue to maintain confidentiality as well as privacy. The main concern is managing the data while communicating as well as storing data from stealing as well as exploiting the data. For this purpose, an appropriate method has been employed for securing the data while transmitting as well as storing the data by utilizing the technique of cryptography. The information is utilized in the form of a plaintext image. The encryption technique converts this plaintext image into an unidentifiable form that cannot be easily decoded. By employing the decryption process, it is possible to get back original information by utilizing the Cipher text method. The main solicitations of this image encryption are many, some of them are military application, medical images, communication via the internet and many more. In this electronic era, cryptography is applied to provide various types of security to different electronic devices [1]. New encryption for images that utilize permutation-substitution network and chaotic systems has provided a better result in terms of both qualitative and quantitative studies that is superior to other techniques [2]. Generally, an intelligent technology that can protect the information from unauthorized person to hack or decode, that the data is immune to attack. This will implicate various techniques such as decipherment in symmetric and asymmetric keys [3] [20]. This process uses the decryption and encryption of data that is equivalent to electronic locking of knowledge [4]. This technique uses a private or secret key that can be shared only with receive for decoding as well as encoding the information. A novel technique based on memristive chaotic for random behavior is used for encryption has resulted in effective results in the decryption and protecting the data [5] [16]. The behavior of chaotic systems is mainly categorized using sensitivity based on dependence on primary situations. By using this chaos for self-synchronization has ignited a huge work in the field of cryptography [6] [15]. For certain initial conditions, some nonlinear functions can generate chaotic or random numbers. By utilizing a one-time pad system and replacing it with proper chaotic function cryptographic results can be improved [7]. Encryption of probabilistic image for LSB plane using random noise provides better diffusion as well as confusion properties with error tolerance [8] [10]. The security analysis is tested for different types of attacks [9] [18]. Decipherment of plain text is called secret information of plain text p and cipher text C is encrypted [11]. Latin square and CNN, for the input image are implemented bitwise and encrypted and decomposed to obtain encrypted image [12] [17]. Input image based on Latin Square creates 256 bits' key, which is operated by ex-or with Latin square to obtain encrypted images [13]. Latin rectangle scrambling using chaotic map reduces the number of iteration for achieving better security [14]. Latin square with S box and chaotic system results in better security [19]. For decryption, decoding the cipher image accompanying lookup table with index numbers results in effective and secured results [21]. Images of multimedia and its application are focused on efficient encryption and decryption to obtain better results [22].

This article is organized as follows, Section 2 provides the background of encryption methods. Section 3 explains the methodology of implementation of Latin square that generates and Substitution for cipher image as well as Whitening of premature images. Whereas Section 4 provides and evaluates the results and Section 5 concludes this paper.

II. Background

With the evolution of information technologies and pandemic situations, present days, people have started to use the internet and images in their daily routine. So encryption of images is gaining huge momentum for the present situation. For security reasons, image encryption converts the original images into unrecognizable images, Medical images, military services as well as its affair application, e-commerce, telecommunication, and many applications that gains the prominence of image encryption. In the present scenario, usage of text data has been reduced because of the advantages of the security in the digital images, since a different enormous amount of information like wide content, maximum data, as well as monotony and an exceptionally authoritative correlativity, can be obtained by the neighborhood pixel.

A. The Objective of the Work

Cryptography deals with mainly three core capacities like integrity, authenticity and confidentiality, which defends the digital data from hackers or unauthorized users from theft. By using encoding, the confidentiality of information is hidden is known as data privacy. The private key is used to extract the image at the transmission of the data. For the decryption of information, the same private key is used that has been used for encryption. Authenticity assurances that a person or a process can be identified receiver by their identification method. If the same data has been transmitted to the receiver ensure that data is integrated, without loss of any information sent by the transmitter.

III. Methodology

Misuse and theft of information are the main threat while storing and transmitting images, to a major amount of care has to be taken to maintain confidentiality and privacy. A technique based on cryptography can be utilized to store information as well as communication of data in plain text data. The process regarding hiding a given message as a fragment of a specific kind with accurateness to protect the data is encoding. The technique of converting cipher content back into plain text without any loss is deciphering.

As a part of security and confidentiality, the encrypting technique has gained momentum in the present situation. Information safety is a significant subject as part of communication purposes and for image storage, encryption is a distinct behavior that can secure the information. For this purpose, encrypting has enormous applications Image encoding includes a huge number of applications Encrypting text messages may lead to unrecognizable information or trash data known as cipher content. This technique is utilized to guarantee that data hidden from others of concern is not intended to, furthermore, individuals can recognize the encoded data. The technique of recovering the cipher text to its reliable plaintext is known as decryption.

A. Latin Square

A Latin Square is a collection N*N is jumble-sale of demonstration of N different elements, that all the elements representation considering exactly just the once in each distinct row along with each distinct column. Fig. 1 represents an example of a Latin square.

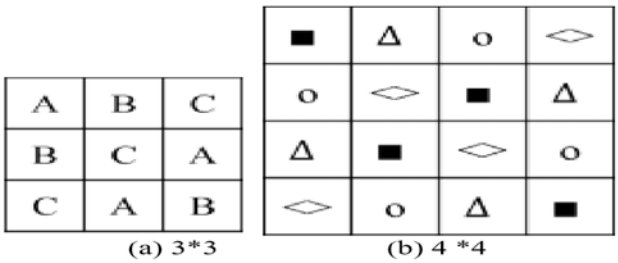


Fig. 1. Latin Square with Dissimilar Sets and Orders.

B. Latin Square Generator

A Latin square is created by using two structures of identical length. If P1 and P2 are two sequences of length equal to N arrangements [11].

LS=LSG (P1, P2)

P1 and P2 are input sequence

Ensure: - LS is a N order Latin square

S1 = Sort Map function (P1)

S2 = Sort Map function (P2)

For i=0 to N-1

LS (i,:)=Row shift(S1, S2(i))

End

P1 and P2 are two input sequences that are aimed to generate a pseudo-random number generator. Sort Map function (S1, S2) is classifying function with index in the middle of an S1 and S2 series, as a sequence of acclivitous order can be obtained by transformed edition S1*, and operated on Row Shift is through (S1, v) which is a periodic transfer of the S1 arrangement with v elements to the left [9][10][11].

C. Substitution-Permutation Network

Input information in cryptography is represented as plaintext and the resultant output information is called cipher text. A substitution-permutation network (SPN) is a sequence of cipher that includes various iterations, for each iteration includes a replacement, a variation and an additional key. It is a sequence of mathematical equations utilized for block cipher or different algorithms. Block ciphers algorithms are designed employing replacing and variation in various forms of SPN. Throughout the N-iteration variation and replacement of the given network, a plaintext will be converted into a new bit series and symbolized as P, which is primary and source content that has to be encoded. SNP that has been encoded to obtain the cipher text in the bit sequence C, by using this technique. Fig. 2 represents the substitution permutation network. A commonly used procedure to persistently synthesis with a given key for the plaintext is known as Latin square

whitening. Replacement of one byte to other bytes in the input sequence is known as Latin Square Substitution. Whereas shuffling bit location in the input sequence is done in Latin Square Permutation. Similarly, the reverse will be implemented in the decryption process of a substitution and permutation network cipher.

D. Latin Square Image Cipher

In the encoding and decoding technique for cipher, the treatment block is by splitting (32*32) bytes of segmented grayscale, for which its definite pixel value with the intensity of 256 levels (1byte). In this algorithm P is a plaintext image is to be defined by (32*32) byte, whereas cipher image C is denoted by (32*32) byte, Latin Square is denoted by L (32 bytes), and the encoding key is denoted by K (32 bytes) encoding key. This suggested technique iterates for 4 and 8 rounds of substitution as well as permutation network structure respectively. Fig. 3 represents Latin square image Cipher with 8-round SPN.

E. Latin Square Whitening

By utilizing a typical SPN method for ciphers block, the mixing of the content of plain text P along with key using rotation is implemented by using logical operations such as XOR procedures. In this proposed algorithm plain text P is composed of 1byte per pixel for the encoding process. In a normal predictable technique, logical whitening using XOR turns out to be intolerable, because of unsuccessful of information in images. Mixing the content of plaintext with that of a sequence is the to generate the Latin square is intended in whitening, for this purpose cipher is reorganized through Finite field or Galois field for plain text image that can be represented as

f = [r + l] 2^8

The reverse of whitening technique to obtain Latin square can be obtained by using this equation,

r = [f + l]2^8

Where f is the whitening result in the byte.

l is the Latin square in the byte.

r is a plain text image in the byte.

2^8 is used to calculate Galois or finite field.

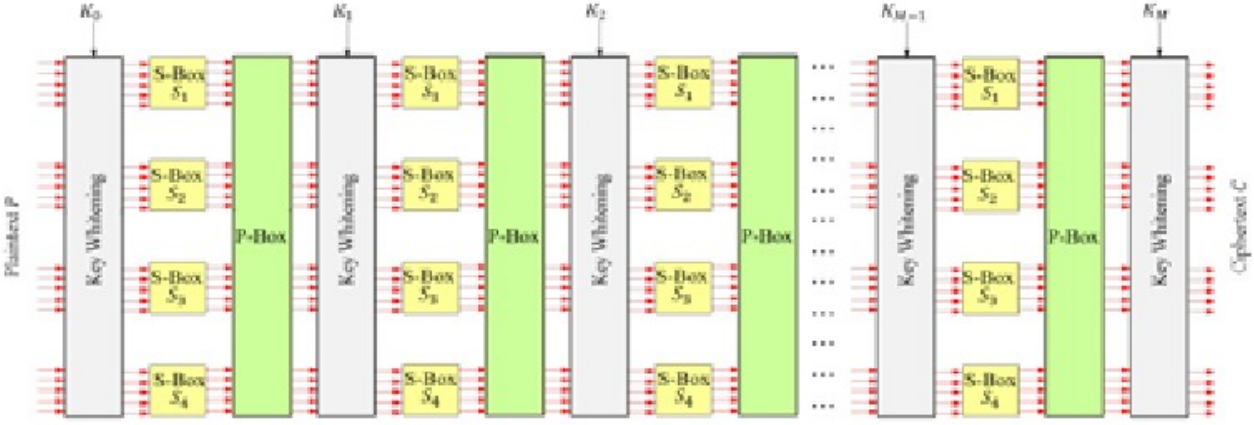


Fig. 2. Substitution-Permutation Network.

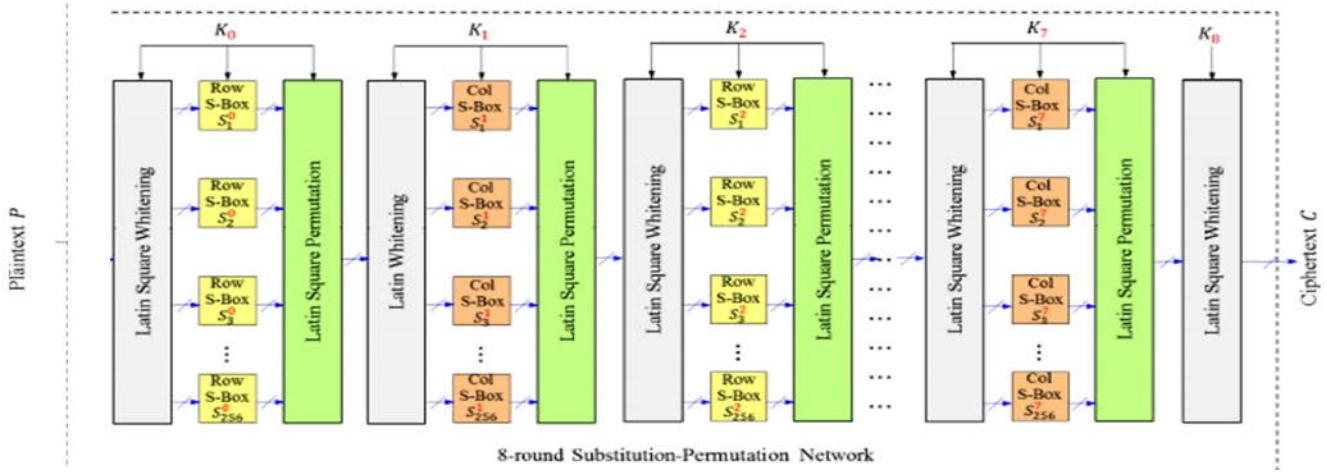


Fig. 3. Block Diagram of Latin Square Image Cipher with 8-round SPN.

F. Latin Square Substitution

S-box is frequently applied to accomplish byte substitution in the calculation of cryptographic images. All S-Box is recognized as bijection, furthermore a definite function that can be mapped as one-to-one. In the encoding process, image pixels are associated with the image as 8-bit or byte. That is a grayscale image of 8-bit that contains intensity with a grayscale of 32-byte along with every intensity is essential of 8-bit. Normally there is a presence of forwarding row and column mapping, this forward column along with row mapping will map to Latin square. Byte replacement is achieved in image cipher by the corresponding mapping which is termed as Latin Square S-box of replacement.

$$\text{LSRS:}\begin{Bmatrix} C = Ecr(L,P) \\ P = Dcr(L,C) \end{Bmatrix}_{ROW}$$

LSRS function in form of a pixel in cipher image is represented in row forwarding plotting function merely by Latin Square L via a function restriction to have cipher text, similarly, IRM is the reverse function to obtain plaintext.

ECR_s^{row}:

$$C(r,c)=\begin{Bmatrix} FRM(L,C(r-1,c),P(r,c)), if\ r \neq 0 \\ FRM(L,0,P(r,c)), if\ r = 0 \end{Bmatrix}$$

DCR_s^{row}:

$$P(r,c)=\begin{Bmatrix} IRM(L,C(r-1,c),C(r,c)), if\ r \neq 0 \\ IRM(L,0,C(r,c)), if\ r = 0 \end{Bmatrix}$$

$$\text{LSCS:}\begin{Bmatrix} C = Ecr(L,P) \\ P = Dcr(L,C) \end{Bmatrix}_{COL}$$

Similarly, for the column pixel function of LSCS, cipher image is represented utilizing FCM function over Latin Square L to obtain ciphertext. Similarly, ICM is a reverse function to obtain the plain test and is represented as.

ECR_s^{col}:

$$C(r,c)=\begin{Bmatrix} FCM(L,P(r,c),C(r,c-1)), if\ c \neq 0 \\ FCM(L,P(r,c),0), if\ c = 0 \end{Bmatrix}$$

DCR_s^{col}:

$$P(r,c)=\begin{Bmatrix} ICM(L,C(r,c),C(r,c-1)), if\ c \neq 0 \\ ICM(L,C(r,c),0), if\ c = 0 \end{Bmatrix}$$

After implementation of Latin square substitution, input plaintext image P will have converted in an unidentifiable image, by applying LSRS and LSCR, the resultant output will be in histogram pattern.

G. Latin Square Permutation

The permutation is mainly required to use P-box for its implementation. This P-Box is referred to as bijection, which specifies one-to-one mapping. This P box of Latin square will plot corresponding integer numbers to each row or column that permutes a series of data in integer. Latin square permutation can achieve both forward and reverse plotting for both row and column mapping function for the r^{th} row and c^{th} column respectively and represented as:

$$\begin{Bmatrix} y = FRM(L,r,x) = L(r,x) \\ x = IRM(L,r,y) = argmax(f(r,z,y)) \end{Bmatrix}$$

$$\begin{Bmatrix} y = FCM(L,x,c) = L(x,c) \\ x = ICM(L,y,C) = argmax(f(r,c,y)) \end{Bmatrix}$$

where x and y, represent input and output relating to plotting function, respectively.

IV. Result and Discussion

Implementation of this work for Encryption as well as Decryption was using MATLAB R2014a, results were simulated and are explained as provided. The results have been divided into four parts as shown below.

A. Encryption and Decryption for 8 Round Grayscale Image

Encryption of grayscale Latin Square Cipher image block for 8 rounds, requires a 32-byte K key, 32*32 bytes, 1byte image block (grayscale) for P to determine C that is of 32*32 byte of grayscale with 1 byte of the block, Fig. 4 represents a Latin square grayscale cipher image block for 8 rounds of encryption.

Similarly, for 8 rounds of decryption value P is determined for the grayscale 1 byte of 32*32 bytes by utilizing the key (K) and Cipher images (C) that is of 32-byte key and grayscale of 1 byte of 32*32 byte. Decryption for 8 round grayscale cipher image using Latin Square is as shown in Fig. 5. For this purpose ship image available in MatLab is utilized for encryption as well as decryption.

B. Encryption and Decryption for 4 Round Grayscale Image

Encryption of grayscale Latin Square Cipher image block for 4 rounds, requires a 32-byte K key, 32*32 bytes, 1byte image block (grayscale) for P to determine C that is of 32*32 byte of grayscale with 1 byte of a block, Fig. 6 represents a Latin square grayscale cipher image block for 4 rounds of encryption.

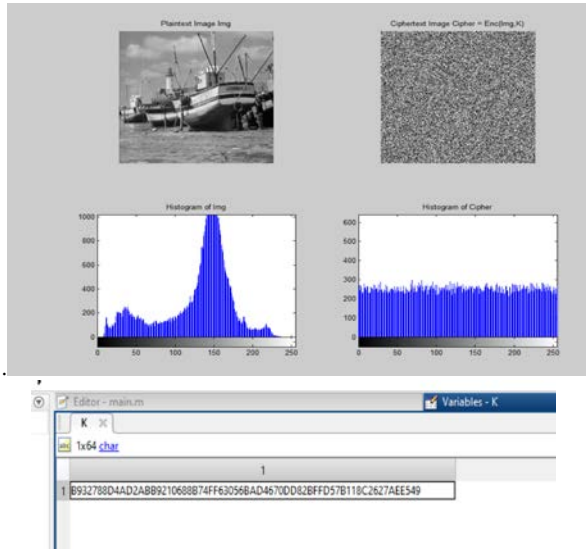


Fig. 4. Encryption of Latin Square Grayscale Cipher Block Image after 8 Rounds.

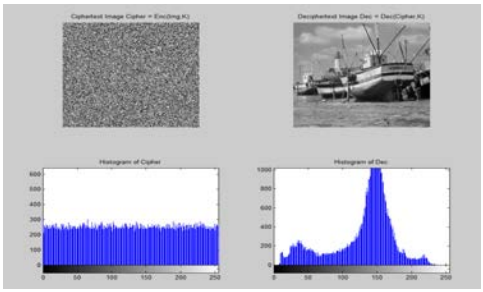


Fig. 5. The Decryption of Latin Square Grayscale Cipher Image after 8 Rounds.

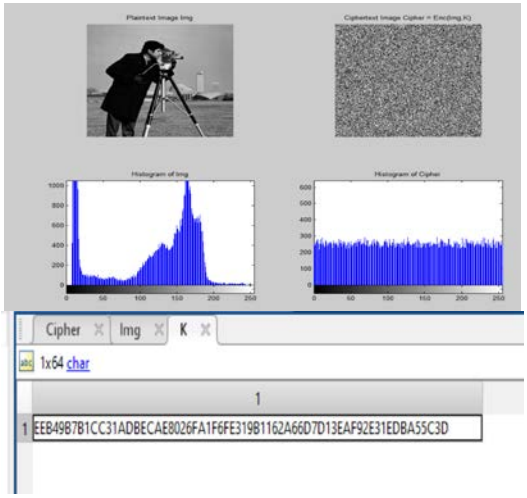


Fig. 6. Latin Square Grayscale Image Cipher block- Encryption for 4 Rounds.

Similarly, for 8 rounds of decryption value P is determined for the grayscale 1 byte of 32*32 bytes by utilizing the key (K) and Cipher images (C) that is of 32-byte key and grayscale of 1 byte of 32*32 byte. Decryption for 8 round grayscale cipher image using Latin Square is as shown in Fig. 7. For this purpose, in internet-based cameramen image utilized that is available in Mat lab for both encryptions as well as decryption.

C. Latin Cube Color Image Cipher block- Encryption and Decryption

Encryption and decryption of grayscale Latin cube color Cipher image require a 32-bit K key, 32*32 bytes, image block (color) for P to determine C and for decryption, it is to find back P image with Key K, Fig. 8 represents a Latin cube color cipher image block for both encryption as well as decryption.

D. Key Change Analysis

1) Encryption and decryption process for Latin Square Grayscale Image Cipher block with the same key (8 rounds): Encryption of grayscale Latin Square Cipher image block for 8 rounds, requires a K key, plain text image P to determine C. For decryption P is determined with the same key and C block, Fig. 9 represents Latin Square Grayscale Image Cipher block- Encryption and Decryption for 8 Rounds. After applying the same key for encryption and decryption with the same K (32 bytes) the resultant image is obtained and is as shown. Fig. 10 represents Latin Square Grayscale Image Cipher block- Encryption and Decryption for 8 Rounds (when a key is same). So the distortion at receive is minimum when data is retrieved.

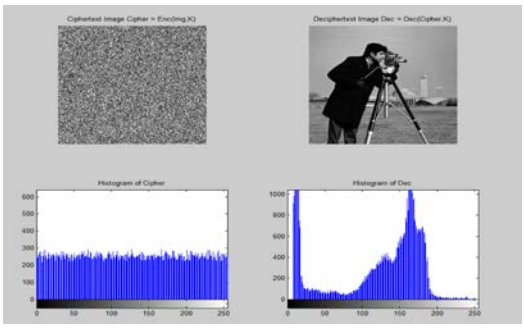


Fig. 7. Latin Square Grayscale Image Cipher for block- Decryption for 4 Rounds.

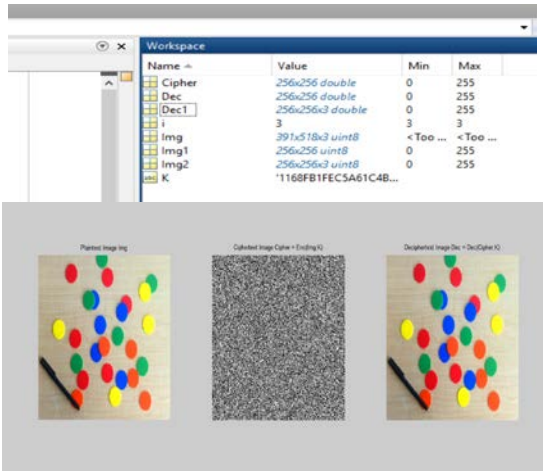


Fig. 8. Latin Cube Color Image Cipher block- Encryption and Decryption.

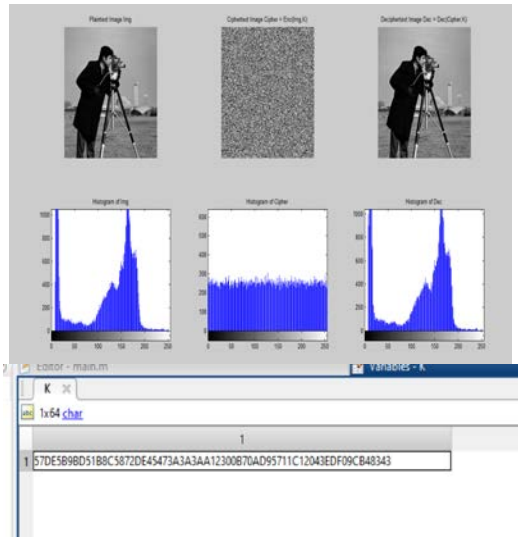


Fig. 9. Latin Square Grayscale Image Cipher Block- Encryption and Decryption for 8 Rounds.

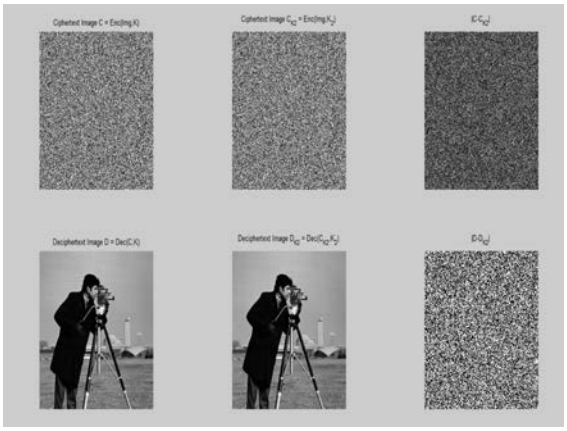


Fig. 10. Latin Square Grayscale Image Cipher Block- Encryption and Decryption for 8 Rounds (When a Key is same).

2) Encryption and decryption process for Latin Square Grayscale Image Cipher block with a different key (8 rounds): For encryption K1 key is required that is of 32 bytes, it is used to determine the value of C that is of size 32*32 bytes by using the cipher image P of size 32* 32 bytes. Similarly, in the decryption stage the value of plain text P employing cipher image C and different key K2, the result is as shown in the diagram. Fig. 11 shows different keys used for encryption and decryption for Latin Square Grayscale Cipher image for 8 rounds. The results can be compared with the images decryption with the same as well as different keys. The images obtained from different key has shown that is similar salt and pepper noise, which shows that the image that has been decrypted is unrecognizable when compared to that of decryption using the same key. The image decryption obtained from the same key has shown a better result when compared to that of a different key.

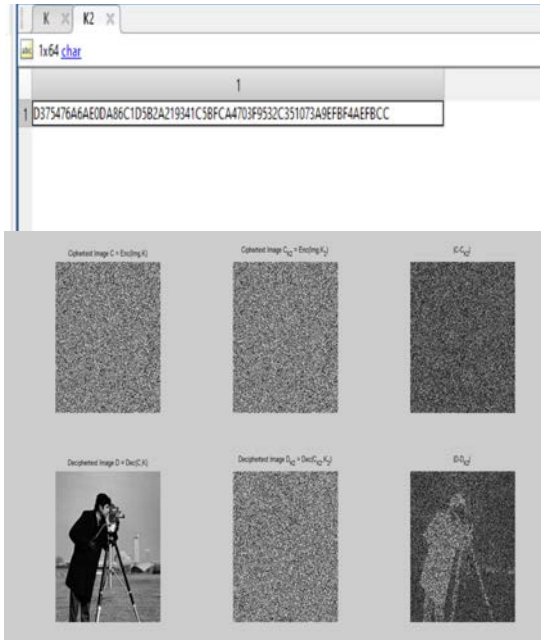


Fig. 11. Different Keys were used for Encryption and Decryption for Latin Square Grayscale Cipher Image for 8 Rounds.

The result can be tabulated as shown in the comparison table. Table I represents the difference between the plain images and deciphers images after decryption image, where decipher images have shown a better result when compared to the normal decipher images. Table II shows the comparison between the present work as well as the base paper for both correlation and entropy are better when compared to the results [1].

TABLE I. COMPARISON RESULTS

Features	Plain image	Decipher image
Contrast	0.587162	0
Correlation	0.922726	Nan
Energy	0.18037	1
Homogeneity	0.8952	1
Mean	118.73	118.73
Standard deviation	62.44	62.35
Entropy	7.009	0

TABLE II. COMPARISON WITH BASE PAPER

Feature	Base paper	Proposed result
Correlation	0.9227	0.9315
Entropy	7.997	7.0097

V. CONCLUSION

In this work, a unique image is constructed that is cipher based on Latin square as well as Latin cube. By using a grayscale image of Latin square plaintext P of 32*32 bytes is encrypted using key K to obtain cipher images of size 32*32, with histogram where Latin square is n*n 2 dimensional attribute. Implementation of Latin cube for the color image is implemented using P and K key, the color image has shown better performance for the Latin cube that has a 3-dimensional attribute. The results obtained are compared with the implementation of the paper by Xu [1] has shown a better performance both in correlation as well as entropy. By using the different keys of the image is decrypted the results have shown a salt and pepper image that cannot be recognized as the original image. Simulation results have shown that the developed algorithm will preserve better security as well as advanced efficiency appropriate for real-world application. LSIC analysis has shown that Latin cube and Square in the cryptographic application have a better safety portion in terms of security as well as privacy from hackers with the change in key. Experimental security analysis with comparison between the plain image as well as decipher image is shown in the table. The distinction between plain image and decode image for the same key was 0.0037%. This result shows same data has been transmitted to the receiver to ensure that data is integrated, without loss of any information sent by the transmitter

REFERENCES

[1] M. Xu and Z. Tian, "An Image Cipher Based on Latin Cubes," Third International Conference on Information and Computer Technologies, pp. 160-168, DOI: 10.1109/ICICT50521.2020.00033, 2020.

[2] Belazi, Akram, Ahmed A. Abd El-Latif, Safya Belghith. "A novel image encryption scheme based on substitution-permutation network and chaos", Signal Processing, pp 155-170, 2016.

[3] Chen, Guanrong, Yaobin Mao, Charles K. Chui. "Asymmetric image encryption scheme based on 3D chaotic cat maps", Chaos, Solitons & Fractals, pp 749-761, 2004.

[4] Zhang, Yong, Yingjun Tang, "A plaintext-related image encryption algorithm based on chaos", Multimedia Tools and Applications, pp 6647-6669, 2016.

[5] Wang Bo, F Zou, Jun Cheng, "A memristor-based chaotic system and its application in image encryption", Optick, pp 538-544, February 2018.

[6] L. Kocarev, "Chaos-based cryptography: a brief overview," IEEE Circuits and Systems Magazine, vol. 1, no. 3, pp. 6-21, 2001, DOI: 10.1109/7384.963463.

[7] Robert Matthews, "On the derivation of a chaotic encryption algorithm", Crypto logia, pp29-42, Vol 13, 1989.

[8] Yue Wu, Yicong Zhou, Joseph P. Noonan and Sos Agaian. "Design of image cipher using Latin squares", Information Sciences, p 317-339, 2014.

[9] Machkour, M., A. Saaidi, M. L. Benmaati. "A novel image encryption algorithm based on the 2-dimensional logistic map and the Latin square image cipher", 3D Research, pp1-18, 2014.

[10] Nema and Tanvi. "A Symmetric-Key Latin Square Image Cipher with Probabilistic Encryption for Grayscale and Color Images", IJCSIT, International Journal of Computer Science and Information Technologies, Vol. 8 (3), pp 380-388, 2017.

[11] Chai, Xiuli, et al. "Medical image encryption algorithm based on Latin square and memristive chaotic system." Multimedia Tools and Applications pp 35419-35453, 2019.

[12] M. Lin, F. Long and L. Guo, "Grayscale image encryption based on Latin square and cellular neural network," Chinese Control and Decision Conference, pp. 2787-2793, DOI: 10.1109/CCDC.2016.7531456, 2016.

[13] S. K. N. Kumar, H. S. S. Kumar and H. T. Panduranga, "Hardware-software co-simulation of dual image encryption using Latin square image," Fourth International Conference on Computing, Communications and Networking Technologies, pp. 1-5, DOI: 10.1109/ICCCNT.2013.6726681, 2013.

[14] S. Chapaneri and R. Chapaneri, "Chaos-based image encryption using Latin rectangle scrambling," Annual IEEE India Conference, p. 1-6, DOI: 10.1109/INDICON.2014.7030358, 2014.

[15] Jude Hemanth and S. Smys Computational Vision and Bio-Inspired Computing", Springer Science and Business Media, Lecture Notes in computational vision and biomechanics, vol 28, 2018

[16] Xu, M. and Tian, Z, "An Image Cipher Based on Latin Cubes." Third International Conference on Information and Computer Technologies, pp. 160-168. March 2020.

[17] Hua, Z., Li, J., Chen, Y, Yi, S, "Design and application of an S-box using complete Latin square". Nonlinear Dynamics, 104(1), pp.807-825, 2021.

[18] Khalid, S. S. Jamal, T. Shah, D. Shah and M. M. Hazzazi, "A Novel Scheme of Image Encryption Based on Elliptic Curves Isomorphism and Substitution Boxes," in IEEE Access, vol. 9, pp. 77798-77810, 2021, DOI: 10.1109/ACCESS.2021.3083151.

[19] Zhongyun Hua, Jiaxin Li, Yongyong Chen and Shuang Yi. "Design and application of an S-box using complete Latin square", Nonlinear Dynamics Vol 104, pp 807-825, 2021.

[20] H. Mou, X. Li, G. Li, D. Lu and R. Zhang, "A Self-Adaptive and Dynamic Image Encryption Based on Latin Square and High-Dimensional Chaotic System,", Third International Conference on Image, Vision and Computing, pp. 684-690, DOI: 10.1109/ICIVC.2018.8492876, 2018.

[21] Subjajith Adhikari and Sunil Karfoma, "A Novel image encryption method for e-governance application using elliptic curve pseudo-random number and chaotic random number sequence", Multimedia Tools and Applications, Springer series 2021.

[22] Ashish S. Dongare, Dr. A. S. Alvi, N. M. Tarbani, "An Efficient Technique for Image Encryption and Decryption for Secured Multimedia Application", International Research Journal of Engineering and Technology, Volume: 04, Issue: 04, 2017.

A Review of Feature Selection Algorithms in Sentiment Analysis for Drug Reviews

Siti Rohaidah Ahmad, Nurhafizah Moziyana Mohd Yusop, Afifah Mohd Asri, Mohd Fahmi Muhamad Amran
Department of Science Computer, Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia, Sungai Besi
Kuala Lumpur, Malaysia

***Abstract*—Social media data contain various sources of big data that include data on drugs, diagnosis, treatments, diseases, and indications. Sentiment analysis (SA) is a technology that analyses text-based data using machine learning techniques and Natural Language Processing to interpret and classify emotions in the subjective language. Data sources in the medical domain may exist in the form of clinical documents, nurse's letter, drug reviews, MedBlogs, and Slashdot interviews. It is important to analyse and evaluate these types of data sources to identify positive or negative values that could ensure the well-being of the users or patients being treated. Sentiment analysis technology can be used in the medical domain to help identify either positive or negative issues. This approach helps to improve the quality of health services offered to consumers. This paper will be reviewing feature selection algorithms, sentiment classifications, and standard measurements that are used to measure the performance of these techniques in previous studies. The combination of feature extraction techniques based on Natural Language Processing with Machine Learning techniques as a feature selection technique can reduce the size of features, while selecting relevant features can improve the performance of sentiment classifications. This study will also describe the use of metaheuristic algorithms as a feature selection algorithm in sentiment analysis that can help achieve higher accuracy for optimal subset selection tasks. This review paper has also identified previous studies that applied metaheuristics algorithm as a feature selection algorithm in the medical domain, especially studies that used drug review data**.

Keywords*—Sentiment analyis; drug reviews; feature selection; metaheuristic

I. Introduction

Sentiment analysis (SA) or opinion mining is a field that analyses opinions, comments, expressions, and views on different entities, such as products, services, organisations, and individuals. It is subjective in sentiment analysis to analyse each comment to identify the types of sentiment polarity, as either positive, negative, or neutral. According to [1], sentiment analysis is widely implemented in the domains of products, restaurants, movies, etc. However, this technique is not widely used in the medical domain, which could probably be due to privacy and ethical issues [1].

The current widespread use of social media has allowed users the freedom to speak their mind by giving their opinions or views in various aspects, such as medical quality, services they received, the effectiveness and side effects of drugs, and medical costs. Users would use social media platforms as a place for them to express their dissatisfaction or satisfaction with the goods or services provided. According to [2], medical documents are classified into six types, namely, nurse's letter, radiological report, discharge summary, drug reviews, Medblogs, and slashdot interviews.

This study has specifically focused on drug reviews, namely, users' comments on drugs in terms of effectiveness, side effects, symptoms, facilities, and the value of the drugs. The main problem in sentiment classification is that features extracted from user comments often contain data that are redundant, irrelevant, or even misleading [3], [4].

According to [5], there are three levels of SA, namely, feature, sentence, and document. The focus of this study has been on feature, which is to identify features embedded in customers' comments, as either positive, negative, or neutral. This study has also identified previous studies that were conducted on the medical domain, which were specifically related to drug reviews. This study has focused on identifying feature selection techniques and techniques for classifying sentiments in customers' comments on drug use. Several combinations of keywords were used ("feature selection + sentiment analysis + drug review or drug) during the search process in standard databases, such as Elsevier, ACM, Google Scholar, Science Direct, Elsevier, SpringerLink, Scopus, Taylor & Francis; and IEEE Xplore.

II. Background Research

According to [2], drug reviews refer to comments on drugs, which are related to their effectiveness, side effects, convenience, and value. User comments can help other users find the best businesses, destinations, or services by sharing opinions and ratings on these drugs.

It is important to analyse these drug reviews and identify users' views or opinions on these drugs, whether they are good or vice versa. The results of this analysis could contribute insights related to the health field to the stakeholders of this field [1]. Apart from that, the results of this sentiment analysis could also help the community understand the effects of drugs on human health. This paper will describe drug reviews, sentiment analysis and feature selection in detail.

A. Drug Review

According to [6], drug reviews consist of posts in social media, where patients express their experiences and opinions about treatments or medicines. According to [7], the

Pharmaceutical Care Network Europe (PCNE) defined medication review as a structured evaluation of a patient's medicines, with the aim of optimising medicine usage and improving health outcomes, in terms of drug-related problems and recommended interventions. As reported by [2], a drug review can be defined as a user's personal perceptions on several drug-related categories, including effectiveness, side effects, convenience, and value. According to [8], a drug review is a patient-written review on various drugs based on their experiences and preference. This kind of review provides a lot of information that can lead to accurate decisions about public health and drug safety.

B. Sentiment Analysis

Sentiment analysis (SA), also referred to as opinion mining, is the area of research that analyses the perceptions, thoughts, opinions, evaluations, behaviour, and emotions of people on anything, for example, products, services, organisations, people, concerns, activities, topics, and their attributes [5]. SA is the process of evaluating a word or a sentence based on their sentiment. Any opinion or emotion expressed in the form of a text would contain a negative, positive, or neutral element [9]. As stated by [2], SA could be used to gather information on the effectiveness of a treatment or medication from social media and health records. According to [6], drug manufacturers could also benefit from SA, particularly in pharmacovigilance, as particular adverse effects of a drug can be found more easily from public repositories or social media posts. A drug review may contain a high proportion of sentiment terms formed from personal impressions and feelings [2]. Therefore, SA can be used to collect useful information that can assist in making accurate decisions on public health and drug safety.

C. Feature Selection

Features are topics or keywords found in users' comments. A feature can be a topic that is being discussed or things that users made comments on. An example of a user's comment sentence: 'This camera is very good': the feature in this sentence is 'camera' and the word sentiment is 'good'. Various definitions of feature selection have been provided by previous studies [10 –14]. Based on studies by [4, 9, 15], it is important to produce an optimal feature subset by reducing feature size to increase classification accuracy. In conclusion, feature selection is a process of selecting and identifying features that are not redundant and relevant to reduce the size of feature dimension and improve the accuracy of sentiment classification. Therefore, this study aimed to identify feature selection techniques used in previous studies to select features in drug review datasets.

III. A Review Feature Selection Algorithms used in Sentiment analaysis for Drug Reviews

The world of social media is full of people who would make various comments, either positive or negative. Social media is full of information regarding users' preferences and experiences when using products or services. This type of information should be utilised by identifying valuable insights in such comments using artificial intelligence technologies, such as sentiment analysis. Big-sized and high-dimensional data are a major problem that can decrease the accuracy of classification performance and complicate the process of obtaining an optimal feature subset. Feature selection in SA is an important step to produce an optimal feature subset [14], without having to change the original meaning of the feature. This study will identify feature selection methods, feature extraction, sentiment classification, data sets, and evaluation standards that are being used to measure the performance of the methods used.

NLP concepts, such as part of speech tagging, n-gram, content words, and function words have also been used to extract features from tweet data [16]. The Penguin Search Optimization (PeSOA) algorithm [16] was also used as a feature selection technique to select optimal features based on the keywords of drugs and cancer in tweet data. The K-Nearest Neighbour (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM) methods were used through MATLAB simulation software to classify the tweet data. The performance metrics used were processing time, accuracy, precision, recall, and F-Measure to measure the performance value of each proposed method. Based on the combined feature selection techniques, which consisted of PeSOA and three classification methods, namely, PeSOA-KNN, PeSOA-NB, and PESOA-SVM, it was found that the combination of PeSOA-SVM was able to produce high accuracy, precision, recall, and F-Measure values compared to the other combinations. Similarly, PeSOA-SVM required less processing time to complete the classification process compared to other combinations. This increased performance was due to the ability of the combined SVM and PeSOA to classify larger data sizes from the search process from multiple dimensions. Their study had only focused on comments that contain the keyword drug, regardless of the type and effect of the drug. According to [1], the Bag of Words (BoW) technique or the term frequency-inverse document frequency (TF-IDF) technique were used to extract important words in a document. Once the keywords have been extracted from the document, the next process was to select an optimal feature subset using the Fuzzy-Rough Quick Reduct (FRQR) technique. By using BoW to determine the value of a feature, the feature selection process was able to significantly reduce the generated feature space. FRQR was able to select 43 optimal features from the 903 original features using the forward search strategy. Meanwhile, 56 optimal features were selected using the backward search. These two resultant feature subsets were tested using four classification methods, namely, the Ripper, Naive Bayes, Random Forest, and Decision Tree. The performance of these methods was measured based on training accuracy, performance of running independent hold-out test, and the time required to build the model. The experimental results showed that the FRQR technique was able to increase sentiment accuracy, as well as reduce the complexity of feature space, and the classification of run-time overheads.

According to [17], machine learning methods are insufficient to address the complex grammatical relationships between words in clauses. Their study applied a linguistic approach to overcome weaknesses in machine learning approaches. The advantage of using a linguistic approach is that this method can determine sophisticated rules for dealing

with various grammatical relationships between words in sentences or clauses. New rules based on linguistics can also be added to the system at certain levels. A comparison between SVM (a machine learning method) and linguistic approach was conducted using a dataset obtained from DrugLib.com. Experimental results showed that the linguistic approach was more effective compared to the SVM method. However, several problems have been identified based on the error analysis. This situation showed that the proposed linguistic approach required improvement.

Satisfaction with drug use was analysed based on drug reviews from www.askapatient.com [18]. Several experimental analyses were conducted on the performance of Probabilistic Neural Network (PNN) and Radial Base Function Neural Networks (RFN) using two different datasets, namely, cymbalta and depo-provera. The results showed that the Neural Network approach surpassed the SVM method in terms of precision, recall, and F-Score values. The RFN method showed a higher performance value compared to the PNN method.

In their research [19] used the Probabilistic Aspect Mining Model (PAMM), which is a method to identify the relationship between features and class labels. Due to the unique features of PAMM, it focuses on finding features related to one class only rather than simultaneously finding features for all classes in each implementation. Apart from finding features, it also has properties that can be distinguished by the class. This means PAMM can be used to differentiate between classes, which help reduce the likelihood of features being formed from mixing different class concepts. Thus, the identified features would be easier to construe. Researchers have argued that this method can avoid features that have been identified as having contents mixed from different classes. Better and more specific features can be identified by focusing on the tasks in one class. This approach is also different from the intuitive approach, whereby reviews were grouped first according to their class label and followed by features for each group. The proposed model used all reviews when finding features that were specific to the target class. This approach helped to distinguish reviews from different classes.

Various sentiment categories for consumer review on drugs have been identified for the introduction to Adverse Drug Reactions (ADRs) [20]. The Weakly Supervised Model (WSM) was introduced using data labelled as weak to pre-train model parameters. Then, WSM was combined with the Convolutional Neural Network (CNN) and the Bidirectional Long Short-term Memory (Bi-LSTM) to produce another model, known as the WSM-CNN-LSTM to implement the sentiment classification process. The experiments showed that the proposed model was able to improve ADR recognition based on accuracy, precision, and F-Score values compared to other models.

According to [8], two deep fusion models have been proposed based on the three-way decision theory to analyse drug reviews. The first fusion model was known as the 3-way fusion of one deep model with traditional models (3W1DT). In 3W1DT, each classic algorithm is combined with a deep learning method separately. For example, Naïve Bayes (NB) was combined with Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), and Three-Way Convolutional Recurrent Neural Network (3CRNN), and known as GRU-NB, CNN-NB, and 3CRNN-NB, respectively. The second combination models were known as the 3-way fusion of three deep models with traditional models (3W3DT) to improve the performance of the deep learning methods. This second model combined three learning algorithms, namely, GRU, CNN, and 3CRNN with traditional algorithms, which were NB, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbour (KNN). These combinations were known as 3W3DT-NB, 3W3DT-DT, 3W3DT-RF, and 3W3DT-KNN. Data sets from Drugs.com were used to test these two models. The 3W1DT and 3W3DT methods showed better results compared to the stand-alone traditional and deep learning methods. Meanwhile, a comparison between 3W1DT and 3W3DT showed that 3W3DT was able to produce higher accuracy and F1-Score values compared to 3W1DT. The study [8] had also intended to apply a metaheuristic feature selection technique and evolutionary algorithm to improve the performance of the proposed fusion models in the future.

In their study, [21] implemented two feature extraction methods, namely, Word Embedding and Position Encoding in Vector Representation to extract features from drug review datasets. The obtained features were tested using four sentiment classification methods, namely, NB, SVM, RF, and Radial Basis Function Network (RBFN). They compared the sentiment classification of the original SentiWordNet (SWN) lexicon with the medical domain-based SentiWordNet lexicon (Med-SWN). Experimental results showed the effectiveness of the proposed method in the feature selection process. Meanwhile, an assessment on the performance of sentiment classification has proven that the features extracted from Med-SWN outweighed those from SWN.

Based on the summary in Table I, the use of metaheuristic techniques as part of feature selection techniques is still in its infancy. Therefore, further research must be conducted to prove that metaheuristic techniques are able to produce optimal feature subsets and help improve the performance of sentiment classification accuracy. The use of metaheuristic feature selection techniques was suggested by [8] to improve the performance of sentiment classification accuracy. However, this situation depends on the data training sets. Tests based on domains could also play an important role in each study.

TABLE I. A SUMMARY OF FEATURE SELECTION ALGORITHMS AND FEATURE EXTRACTION METHODS FOR DRUG REVIEWS

Author	Feature Extraction	Feature Selection	Classification	Measurement
[1]	Bags of Words (BoW) or term frequency-inverse document frequency (TF-IDF)	Fuzzy-Rough Quick Reduct (FRQR).	Ripper, Naive Bayes, random forest and decision tree	Performance based on training accuracy, performance of running independent hold-out test, and the time required to develop the model.
[8]	Not mentioned in paper.	Not mentioned in paper.	CNN-NB, GRU-NB dan 3CRNN-NB. 3W3DT-NB, 3W3DT-DT, 3W3DT-RF dan 3W3DT-KNN	Precision, recall, and F-Score
[16]	Part of speech tagging, n-gram, content words, function words	Penguin Search Optimization (PeSOA)	K-Nearest Neighbour (KNN), Naïve Bayes (NB), and support vector machine (SVM)	Accuracy, precision, recall, and F-Measure
[17]	Not mentioned in paper.	Not mentioned in paper.	Rule-based Linguistic	Precision, recall, accuracy, and F1-score
[18]	Not mentioned in paper.	Not mentioned in paper.	Probabilistic neural network (PNN), and radial basis function neural networks (RFN)	Precision, recall, and F1-Score
[19]	The specific type of feature extraction was not mentioned.	Not mentioned in paper.	Probabilistic aspect mining model (PAMM)	Mean Pointwise Mutual Information (PMI) and accuracy
[20]	Not mentioned in paper.	Not mentioned in paper.	Weakly supervised model (WSM), convolutional neural network (CNN), and bidirectional long short-term memory (Bi-LSTM)	Accuracy, precision, and F1-Score
[21]	Word embedding and position encoding in vector representation	Not mentioned in paper.	Naive Bayes, SVM, RF, and RBFN	Precision, recall, and F-Score

IV. A SURVEY OF FEATURE SELECTION USING METAHEURISTIC ALGORITHMS IN SENTIMENT ANALYSIS

This section will briefly present feature selection techniques that use metaheuristic algorithms in sentiment analysis. Metaheuristic techniques can solve various problems with satisfactory solutions in a reasonable time. According to [22], metaheuristic techniques have been used for over 20 years in numerous applications. Most applications that use this technique demonstrated efficiency and effectiveness for solving large and complex problems.

These techniques are a high-level strategy and iteration generation process, which can guide the process of exploring the search space using different techniques. Metaheuristic techniques may include ant colony optimization (ACO), artificial immune system (AIS), bee colony, genetic algorithm (GA), particle swarm optimization (PSO), and genetic programming [22], [23]. According to the study by [23], metaheuristic characteristics are as follows:

1) A strategy that provides guidance in the search process.

2) Able to effectively explore the search space and find the optimal solution; and.

3) A simple local search procedure for complex learning processes.

Metaheuristic techniques have been used as feature selection techniques by [4], [24], [25], and [26]. Table II lists several studies in other domains that similarly used metaheuristic algorithms, such as particle swarm optimization, ant colony optimization, hybrid cuckoo search, and artificial bee colony that have been proven to show good results based on precision, recall, F-measure or accuracy values.

TABLE II. A SUMMARY OF FEATURE SELECTION USING METAHEURISTIC ALGORITHMS IN SENTIMENT ANALYSIS

Author	Feature Selection	Domain	Result
[4]	Ant Colony Optimization	Customer Review	Precision = 81.5%; Recall = 84.2%; and F-score = 82.7%
[26]	Multi-Swarm Particle Swarm Optimization	Online Course Reviews	Micro-F-measure = 88%
[27]	Particle Swarm Optimization	Movie review	Accuracy level from 71.87% to 77%.
[28]	Multi Objective Artificial Bee Colony	Movie Review	Accuracy 93.8%
[29]	Particle Swarm Optimization	Laptop and Restaurant	F-measure values = 81.91% and 72.42% for aspect term extraction classification. Accuracies = 78.48% (restaurant) and 71.25% (laptop domain).
[30]	Fitness Proportionate Selection Binary Particle Swarm Optimization	Hotel Reviews And Laptop Reviews	Accuracy = 93.38%
[31]	Particle Swarm Optimization	Cosmetic Products Review	Accuracy from 82.00% to 97.00%
[32]	Hybrid Cuckoo Search	Twitter Dataset	Not mentioned the value of accuracy.
[33]	Ant Colony Optimization	Twitter Dataset	Accuracy = 90.4%

Next, the search for research papers on feature selection using metaheuristic algorithms that use drug review data was based on the following combinations of keywords:

1) ("Feature selection + sentiment analysis + metaheuristic + drug review);
2) ("Feature selection + sentiment analysis + optimization + drug review); and
3) ("Feature selection + sentiment analysis + swarm intelligence + drug review).

Searches in benchmark databases, such as ACM, IEEE Xplore, Elsevier, SpringerLink, Scopus, Google Scholar Taylor & Francis; and Science Direct showed no results. However, when the keyword combination has no 'sentiment analysis' and 'drug review', several papers were found containing the following keywords:

1) ("Feature selection + swarm intelligence + medical);

and

2) ("Feature selection + swarm intelligence + health).

Brief descriptions on each paper are given in the following section. In their work, [34] studied feature selection techniques for the classification of medical datasets based on Particle Swarm Optimisation (PSO). Their research was focused on multivariate filter and wrapper approaches, combined with PSO using medical dataset. PSO was used as a filter and CFS was used as a fitness function. They also proposed using the wrapper approaches with PSO on five classifiers, namely, decision tree, Naïve Bayes, Bayesian, radial basis function, and k-nearest neighbour to increase classification accuracy. This method had been tested for feature selection classification on three medical datasets, which were the breast cancer dataset, the Statlog (Heart) dataset, and the dermatology datasets. A comparison was performed between the proposed approaches with the feature selection algorithm based on genetic approach. The results showed that the PSO_CFS filter was able to improve classification accuracy, while the proposed wrapper approaches with PSO showed the best classification accuracy. However, two studies had identified that GA_CFS is more reliable than the proposed method, which would be when KNN and RBF classifiers were applied to the to Statlog (Heart) datasets.

Confidence-based and cost effective feature selection (CCFS) methods were proposed using binary PSO on UCI lung cancer dataset [35]. The results showed that the proposed algorithm demonstrated effectiveness in terms of accuracy and cost of feature selection. Additionally, [36] applied the Binary Quantum-Behaved Particle Swarm Optimisation (BQBPSO) algorithm as a feature selection technique for selecting optimum feature subsets for a microarray dataset that contains five types of data set, namely, Leukaemia, Prostate, Colon, Lung, and Lymphoma. The BQBPSO showed more significant results in terms of accuracy and optimal feature subset compared to two comparison algorithms, namely, Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA).

An ontology-based two-stage approach to medical text classification, with feature selection using particle swarm optimization research was conducted by [37]. They developed a two-stage methodology to analyse domain principles and identify which concepts are discriminatory to a classification problem. This research used a set of clinical text, known as the 2010 Informatics for Integrating Biology and Bedside (i2b2) dataset. This dataset must go through an ontology-based feature extraction during the first stage. The MetaMap tool was then used to send the document to Unified Medical Language System (UMLS) to extract all features with meaningful phrases. A simple idea was applied in the concept section set and finally, a tf-idf measure was used to transform the feature into a vector. In the second stage, PSO was used to further remove redundant and unwanted features. To test the accuracy of the suggested method, five classifiers were used, namely, Naive Bayes (NB), Linear Support Vector Machine (LSVM), K-Nearest Neighbour (KNN), Decision Tree (DT), and Logistic Regression (LR). The results showed that the two-stage approach was able to extract meaningful features, reduce the number of features, and improve classification accuracy.

Based on the summary of previous studies in Table III, metaheuristic algorithms have been used as a feature selection algorithm in the medical or healthcare domain. However, these experiments had only included other disease datasets, such as breast cancer, lung cancer, and leukaemia. Experiments using drug review data were not found in this literature review.

Therefore, further research should be conducted using drug review datasets as research data to implement the use of metaheuristic algorithms. Additionally, studies should be conducted to identify metaheuristic algorithms that would be appropriate for drug review datasets.

TABLE III. A SUMMARY OF FEATURE SELECTION USING METAHEURISTIC ALGORITHMS IN MEDICAL OR HEALTHCARE DOMAIN

Author	Feature Selection	Domain	Dataset
[34]	Particle Swarm Optimization	Medical	Breast Cancer, Heart, Dermatology
[35]	Binary Particle Swarm Optimization	Healthcare	Lung Cancer
[36]	Binary Quantum-Behaved Particle Swarm Optimization	Medical	Leukaemia, Prostate, Colon, Lung, and Lymphoma
[37]	Particle Swarm Optimization	Medical	Medical Notes
[38]	Confidence-based Cost-effective + Binary Particle Swarm Optimization	Healthcare	UCI datasets

V. CONCLUSION AND FUTURE WORK

The literature review in this research paper was conducted in three parts. The first part was to identify which feature selection algorithms were used for drug review data in

previous studies. Table I shows the results of the first search, whereby natural language processing, machine learning, and metaheuristic algorithms were used. However, several studies did not state the type of feature selection techniques used in their study. Table I also shows that the use of metaheuristic algorithms as a feature selection technique is still lacking in the domain of drug reviews. Next, this study identified the use of metaheuristic algorithms as a feature selection technique in sentiment analysis in general. The results are summarised into Table II, which shows several studies in different domains using metaheuristic algorithms as a feature selection technique. Table II also shows excellent experimental results based on the measured values. Next, this study searched for previous research papers in a list of standard databases that applied metaheuristic algorithms as a feature selection technique in the drug review domain. However, no matches were found. Then, the keyword combinations were changed, which were "feature selection + swarm intelligence + medical" or "feature selection + swarm intelligence + healthcare". Several research papers have been found using these keywords, as listed in Table III. Table II and Table III show that metaheuristic algorithms can be used as a feature selection technique in the domains of movies, customer reviews, tourism, medical, and healthcare, with excellent experimentation results. SA plays an important role in the decision-making process. Health-based organisations or services would have to make decisions on the use of drugs, side effects or services provided based on user comments. Numerous approaches can be used in SA. Metaheuristic-based feature selection techniques can assist in the selection of optimal features, with higher accuracy. The literature review has shown that research to implement metaheuristic algorithms as feature selection in the medical domain have great potential which would benefit from further studies on drug review data. Researchers also need to identify the advantages and disadvantages of metaheuristic algorithms that would be used as feature selection algorithms in further studies. More studies are needed to identify previous studies that applied metaheuristic techniques. More experiments should be conducted to identify metaheuristic techniques that are appropriate for future drug review data.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Pertahanan Nasional Malaysia, and the Skim Geran Penyelidikan Jangka Pendek Fasa 1/2021 for supporting this research project through grant no. UPNM/2021/GPJP/ICT/1.

REFERENCES

[1] T. Chen, P. Su, C. Shang, R. Hill, H. Zhang, and Q. Shen, "Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection," IEEE Int. Conf. Fuzzy Syst., vol. 2019-June, pp. 1–6, 2019, doi: 10.1109/FUZZ-IEEE.2019.8858916.

[2] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," Artif. Intell. Med., vol. 64, no. 1, pp. 17–27, 2015.

[3] H. Arafat, R. M.Elawady, S. Barakat, and N. M.Elrashidy, "Different Feature Selection for Sentiment Classification," Int. J. Inf. Sci. Intell. Syst., vol. 1, no. 3, pp. 137–150, 2014.

[4] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "Ant colony optimization for text feature selection in sentiment analysis," Intell. Data Anal., vol. 23, no. 1, pp. 133–158, 2019.

[5] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[6] D. Cavalcanti and R. Prudêncio, "Aspect-based opinion mining in drug reviews," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10423 LNAI, no. August, pp. 815–827, 2017.

[7] N. Griese-Mammen et al., "PCNE definition of medication review: reaching agreement," Int. J. Clin. Pharm., vol. 40, no. 5, 2018, doi: 10.1007/s11096-018-0696-7.

[8] M. E. Basiri, M. Abdar, M. A. Cifci, S. Nemati, and U. R. Acharya, "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques," Knowledge-Based Syst., vol. 198, 2020, doi: 10.1016/j.knosys.2020.105949.

[9] A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web forums," ACM Trans. Inf. Syst., vol. 26, no. 3, pp. 1–34, 2008, doi: 10.1145/1361684.1361685.

[10] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in Machine Learning: Proceedings Of The Eleventh International Conference, 1994, pp. 121–129.

[11] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 2, 1997, doi: 10.1109/34.574797.

[12] P. Koncz and J. Paralic, "An Approach to Feature Selection for Sentiment Analysis," in International Conference on Intelligent Engineering System (INES 2011), 2011, pp. 357–362, doi: 10.1109/INES.2011.5954773.

[13] B. Agarwal and N. Mittal, "Sentiment Classification using Rough Set based Hybrid Feature Selection," in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2013), 2013, no. June, pp. 115–119.

[14] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection techniques in sentiment analysis," Intell. Data Anal., vol. 23, no. 1, pp. 159–189, 2019.

[15] M. Abulaish, Jahiruddin, M. N. Doja, and T. Ahmad, Feature and Opinion Mining for Customer Review Summarization, vol. 5909. SPRINGER-VERLAG BERLIN, 2009.

[16] T. Anuprathibha and C. S. Kanimozhiselvi, "Penguin search optimization based feature selection for automated opinion mining," Int. J. Recent Technol. Eng., 2019, doi: 10.35940/ijrte.B2629.098319.

[17] J. C. Na, W. Y. M. Kyaing, C. S. G. Khoo, S. Foo, Y. K. Chang, and Y. L. Theng, "Sentiment classification of drug reviews using a rule-based linguistic approach," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, doi: 10.1007/978-3-642-34752-8_25.

[18] V. Gopalakrishnan and C. Ramaswamy, "Patient opinion mining to analyze drugs satisfaction using supervised learning," J. Appl. Res. Technol., vol. 15, no. 4, 2017, doi: 10.1016/j.jart.2017.02.005.

[19] S. G. Pooja Gawande, "Extraction of Aspects from Drug Reviews Using Probabilistic Aspect Mining Model," Int. J. Sci. Res., vol. 4, no. 11, 2015, doi: 10.21275/v4i11.nov151059.

[20] Z. Min, "Drugs Reviews Sentiment Analysis using Weakly Supervised Model," in Proceedings of 2019 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2019, 2019, doi: 10.1109/ICAICA.2019.8873466.

[21] S. Liu and I. Lee, "Extracting features with medical sentiment lexicon and position encoding for drug reviews," Heal. Inf. Sci. Syst., vol. 7, no. 1, 2019, doi: 10.1007/s13755-019-0072-6.

[22] E. G. Talbi, Metaheuristics: From Design to Implementation. John Wiley and Sons, 2009.

[23] C. Blum and A. Roli, "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison," ACM Comput. Surv., vol. 35, no. 3, pp. 268–308, 2003, doi: 10.1145/937503.937505.

[24] J. Zhu, H. Wang, and J. T. Mao, "Sentiment Classification using Genetic Algorithm and Conditional Random Field," in Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, 2010, pp. 193–196.

[25] P. Kalaivani and K. L. Shunmuganathan, “Feature Reduction Based on Genetic Algorithm and Hybrid Model for Opinion Mining,” Sci. Program., p. 15, 2015, doi: 10.1155/2015/961454.

[26] Z. Liu, S. Liu, L. Liu, J. Sun, X. Peng, and T. Wang, “Sentiment recognition of online course reviews using multi-swarm optimization-based selected features,” Neurocomputing, vol. 185, pp. 11–20, Apr. 2016.

[27] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization,” in Procedia Engineering, 2013, doi: 10.1016/j.proeng.2013.02.059.

[28] T. Sumathi, S. Karthik, and M. Marikkannan, “Artificial bee colony optimization for feature selection with furia in opinion mining,” J. Pure Appl. Microbiol., 2015.

[29] D. K. Gupta, K. S. Reddy, Shweta, and A. Ekbal, “PSO-asent: Feature selection using particle swarm optimization for aspect based sentiment analysis,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, doi: 10.1007/978-3-319-19581-0_20.

[30] L. Shang, Z. Zhou, and X. Liu, “Particle swarm optimization-based feature selection in sentiment classification,” Soft Comput., vol. 20, no. 10, pp. 3821–3834, 2016.

[31] D. A. Kristiyanti and M. Wahyudi, “Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review,” in 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017, 2017, doi: 10.1109/CITSM.2017.8089278.

[32] A. Chandra Pandey, D. Singh Rajpoot, and M. Saraswat, “Twitter sentiment analysis using hybrid cuckoo search method,” Inf. Process. Manag., 2017, doi: 10.1016/j.ipm.2017.02.004.

[33] L. Goel and A. Prakash, “Sentiment Analysis of Online Communities Using Swarm Intelligence Algorithms,” in Proceedings - 2016 8th International Conference on Computational Intelligence and Communication Networks, CICN 2016, 2017, doi: 10.1109/CICN.2016.71.

[34] H. M.Harb and A. S. Desuky, “Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization,” Int. J. Comput. Appl., vol. 104, no. 5, 2014, doi: 10.5120/18197-9118.

[35] Y. Chen, Y. Wang, L. Cao, and Q. Jin, “An Effective Feature Selection Scheme for Healthcare Data Classification Using Binary Particle Swarm Optimization,” in Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018, 2018, doi: 10.1109/ITME.2018.00160.

[36] M. Xi, J. Sun, L. Liu, F. Fan, and X. Wu, “Cancer Feature Selection and Classification Using a Binary Quantum-Behaved Particle Swarm Optimization and Support Vector Machine,” Comput. Math. Methods Med., vol. 2016, 2016, doi: 10.1155/2016/3572705.

[37] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li, “An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation,” in 2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings, 2019, doi: 10.1109/CEC.2019.8790259.

[38] Y. Chen, Y. Wang, L. Cao, and Q. Jin, “CCFS: A Confidence-based Cost-effective feature selection scheme for healthcare data classification,” IEEE/ACM Trans. Comput. Biol. Bioinforma., 2019, doi: 10.1109/tcbb.2019.2903804.

Detection of Covid-19 through Cough and Breathing Sounds using CNN

Evangeline D[1], Sumukh M Lohit[2], Tarun R[3], Ujwal K C[4], Sai Viswa Sumanth D[5]
Assistant Professor, Department of ISE, M S Ramaiah Institute of Technology, Bangalore, India[1]
Student, Department of ISE, M S Ramaiah Institute of Technology, Bangalore, India[2, 3, 4, 5]

***Abstract*—Covid-19 is declared a global pandemic by WHO due to its high infectivity rate. Medical attention is required to test and diagnose those with Covid-19 like symptoms. They are required to take an RT-PCR test which takes about 10-15 hours to obtain the result, and in some cases, it goes up to 3 days when the demand is too high. Majority of victims go unnoticed because they are not willing to get tested. The commonly used RT-PCR technique requires human contact to obtain the swab samples to be tested. Also, there is a shortage of testing kits in some areas and there is a need for self-diagnostic testing. This solution is a preliminary analysis. The basic idea is to use sound data, in this case, cough sounds, breathing sounds and speech sounds to isolate its characteristics and deduce if it belongs to a person who is infected or not, based on the trained model analysis. An Ensemble of Convolution neural networks have been used to classify the samples based on cough, breathing and speech samples, the model also considers symptoms exhibited by the person such as fever, cold, muscle pain etc. These Audio samples have been pre-processed and converted into Mel spectrograms and MFCC (Mel Cepstral Coefficients) are obtained that are fed as input to the model. The model gave an accuracy of 88.75% with a recall of 71.42 and Area Under Curve of 80.62%.**

Keywords*—Coronavirus; cough sounds; mel frequency cepstral coefficients; convolutional neural network; reverse transcription–polymerase chain reaction (RT-PCR)

I. Introduction

Covid19 is caused by the SARS-CoV2 Virus and was declared a pandemic across the world on the 11th of February 2020. Majority of Covid-19 patients experienced fever, dry cough and fatigue. Other symptoms experienced by the patients include aches and pains, sore throat, diarrhoea, conjunctivitis, headache, loss of taste or smell, a rash on skin, or discoloration of fingers or toes. This pandemic has affected more than 17 crores worldwide and has resulted in the death of 38 lakhs as of June 2021. Testing has become one of the most important requirements for starting the treatment, allocation of Beds, procurement of specific medicines etc. The current methods which are RT-PCR tests are conducted and samples are sent to a lab for disease detection. While Lateral Flow Tests (LFTs) can diagnose Covid-19 immediately, it is not as precise as RT-PCR. Antibody tests also cannot effectively detect Covid-19, but can determine an individuals' immunity to Covid-19. There could be false negatives up to 30% in RT-PCR tests. It means the presence of an infection could be done in far better way than giving a patient the all-clear negative report. There is possibility of false positive results because of detecting dead and deactivated viruses in the body of a pateint recovered from Covid-19.

The paper focuses on using deep learning techniques to detect Covid 19 using methods that don't involve any incision or wound to the patient while still finding ways to detect Covid-19. Cough and breathing based analysis to isolate cough sound snippets and convert this data that can be used for analysis and training a model based on the characteristics obtained. Convolutional neural network is used to analyze the images which are derived after preprocessing and Ensemble of models are used to increase the accuracy. For testing performed the performance will be measured based on the recall and accuracy rates. The applications would include testing methods within web applications, and sophisticated testing without one's aid.

The rest of the paper is organized as follows: Section II highlights the contemporary works carried out for Covid-19 detection in various countries. Section III discusses the materials and methods used in this work. Section IV gives the results obtained on carrying out the work. Section V summarizes the work carried out for Covid-19 detection.

II. Literature Review

In [1], the authors had analyzed the features for cough breathing and voice of the patients. Long term dependencies can be remembered in LSTM. Dataset consists of 60 healthy speakers and 20 Covid-19 patients. In comparison between the three through performance metrics, it is found that patients cough and breathing sounds are effective to diagnose infection as they have high recall. The Limitations of the papers include inefficient preliminary results due to time constraints, collected data is small and lacks control on other patients suffering from other kinds of respiratory illness.

In [2], a CNN model to detect Covid-19 from breath and cough sounds was proposed. Spectrogram extraction to obtain visuals of audio frequencies against time. CNN variant CIdeR is based on ResNets which alleviates the vanishing gradient problem. The output is later given to a sigmoid layer. A score, thus obtained can be used to determine if a person is Covid-19 positive or not. Dataset used is 517 coughing and breathing recordings from 355 people, of which 62 participants had tested positive within 14 days of the recording. The Technique used is CNN-ResNet. The prime limitation of this study is size and demographics of the dataset.

The proposed work in [3] identified coughing is not just one predominant symptom of Covid-19 but a symptom of more than 100 diseases. Machine learning techniques can be applied on global smartphone recordings to detect Covid-19. Dataset used for training is the Coswara dataset and the Sarcos dataset. Techniques used are Logistic regression, SCM, MLP, LSTM, Resnet50.

In [4], the authors suggested that the only screening method for Covid-19 is a thermometer but only 45% of the mild-moderate patients had fever, their study suggested that Covid patients who are especially asymptomatic could be classified with a good accuracy as positive or negative with just the forced-cough recordings. The data was collected from opensigma.mit.edu and it had 5320 cough recordings. The cough recordings were converted using Mel Frequency Cepstral Coefficient (MFCC) and later on, fed into CNN containing one Poisson biomarker layer with biomarkers like Muscular degradation, Vocal cords, Sentiments, Lungs and Respiratory Tract and 3 pre-trained ResNet50's in parallel; 4256 recordings were used for training and the remaining 1064 was used for testing. The model had a recall of 98.5% and a specificity of 94.2%. For asymptomatic persons, it achieved recall of 100% with a specificity of 83.2%.

It was proposed in [5] that the lung volume and oxygenation can be modelled and approximated with a good accuracy. Energy of acoustic signal of respiration in each phase of airflow to/from the lungs from the breathing sounds is considered. Rhonchus, squawk, and stridor are also considered in the inspiration phase. All of the above characteristics are extracted from the lung function augmentation graph which is calculated against time and amplitude.

In [6], the authors have analyzed the computation of Mel spectrogram from cough sounds. Deep transfer learning based multi class / binary classifier along with classical machine learning based multi class classifier were applied to differentiate cough due to Covid-19 from cough due to other respiratory infections.

In [7], the authors have proposed a Radiographic scoring model applied by assessing disease severity using severity score. Score was given based on the extent of lung infection. Statistical analysis was performed by applying Student's T test, Mann Whitney's test, Chi-square test and Fisher's exact test. Some limitations were lack of retrospective analysis and correlation between CXR severity score and patient co morbidities.

The authors in [8] have made a comparative study of twelve deep learning algorithms using a multi-center dataset, including open-source and in-house developed algorithms. The 12 methods which are Benchmarked and compared include Lung segmentation for severe pathologies, 3D Lung segmentation Lobe segmentation, CT Angel software for lung segmentation and binary lesion, CovidENet, 2D Unet, 3D multiclass segmentation, Inf-Net, WASS, UNWM and Majority voting.

In [9], it was proposed that RT-PCR is probably a more accurate and sensitive strategy. GradCAM mappings and 3-D model were applied. Pretest background prevalence determines test success metrics. Testing practices may differ depending upon exposure rates and pandemic phases. The work was limited in truly evaluating the generalizability of this model to an independent population, because positive and negative cases were obtained from separate populations.

In [10], the authors have proposed some Non Clinical techniques such as machine learning, data mining, expert system and other artificial intelligence techniques must play significant roles in diagnosis and containment of the Covid-19 pandemic to detect asymptomatic cases and for accelerating the testing process. Supervised ML models were developed using decision tree, logistic regression, Naive Bayes, SVM and ANN.

Detection of Covid-19 by using both cough recording and the uttering of the vowel sounds was proposed in [11]. Many classifiers like decision trees, support vector machines, K-Nearest-Neighbor, Random Forest (RF) and XGBoost were used. The best performance was shown by the weighted XGBoost classifier. A larger number of X-ray images with a wider distribution could have been used.

The authors in [12] have proposed detection of Covid-19 using cough and breathing sounds. Crowdsourced dataset was used for analysis. The model was trained using two feature sets - one was the handcrafted features like the MFCC (Mel Frequency Cepstral Coefficients) and other statistical features and the other feature set was obtained from the pre-trained models [14] [17][21].

III. Materials and Methods

A. Materials

The proposed solution provides a web application that takes cough, breathing and speech sounds along with the symptoms shown by the person as input and predicts the likelihood of him/her having Covid-19. The training process of the machine learning model begins with pre-processing techniques like loading the audio samples and removing null values. The features of the audio samples like MFCCs and Mel Spectrogram images are further extracted from the audio samples and given to the input generator function. Mel-Frequency Cepstrum Coefficients (MFCCs) are coefficients that provide a representation of the short-term power spectrum of a sound.

A Mel spectrogram is a spectrogram where the frequencies are transformed to Mel scale. This converts the audio samples to image form. The input generator function divides the training data into batches and shuffles the order of the examples so that batches between epochs do not look alike. The data generated from the input generator function is then fed to the machine learning model. Next, an ensemble of three convolutional neural network models and four dense neural network models are used to analyze the pre-processed data and predict the desirable output. Each CNN model consists of 3 2D convolutional layers, average pooling, batch normalization, Relu activation function and a dropout model. Further, the input is flattened using the Flatten class. Glorot Uniform initializer is used to initialize layer weights. Each dense neural network consists of two dense layers using Relu

activation function and Glorot Uniform initializer and a dropout layer. The concatenated ensemble model is then fed through two hidden layers and an output layer to finally generate the output.

The outline of the overall Workflow of the model from start to finish is as follows: The flow starts with the user recording his voice and the symptoms shown by him, the application after receiving the sample runs separate pre-processing steps for voice and categorical features, audio features are converted to Mel spectrograms, MFCC, along with the binary symptoms are fed into the ensemble model which tells if the patient is positive or not.

Fig. 1 gives a brief overview of the Machine Learning model employed in the proposed work. It is an ensemble model of 3 CNN models and 4 dense neural network models. The 3 CNN models take Mel Spectrogram images as input, 3 dense neural network models take MFCCs as inputs and the other dense neural network takes symptoms as input. The ensemble model is then fed through the two hidden layers and an output layer to derive the output. The model consists of an ensemble model of 3 CNN models and 4 dense neural network models. The 3 CNN models each consist of 3 Conv-2D layers, 3 Average Polling 2D layers, 3 Batch Normalization layers, 3 dropout layers and 3 Relu activation functions. The CNN models flatten the input before sending the input through the final dropout layer, dense layer and activation function. The 4 dense neural network models consist of 2 dense layers and 2 dropout layers [16],[18],[26].

Three CNN models were selected because out of the 7 inputs,

The inputs were image inputs i.e Mel spectrogram images of Cough, Breathing and Speech. Since CNN's are very effective in Image classifications because Images have higher dimensions and CNN is very good at reducing the number of parameters without reducing the quality makes it the best option to consider as the model. The 3 CNN models take images as input, 3 dense neural network models take MFCCs as input and the other dense neural network models take symptoms as input. Glorot Uniform is used as the initializer for each of the models. The ensemble model is then passed through 2 dense hidden layers which then feed the input into the dense output layer. The hidden layers use Relu activation function. The output layer finally makes the prediction for the input. The output layer uses the sigmoid activation function.

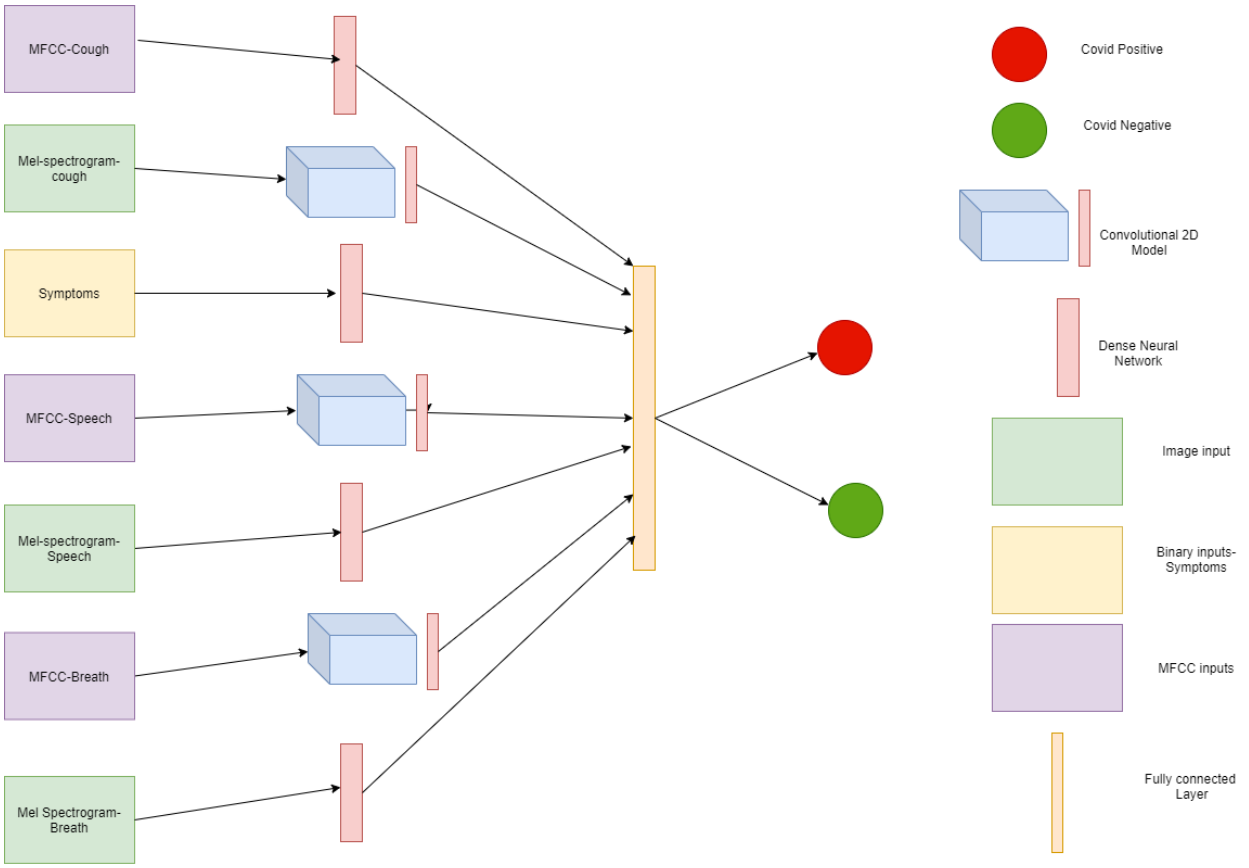


Fig. 1. Brief overview of the Model.

B. Dataset

DiCOVA challenge employs a dataset for respiratory heath diagnosis by speech and audio processing [13]. In this paper, the data used is taken from Coswara data which was collected by Indian Institute of Science (IISc) Bangalore. The dataset includes voice samples including fast and slow breathing sounds, deep and shallow cough sounds, phonation of sustained vowels, and counting numbers at slow and fast pace. Collected metadata includes the participant's age, gender, location country, state/ province of the participant, current health status that could be healthy, exposed, cured or infected and the presence of comorbidities like pre-existing medical conditions. The dataset consists of 1645 samples of individuals. The data was sampled from all the continents except Africa and more than 88% of the samples were from India.74% of samples are male and around 26% are females. 76.8% of samples are healthy individuals; 8.4% of samples are positive and the rest can't be identified. Majority of the samples are those between 20-30 years of age followed by 30-40 years and followed by 40-50 years. Fig. 2 shows Pearson's correlation heatmap for the dataset. It can be inferred from the above figure that the most important symptoms of Covid-19 status are asthma, fever, cough, sore throat, breathing difficulty, cold, fatigue, muscle pain, loss of smell and pneumonia in that order.

C. Data Preprocessing

Sound, represented as an audio signal possesses characteristics like frequency, bandwidth, decibel, etc. Usually, such signals are given as a function of amplitude and time. Audio processing involves extraction of acoustics features relevant to a task. Librosa [19][23][24], a Python for music and audio analysis facilitates various options to construct music information retrieval systems. A Mel spectrogram is a spectrogram where the frequencies are transformed to the Mel scale. The Fourier transform maps continuous time into a frequency spectrum, but an inverse is performed over its log to make it perceptible by humans. General data preprocessing is finally done using Python libraries [22]. In brief, the process of obtaining Mel – Spectrogram is as follows:

1) The samples of air pressure are collected at different instances of time and the same digitally represents an audio signal.

2) The audio signal is transformed from time domain to frequency domain using the Fast Fourier transform.

3) Frequency on y axis is converted to a log scale and amplitude is converted to decibels and the spectrogram is subsequently formed.

4) Frequency is again mapped onto Mel scale and Mel spectrogram is obtained.

D. Mel-Frequency Cepstral Coefficients (MFCCs) and Feature Extraction

MFCCs consist of 10–20 features usually. These features give the general shape of a spectral envelope and models the characteristics of the human voice. To get MFCC, DCT on the Mel-spectrogram is computed. To obtain MFCC, the following steps could be carried out.

1) A windowed excerpt of a signal is taken and Fourier Transform is applied on it.

2) Powers of the spectrum thus obtained are mapped onto Mel scale using triangular overlapping windows.

3) Logarithms of the powers at each Mel frequency are computed.

4) DCT of such MEL log powers is calculated.

5) Amplitudes of the resulting spectrum give MFCCs

Fig. 2. Pearson Correlation Heatmap for the Dataset

A

```
Data Generation
Class CustomDataset
    Initialise batch size, labels,dimensions, image files, channels and scale
    FUNCTION len()
        return number_of_batches
    END FUNCTION
    FUNCTION get_item(index)
        indices<- batch with indices; X,y<- data_generation(indices)
 return X,y
    END FUNCTION
    FUNCTION on_epoch_end()
        update indices after each epoch
    END FUNCTION
    FUNCTION data_generation(list_ids,temp)
        X<- n_samples, dimensions, channels; Y<- n_samples
        for i, id in (list_ids,temp)
            X<- store sample; Y<- store class
        return X,y
    END FUNCTION
```

B

```
FUNCTION feature_extractor()
Load audio sample
    mfcc(sudio_sample); Scale mfccs;        melspectrogram(audio_sample);
    convert mel-spectrograms to dB units
    return mfccscaled,melspectrogram
ENDFUNCTION
for row in tqdm(Data)
    feature_extractor(cough_samples in Data)
for row in tqdm(Data)
    feature_extractor(breathing_samples in Data)
for row in tqdm(Data)
    feature_extractor(speech_samples in Data)
Remove_Nans(dataset)
```

Fig. 3. Psuedocodes (A) Data Generation (B) Feature Extraction.

The following is the six step process to obtain the features.

1) The signals are split into short-time frames.

2) Windowing is applied.

3) An NN-point FFT on each frame is applied and frequency spectrum also called Short-Time Fourier-Transform (STFT) is obtained.

4) Filter Banks are applied. Actually, a set of 20–40 triangular filters is employed.

5) Logarithm of these spectrogram values is applied to get log filter bank energies.

6) DCT (Discrete Cosine Transform) is applied.

E. Convolution Neural Network

In Covolutional 2D filters, Keras Conv2D parameter determines the number of kernels to convolve with the input volume resulting in a 2D activation map. Average Pooling calculates the average value for patches of a feature map and creates a down sampled (pooled) feature map.

```
Model-training
FUNCTION build_model()                                                    C
        //First, third, fifth and seventh model,
        Input()
        Dense(activation='relu',kernel_initializer='GlorotUniform')
        Dropout()
        Dense()
        Dropout()
//Second, Fourth and sixth model
        Input()
        Conv2D()
        AveragePooling2D()
        BatchNormalization()
        Activation('relu')
        Conv2D()
        AveragePooling2D()
        BatchNormalization()
        Activation('relu')
        Conv2D()
        AveragePooling2D()
        BatchNormalization()
        Activation('relu')
        merge<- Concatenation of seven models
        hidden1<-Dense(activation='relu')(merge)
        hidden2<-Dense(activation='relu')(hidden1)
        output<-Dense(activation='sigmoid')(hidden2)
for all splits:
        build_model()
        Compile_model(metric='AUC',optimizer = 'Adam',loss = 'BinaryCrossentropy')
        Split data into train,validation and test sets of ratio 0.7:0.15:0.15
        SevenInputGenerator(train_data)
        SevenInputGenerator(validation_data)
        SevenInputGenerator(test_data)
        model.fit(training_data)
END FUNCTION
predict(model,train_data); predict(model,validationt_data); predict(model,test_data)
Evaluate metrics on train, test and validation data; Visualise the results
```

Fig. 4. Psuedocode (C) Model Training.

In Average Pooling layer the average for each block is computed rather than a max value, which is the case with max pooling. Batch Normalization (BN) is performed as a solution to speed up the training phase of deep neural networks through the introduction of internal normalization of the input values within the neural network layer. Dropouts can prevent overfitting in the model and can be added to randomly switch some percentage of neurons of the network. When the neurons are switched off, the incoming and outgoing connection to those neurons is also switched off. The dense layer is a deeply connected neural network layer with each neuron in the dense layer receiving input from all neurons of the preceding layer. Thus the dense layer results in an 'm' dimensional vector.

```
Class CustomPipeline
    Initialise batch size, labels,dimensions, image files, channels and scale
    FUNCTION len()                                                          D
        return number_of_batches
    END FUNCTION
    FUNCTION get_item(index)
        indices<- batch with indices
        X,y<- data_generation(indices)
        return X,y
    END FUNCTION
    FUNCTION on_epoch_end()
        update indices after each epoch
    END FUNCTION
    FUNCTION data_generation(list_ids,temp)
        X<- n_samples, dimensions, channels
        Y<- n_samples
        for i, id in (list_ids,temp)
            X<- store sample
            Y<- store class
        return X,y
    END FUNCTION
Class SevenInputGenerator
    X1<- CustomPipeline(X1,Y)
    X2<- CustomDataset(X2,Y)
    X3<- CustomPipeline(X3,Y)
    X4<- CustomDataset(X4,Y)
    X5<- CustomPipeline(X5,Y)
    X6<- CustomDataset(X6,Y)
    X7<- CustomPipeline(X7,Y)
FUNCTION get_item(index)
    X1_batch, Y_batch = X1.getitem(index)
    X2_batch, Y_batch = X2.getitem(index)
    X3_batch, Y_batch = X3.getitem(index)
    X4_batch, Y_batch = X4.getitem(index)
    X5_batch, Y_batch = X5.getitem(index)
    X6_batch, Y_batch = X6.getitem(index)
    X7_batch, Y_batch = X7.getitem(index)
    X_batch=[X1_batch,X2_batch,X3_batch,X4_batch,X5_batch,X6_batch,X7_batch]
return X_batch_Y_batch
    END FUNCTION
```

Fig. 5. Psuedocode (D) Processing.

Units are the fundamental parameter that takes positive integers usually. This parameter determines the size of the weight matrix along with the bias vector. The activation parameter helps to apply element-wise activation function in a dense layer. By default, Linear Activation is used but any of the options that Keras [16][18][24] provides can be used for this. In the model Relu has been used as the activation function and it is used for activation to avoid gradient descent,

It also avoids vanishing gradient problem from sigmoid function. The rectified linear activation function or ReLU for short is a piecewise linear function that maps input to output for positive values. The initializer parameter deals with initialization of values in the initialization layer. In the Dense Layer, the weight matrix and bias vector has to be initialized. The initializer used here called glorot_uniform draws samples from a uniform distribution within [-limit, limit] where,limit=sqrt(6/fan_in+fan_out) where, fan_in is the number of input units in the weight tensor and fan_out is the number of output units. The algorithms employed in this work are mentioned in Fig. 3, Fig. 4 and Fig. 5.

IV. Results and Discussion

Through the analysis of the Pearson's correlation matrix we found that a few symptoms are particularly highly correlated and other symptoms didn't have much weightage so only highly influential symptoms were provided as the option. The inclusion of the 7 inputs in itself has the information equal to that of many inputs because the inputs like Mel spectrogram and MFCC have a lot of parameters associated with them , each parametre represent different feature which makes it so that we are considering enough inputs without getting to the scale of overfitting. The 7 inputs include Mel spectrogram and MFCCs of cough, speech and breathing samples respectively along with another input that included symptoms. MFCC inputs further have 39 coefficients each.

The user interface for the web application is developed [15] [20]. It shows the input given to the Covisound-Covid detection website. Symptoms are given as inputs. The symptoms considered are fever, muscle pain and respiratory problems (asthma, breathing difficulty, cold and cough). It also shows the result displayed on the Covisound-Covid detection website. The website displays the result of the Covid-19 test, negative in this case and how likely it is that the person has Covid-19 is 5.82%.

Fig. 6 shows the input given to the Covisound-Covid detection website. Breathing sounds, cough sounds, speech sounds and symptoms are given as inputs. The application shows the result displayed on the Covisound-Covid detection website. 60% of the samples were used for training, 20% for testing and 20% for validation. The total number of samples considered after preprocessing are 1605. Therefore, the number of training and testing samples is 963 and 321 respectively. The website displays the result of the Covid-19 test, negative in this case and the likelihood of the person having Covid-19 is 72.37%. Evaluation metrics considered here include accuracy, recall, AUC Score, ROC Curve, False Positive Rate and False Negative Rate [25]. The five models used for analysis have a few attributes and parameter changes between them; this is performed to see how the prediction varies based on the values set. We set a different training environment for the data to deduce the best possible trade-off over different variations.

Five models are trained by having a binary cross entropy loss function with different metrics for optimising the model. The first model uses Area under the curve as the metric for optimising the model and thereby maximising the same parameter. For imbalance data handling, class weights parameter was provided in the ratio of 1:10 (each negative sample gets ten times the weightage of a positive sample). A high AUC value is obtained here. In our second model the metric that was tuned used is accuracy so that the model focuses on the accuracy of each function and it maximises the same in each iteration. Here the trade-off made is recall which was very low and not favourable for this particular paper as the main aim is to have a high recall to reduce false negatives. In the Third model, the metric tuned was true negatives, the model gave considerable high accuracy and true negatives but it had one of the lowest recall when compared to all the other models. The false negatives rates obtained were also high, so this model was not very favourable. In the fourth model the metric used for optimising the model is recall, this attribute describes the number of positive predictions determined out of all True positive predictions. Recall also shows how well the positive class is covered in the predictions. The fifth model uses Area under the curve as the metric for optimising the model and maximises the parameter. The reason being this parameter shows the true variance between true positives and true negatives which leads to better distinguishability between them. Higher AUC means the model is able to deduce the right class it belongs to. Scoring metrics for different models are tabulated below in Table I. In Fig. 7A, ROC curves of a few models trained on the dataset in different epochs is shown. It also shows the AUC for the respective models. Fig. 7B shows the variation in AUC values vs different epochs for the model. The graph shows a steady improvement in AUC value as the epoch number increases.

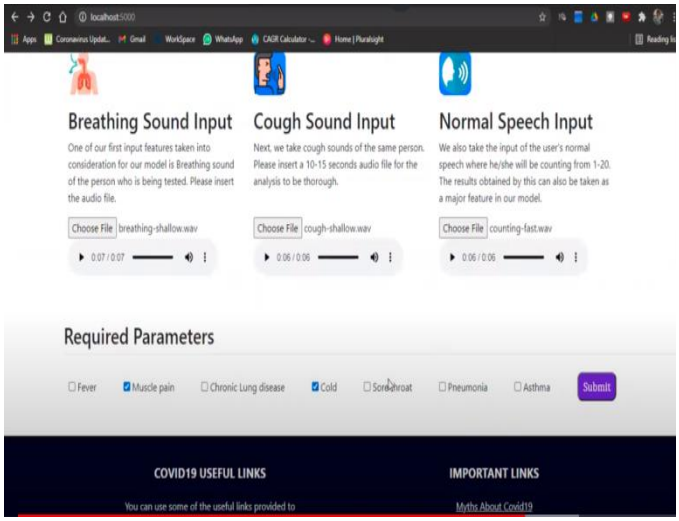


Fig. 6. Covisound- Covid Detection Website Inputs-Positive Sample.

TABLE I. Scoring Metrics for different Models

Model	Accuracy	Recall	AUC-Score	False positive rate	False negative rate
Model-1	88.75%	71.42%	80.62%	10.18%	28.57%
Model-2	95%	44.44%	71.77%	0.9%	55%
Model-3	92.91%	60%	77.95%	4%	40%
Model-4	83.75%	75%	79.86%	15.27%	25%
Model-5	90.42%	63.16%	77.96%	7.23%	36.84%

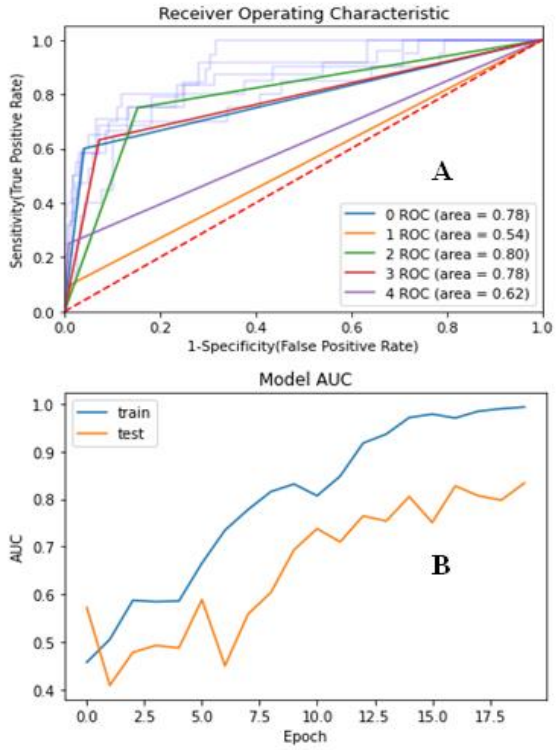


Fig. 7. (A) ROC Curves of a Few Trained Models (B) AUC Value vs epoch Numbers.

The following are the observations made after analysing the results: The best model for the paper was chosen based on AUC and recall. Model 1 gave the best recall and AUC values of 71.42% and 80.6% respectively. It gives an accuracy of 88.75%, a false negative rate of 28.57% and a false positive rate of 10%. A slight trade-off between accuracy and recall was observed while choosing the best model. Some of the models had an accuracy of 96.67% and 97% but had low recall whereas some of the other models had a high recall of >80% and close to 90% but they had low accuracy. So, the best model had to be chosen in such a way that it gave decent values for both accuracy and recall. AUC was chosen as the metric while training the model. Training AUC of 99.8%, validation AUC of 91% and a test AUC of around 81% were observed.

Binary cross entropy was chosen as the loss function while training the model. As the epoch number increased, the AUC value increased and the loss value-binary cross entropy decreased for test, train as well as the validation data. Choosing class weights while training increased the overall performance of the models but the best model was obtained while training without the class weights. The model performs much better than the baseline model in which had an AUC of 70%. The work shows that Covid-19 can be determined using the characteristics of an individual related to speech, breathing and cough sounds. The symptoms and existing conditions can further help in determining Covid-19 status. The model is able to predict the Covid-19 status of an individual reasonably and also the likelihood of the individual having Covid-19.

V. Conclusion

The proposed work showcases the possibility of using an ensemble of Convolutional Neural Networks as a testing method for Covid-19. An encouraging result was obtained in the paper by taking speech, cough and breathing sounds as inputs. Coswara data has been used in this work. MFCCs and Mel spectrogram images were obtained to extract features from the audio samples.

Seven inputs were fed to the model- MFCC cough, MFCC speech, MFCC breath, Mel spectrogram cough, Mel spectrogram breath, Mel spectrogram speech and symptoms. An ensemble of three CNN models and four dense neural network models was used for the purpose of classification. The machine learning model is able to identify Covid-19 patients with a recall of 71.42% and an AUC of 80.62%. The model performs much better than the baseline model in which had an AUC of 70%. An easy-to-use web application with a friendly user interface has also been developed as part of the work. The web application predicts whether the individual is Covid-19 positive or negative and also determines the likelihood of the individual having Covid-19. The tradeoffs between recall and the overall accuracy have also been compared in this paper. Some models had high accuracy but low recall whereas some had high recall but low accuracy. So, the best model was chosen in such a way that both recall and accuracy had decent values. Use of class weights was also attempted as part of the work to improve data imbalance handling. This work is able to reduce false negative rate and improve recall appreciably and is a good preliminary analysis tool for distinguishing Covid-19 affected individuals from healthy individuals.

Recall and false negative rate can be improved further. Dataset used is relatively small and imbalanced; a larger dataset may lead to improved results. As the dataset size increases, the model can be retrained with new data. Dataset was heavily imbalanced with positive samples contributing to less than 9% of the total data and less than 150 samples. Imbalanced data handling can be further improved in the future.

References

[1] Abdelfatah Hassan, Ismail shahin, mohamed Bader Alsabek. Covid 19 Detection using RNN (through breathing, cough sounds and talking. 2020.

[2] Alexander Gaskell, Panagiotis Tzirakis, Alice Baird, Lyn Jones, Björn W. Schuller. End-2-End COVID-19 Detection from Breath & Cough Audio,Harry Coppock. 2021.

[3] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings. 2020.

[4] Jordi Laguarta, Ferran Hueto, Brian Subirana. COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. 29 September 2020.

[5] Miad Faezipour and Abdelshakour Abuzneid. Smartphone-Based Self-Testing of COVID-19 Using Breathing Sounds. 1 Oct 2020.

[6] Ali ImranabIryna, Posokhovabc, Haneya, N.Qureshia, Usama, Masooda,Muhammad, Sajid, Riaza Kamran Ali, d, Charles,N.Johna MD, IftikharHussain, be Muhammad Nabeel. AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. 2020.

[7] Rabab Yaasin and Walaa Gouda. Chest X-ray findings monitoring COVID-19 disease course and severity. 2020.

[8] Sofie Tilborghs, Ine Dirks, Lucas Fidona, Siri Willems, Tom Eelbode, Jeroen Bertels, Bart Ilsen, Arne Brys, Adriana Dubbeldam, Nico Buls, Panagiotis Gonidakis, Sebastian Amador Sanchez, Annemiek Snoeckx, Paul M. Parizel, Johan de Mey, Dirk Vandermeulen, Tom Vercauteren, David Robben, Dirk Smeets, Frederik Maes, Jef Vandemeulebroucke, Paul Suetens. Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients. 2020.

[9] Stephanie A. Harmon, Thomas H. Sanford, Sheng Xu, Evrim B. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. 2020.

[10] L. J. Muhammad, Ebrahem A. Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty & I. A. Mohammed. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. 2020.

[11] Pauline Mouawad, Tammuz Dubnov, Shlomo Dubnov. Robust Detection of COVID-19 in Cough Sounds. 12th Jan 202.

[12] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat Dimitris Spathis, Tong Xia, Pietro Cicuta Cecilia Mascolo. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. 18th Jan 2021.

[13] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy et al. Dicova Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. 2021.

[14] https://www.sciencedirect.com/topics/computer-science/cepstral-coefficient.

[15] https://flask-doc.readthedocs.io/en/latest/.

[16] https://keras.io/api/layers/convolution_layers/convolution2d/.

[17] https://towardsdatascience.com/learning-from-audio-the-mel-scale-mel-spectrograms-and-mel-frequency-cepstral-coefficients-f5752b6324a8

[18] https://keras.io/api/layers/core_layers/dense/.

[19] https://librosa.org/doc/latest/index.html.

[20] https://blog.getbootstrap.com/.

[21] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[22] https://scikit-learn.org/stable/modules/preprocessing.html.

[23] https://towardsdatascience.com/how-to-apply-machine-learning-and-deep-learning-methods-to-audio-analysis-615e286fcbbc.

[24] https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly.

[25] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38.

[26] https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/.

An Empirical Study on Fake News Detection System using Deep and Machine Learning Ensemble Techniques

T V Divya[1]
Research Scholar, Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation
Aziz Nagar, Hyderabad, Telangana

Dr Barnali Gupta Banik[2]
Associate Professor, Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation
Aziz Nagar, Hyderabad, Telangana

***Abstract*—With the revolution that happened in electronic gadgets in the past few years, information sharing has evolved into a new era that can spread the news globally in a fraction of minutes, either through yellow media or through satellite communication without any proper authentication. At the same time, all of us are aware that with the increase of different social media platforms, many organizations try to grab people's attention by creating fake news about celebrities, politicians (or) politics, branded products, and others. There are three ways to generate fake news: tampering with an image using advanced morphing tools; this is generally a popular technique while posting phony information about the celebrities (or) cybercrimes related to women. The second one deals with the reposting of the old happenings with new fake content injected into it. For example, in generally few social media platforms either to increase their TRP ratings or to expand their subscribers, they create old news that happened somewhere years ago as latest one with new fake content like by changing the date, time, locations, and other important information and tries to make them viral across the globe. The third one deals with the image/video real happened at an event or place, but media try to change the content with a false claim instead of the original one that occurred. A few decades back, researchers started working on fake news detection topics with the help of textual data. In the recent era, few researchers worked on images and text data using traditional and ensemble deep and machine learning algorithms, but they either suffer from overfitting problems due to insufficient data or unable to extract the complex semantic relations between documents. The proposed system designs a transfer learning environment where Neural Style Transfer Learning takes care of the size and quality of the datasets. It also enhances the auto-encoders by customizing the hidden layers to handle complex problems in the real world.**

Keywords*—Transfer learning; GANS; glove algorithms; word2vec; ensemble techniques; auto encoders; pre-trained models; word embeddings; BERT models

I. Introduction

A few decades back, researchers started working on fake news detection topics with the help of textual data. In the recent era, few researchers worked on images and text data using traditional and ensemble deep and machine learning algorithms, but they either suffer from overfitting problems due to insufficient data or unable to extract the complex semantic relations between documents. The proposed system designs a transfer learning environment where Neural Style Transfer Learning takes care of the size and quality of the datasets. It also enhances the auto-encoders by customizing the hidden layers to handle complex problems in the real world. The identification of fake news or phony news does treat as a sort of "Spam Detection," for which the reliability or accuracy of the model generally depends on the processing of the textual and image data. In this section, article will discuss a few popular textual and image pre-processing techniques. Initially, for the multi-media content, system have to work with images first. The dataset consists of three types of images, namely, a) Fake Image (tampered or morphed), b) Pristine Image (Resizing operation does perform to maintain standard size for all images), c) Image Splicing (it adds new content to the image either in image or text). The samples of fake news about celebrities that are have posted on Facebook does represent in Fig. 1. One of the fake news here is posted about celebrity with name "Tina Turner" stating that she is one of the top most among the haters list released by reputed organization, which is later found to be false. Like this many fake stories are released by few media and social networks to get the attention of viewers. The popular example for these type of examples is you tube people post false thumbnails on their videos irrespective of the content that might be displayed in the videos.

Any algorithm for working with these type of images, any model first need to separate the text and image as shown in Fig. 2.

Fig. 1. Fake News Published on Facebook.

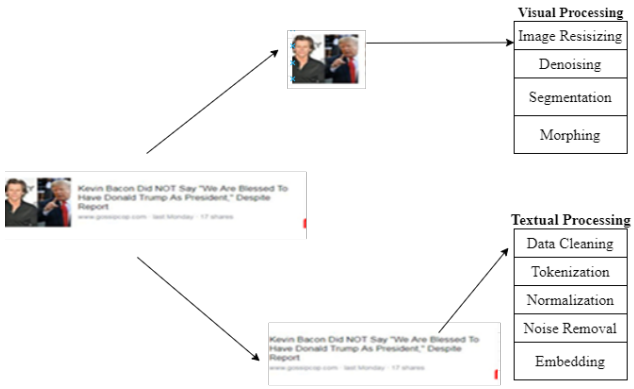


Fig. 2. Working with Multi-Media Content.

A. Image Pre-processing Techniques

This section will discuss mainly four types of processing techniques, as shown in Fig. 2. The image resizing does perform to standardize all the images to equal width and height because different articles will have different dimensions, then it becomes difficult for the neural network to give training based on the dimensions. The Denoise removal of noise steps helps the system improve the image's quality, majorly for the images captured with a low-quality lens or blur images posted due to the motion that occurred while capturing the image. Noises can exist in multiple forms, as shown in Fig. 3.

The segmentation process [22] helps find the interesting regions in the image to subtract the unwanted or uninterested parts from the original image. This process reduces the complexity as well as time to work with the image. The major pre-processing in the phony news detection system is finding the linkage or sequence in between the articles with the same headlines or continuing to a news article published a long back. These scenarios are taken care of by defining the transformation and morphing points or lines across the different images available in the dataset. All these methods mentioned are traditional, so to achieve an excellent accurate system, the researchers can replace these mechanisms with deep learning techniques like CNN [23], LSTM, and others, which are described in Table I.

Apart from these possible techniques, the best way to determine the quality of model is usage of GAN's to generate the duplicate images of different orientations by defining two types of layers one for generator and second layer for discriminator. The purpose of the generator is to create more number of relevant images and discriminator tries to identify the fake images.

B. Related Knowledge

In the implementation of GAN, there exists two components, namely, generator and discriminator. The major focus of the generator is to make the discriminator to believe the information generated about the fake images it has created. Since, the proposed model involves both text and images; the best case of GAN to implement is neural style GAN [21] because style GANS majorly focus on the tampered or morphed images rather than the real objects that exists in the image. The shape, color and edges are the crucial elements in the image processing. The extraction of these features should not impact the other layers of the network, so it's better to take the help from encoding mechanisms.

C. Text Pre-processing Techniques

In this section, the article discusses the basic techniques of NLP to deal with textual data extracted from the image. Data cleaning deals with removing stop and abuse words from the content, handling special characters, emojis, performing either stemming or lemmatization so that important base words are maintained, and others eliminate them from the content. The next step is to count the occurrences or frequencies in various new articles starting from character to sentence because the higher the frequency higher the priority to maintain the element in the content. The crucial text processing step is normalization because the semantic relation between the words (or) sentences does establish. The popular normalization technique is "POS Tagging," which labels each word in the sentence with parts of speech so that all the words might not be reduced, and sometimes over lemmatization might change the meaning of the entire sentence. The word embedding tries to find similar words with different representations and merge them to have a unique representation in the entire article so that the reader finds fewer complications with the interchange of similar words. The overall tentative proposed system does represent in Fig. 4, which can solve the major difficulties identified in the previous systems.

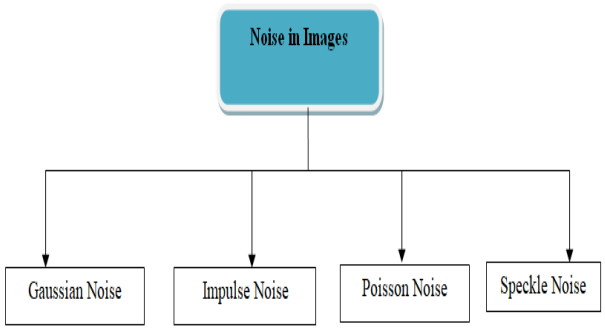


Fig. 3. Types of Noises in Image Processing.

TABLE I. DESCRIPTION OF NEURAL NETWORK TECHNIQUES

S.NO	Name	Description
1	CNN	In general images are represented in the form of 2D matrix, to explore all possible features; the CNN represents them as a 3D matrix. The extraction of features are performed with the help of hidden layers
2	LSTM	It tries to identify the features from the previous predictions which are arranged in an sequential manner taking the help of circuit gates
3	MLP	It is a fully connected layer and data units are transferred through all the layers for processing. These type of networks are applied for applications where speech act as input and to work with complex classification systems
4	RBFNN	In this model, the input vectors are compared to find the similar data points and every vector is compared against all the possible classes of the neural networks.

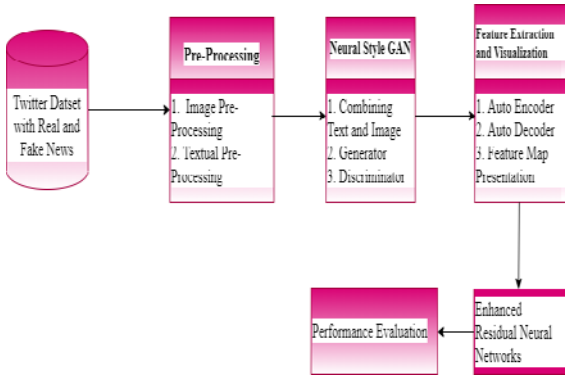


Fig. 4. Brief Overview of the Fake News Detection System.

In the below sections, the paper is organized as follows, the literature survey provides a glance of all the previous existing mechanisms and results and discussion section discusses their advantages and disadvantages to find the limitations that exists in the previous works along the datasets they have used. The observations section identifies the accuracies of all the models and describes about the best model it identified.

II. Literature Survey

In [1], Xinyi Zhou et al. explored a multi-modal analysis to detect fake news using both text and images. First, it extracts the information from the image and runs the similarity matching algorithm with different types of contents. The textual features extract using the text-CNN, in which a fully connected layer does attach to the normal CNN, and the important contents did extract after passing through each layer. For processing the images, a pre-trained model known as "image2sentence" is implemented. An independent model is developed by performing the cross-entropy modal. Every image does label with multiple classes, and the probability distribution function access the correct prediction label by using the concept of majority voting schema.

In [2], Priyanshi Shah et al. designed a framework by analyzing the tweets, which are made available in public datasets. The textual extraction performs sentiment analysis, i.e., by identifying the polarity value for all the tweets. In visual extraction, segmentation is applied by combining wavelets with K-Means and extracting nine important images. The features extracted from text and visual are represented as a matrix and passed as input to the optimal algorithm proposed by the author. Then it identifies the relevant contents and is passed to the SVM classifier with RGBF kernel function to classify the image.

In [3], Anastasia Giachanou et al. have studied multi-model evaluation using two datasets and by combining the neural networks with semantic search. The tags on images were observed and did pass as input for the local binary patterns algorithm, which finds the similarity between the text and images. The model can identify the different images with the same content. In [4], Deepak Mangal et al. proposes text cues to extract features from images by implementing the LSTM technique in neural networks. It removes the information from tag line and compares it with various news articles by using the cosine similarity mechanism. The visual features do train using VGG, and text features do train using word2vect corpus. The LSTM uses three gates to generate the classification by using useful information. The non-linearity function helps the layer identify any short sequences that occur but has a great role to play in predicting.

In [5], Tong Zhang presented a novel approach known as "BERT" implementation in neural networks for identifying the essential domain features. In this model, a component called "BDANN" identifies irrelevant features and integrates with the remaining modules to focus on the event-specific tasks. The model works in both forward and backward directions, and it works on the domain classifier by making them into sub-groups with defining specific patterns for each group to identify in the future quickly. This model has achieved an accuracy of 98.1% and presented well even in the case of true label prediction.

In [6], SomyaRanjanSahoo et al. experimented with an automatic system that can detect multiple features to identify the fake news posted on social networks. This model first analyzes the profiles of the users through a web crawler attached to the browser to get the features associated with the user. These extracted features apply machine learning like KNN, SVM, Logistic Regression, and deep learning techniques like LSTM to classify real and fake news. The model has divided features into two significant categories: contents related to profile and contents newly created, and tries to generate a mapping technique between the previous posts and new ones.

In [7], Aman Agarwal et al. explored a new approach in neural networks by blending different deep learning algorithms. The model tries to establish a relationship in terms of language by constructing a word vector using a novel algorithm known as "GLOVE" to process the textual data. The beauty of word embedding lies in its representation of sentimental meaning and their corresponding relation in geometric shapes, which helps visualize the vector space. The reason behind this shape utilization is, it marks how far two different words did match one against the other. Later, these different vector spaces are organized into a bag of words, using Keras as a pre-trained model to construct a sequential model. The major work in this system focuses on turning the parameters associated with convolution and recurrent neural networks. Using the CNN model and embedded trained matrix, N-gram higher-order representation does construct to find the co-relations between the different sequences to construct a feature vector map. With this matrix, there is a possibility of obtaining non-linear transformations in the model, which the LSTM helps classify the temporal and essential features among the text data available. This model also helps in handling the unlucky data that might be present in the model by any chance by defining the threshold values, by adding the generalized layers like dropout, normalizations, and others.

In [8], Juan Cao et al. discussed detecting fake news in different multi-media channels. The model has designed deep learning techniques incorporated with image processing compression techniques to identify the manipulations or

tampered text within the images. The major role of CNN is to extract the semantic and content features associated with the images, and later it performs statistical measures to perform feature engineering. In this system of adversarial image reconstruction, the first part of the algorithm deals with metadata generation from the top k-features extracted from the reference dataset, and the second part of the algorithm act as a consistency verifier to generate the evidence that is against the tampered image because it is essential to know the relationship between the image and text to make some inferences and decisions.

In [9], Yin-Fu Huang et al. modeled an Adaptive Harmony based genetic algorithm using ensemble techniques of deep learning. The highlight associated with this model is, it has successfully processed data based on their dimensions. For example, suppose the tokenization of words has to perform, then the embedded LSTM takes care by extracting the semantics in the form of 300 dimensions. Similarly, the depth of the sentence and parsing of grammar is taken care of by depth LSTM. From these different LSTM's, an optimized LSTM with minimum weights is constructed. The classification of news is taken care of by heuristic genetic algorithm, which consists of first random selection, whose performance computed using the objective function, which defines as the minimum weight of the output node then in the second step, the value of the node updates if and only if the memory consideration accepted. In every iteration, it adjusts the pitch such that every time it selects the elements that are less than the threshold value and simultaneously reduces the loss function by wrapping the MSE error function with a harmony algorithm.

In [10], AnshikaChoudhary et al. designed a framework based on language to detect fake news. This system uniquely concentrates on syntax related to the sentence instead of semantics by computing the density of each word. During this process, it takes three statistical measures into account; one is count, which computes the occurrence of each word, second one deal with sentences to extract the sentiment associated with it in the form of polarity, which represents the negative and positive words and subjectivity, which assigns the score as either 0 or 1. The last one takes care of the properties associated with grammar by computing the readability of the sentence based on the pattern of writing. In deep learning algorithms, all the features are standardized and passed as input to the network with two hidden layers, which designs a variant of LSTM by taking the input gate's input because it determines the number of updated values do add to the four neurons. The layer might feel some of the features are unimportant in the next phase, so the forget gate can control this. Finally, using the activation function tanh and the dot product of weighted sums, the memory gate computes the final value at the output node.

In [11], JiangfengZeng computed deep semantic correlations between the textual data and multi-media images by using pre-trained models like VGG to deal with variations in hierarchical representation systems. The model knows the complexities involved in training with the machine learning model, so it has opted for the computation of Eigenvectors to train the model, which in turn helps in the image enhancement process. The system takes news articles, i.e., text documents and the image, associated with annotated labels as input. The system uses the concept of word embedding for processing the textual data, for converting the words which are semantically closer as a group of vectors, and VGG-19 to train the features related to the images in the transfer learning environment because, in the end, they produce the output as an encoded image. The bi-directional LSTM integrated with soft attention encodes the word to extract the complete sentence in terms of grammar and other notations and represents them in a hierarchal format known as "DOM"; these can traverse in both directions. At last, the visual content connects with two dense, fully connected layers to perform decoding of both textual and visual features.

In [12], Rohit Kumar Kaliyar et al. designed a tensor decomposed matrix to work with the BuzzFeed dataset by integrating the machine learning algorithm "eXtreme Gradient Boost" algorithm with a proposed algorithm known as "DeepFake." This system involves only detecting fake in terms of social context, the construction of which needs two major inputs is contents related to the article, and the second is a matrix that defines the relation of users with different communities. The contents convert into an N-grams count vector, representing the occurrence of a particular word in different news articles. The relationship matrix converts into couple matrix factorization and mode-1 participation by constructing echo chambers, where people share their opinions about it a particular post. These matrices help compute the interests of different communities based on their comments and relationships in the network. The DeepFake architecture has designed a 4-layered network with correct regularizations.

In [13], Muhammad Umer et al. experimented with stance detection using multi-classification techniques with the help of neural networks. The headlines associated with the news article are converted into a 100-dimensional vector so that all the details about the report are stored clearly and projected into n-dimensional space, which is geometrically represented as a dense shape and is taken care of by chi-square PCA model, to reduce the low latent space vectors. The covariance, a relation between the dimensions and observations, helps the PCA reduce the number of features considered for constructing a classifier. The content in the images is categorical data, so the computation of statistical measures is done by chi-square test. The model has to predict one output out of 4 class labels, so; the last constructs, with the help of a densely connected layer, attach with softmax activation function, which is popular in handling the multi-classification, reduces the burden of the overfitting problem.

In [14], Stephane Schwarz et al. designed EMET, which considers users' responses for a particular post by performing multi-classification as true, false, and unknown. These three values are projected onto a latent space using NLP techniques; corresponding transformers did generate. The author has chosen LS representation to draw the inferences, decisions, and finally to construct a two-dimensional matrix between the post, comments, and users. The input signal transformation has marked its place in the NLP world because it generates bitextual pairs, draws the semantic relations between the closest words, and answers for a particular search query

displays with the rules that match with expert system shells. The model has used a hyper tuned CNN to classify the images by finding the best number of filters, strides, filter size, and several neurons.

In [15], SawinderKaur et al. used the concept of voting at multi-levels to identify fake news. This model has experimented with three features extraction techniques implemented as a step in ML algorithms LR and Linear SVC. All the traditional cleaning steps take care of initial processing, and mean values for different types of articles present in different datasets are measured. The System generates features; a sparse matrix does publish, using different computations like TF-IDF, hashing, and count vectors. In the TF-IDF vector, the relationship does base on the count and its associated weight; for every word, the probability of word occurrence in each document is represented in the interaction cell.

In contrast, the count vector is a numerical representation. The hashing is a novel approach implemented to store the values in buckets based on the calculated remainder value; the major advantage of this approach is that it efficiently handles the memory space during the extraction process. The reason for selecting a multi-level algorithm is to reduce the training time and to generate the outputs in parallel by using the voting classifier. The voting classifier picks the top 3 algorithms with high false rates and ensemble them because the higher the false rate, they are weak the classifiers. The model once again constructs the soft voting classifier and predicts the label based on their false positive rate from the obtained results.

In [16], Mohammad HadiGoldani et al. incrementally designed a margin loss CNN. This research aims to create a less error-prone function that reduces the cross-entropy generated by the softmax activation function. In general, the softmax functions don't adapt to the new environment, so to overcome this drawback; this research has extended the activation function with reinforcement learning. This mechanism helps in generating related featured during the process of adaptive learning. The model uses intracluster property while performing the compaction technique and inters cluster property for separating the dissimilar characteristics. The defined lambda value specifies the margin value empirically predicts the class labels.

A. Gaps Identified

Few researchers worked on the textual data to identify the fake news and few worked on both images and text. The working on images and text involves complex operations so to simplify the model, it is needed a design an architecture which reduces the cost of operations that are involved in image processing. It is observed that most of the systems find difficulty in finding the semantic relation between the different images along with their text. The popular technique for feature reduction is PCA, but it cannot handle the images where the text is embedded in the image so the usage of variant auto encoders and decoders might help the system to expose both the image features and textual features. Most of the researchers utilized the predefined architectures due to which sometimes it may suffer from bottle neck problem. The solution to this problem might be provided with the usage of transfer learning.

III. Results and Discussion

All the existing systems have successfully implemented the detection of fake news in different social media platforms. Table II compares the methods used in previous case studies and their usage of datasets and identifies the limitations found to address the research gap for implementing a better system with efficient metrics in terms of all objective functions that a deep learning system needs.

To understand the flaws and advantages in terms of various metrics, article has tabulated all the results obtained from different old research systems in Table III.

TABLE II. Comparative Study on Existing System of Fake News Detectors

Author	Approach Implemented for Feature Extraction	Algorithms Implemented for Classification	Dataset (s) Used	Limitations
Zhou[1]	Multi-Modal	Similarity Aware Capture Model	1. PolitiFact 2. GossipCop	Articles that are in pair format from different sources but published differently are not taken care
Shah[2]	Cultural Algorithm	SVM	1. Weibo 2. Twitter	The model cannot handle the complex relations between the text and visual, if any
Giachanou[3]	Google Word Embedding Layer + LBP	Inception+Xception	1. PolitiFact 2. GossipCop 3.MediaEval	It is not able to produce the decision with neutral polarity.
Mangal[4]	Textual- word2vec Visual- VGG Net	LSTM+ Cosine Similarity	Twitter	Semantically irrelevant data didn't consider here.
Zhang[5]	Multi-Modal	Domain Classifier (BDANN)	1. Weibo 2. Weibo Filtered	The domain classification rate further extended by computing the probability rate
Sahoo[6]	Web Crawler in browser	1. Group of ML Algorithms 2. Group of DL Algorithms	Own Dataset	In ML algorithms, ensemble techniques, and the DL algorithms, Hyper tune parameters can yield better results.
Agarwal[7]	N-gram through Convolution	LSTM	Kaggle	The dataset with an early annotation mechanism helps in the good clustering of characteristics.
Cao J[8]	CNN+word2vec	LSTM with multiple packages, AIRD	Own Dataset	The Metadata generator takes much more time.
Yin-Fu Huang[9]	Ensemble LSTM+ N-Gram CNN	Self Adaptive HS	Cross-Domain Dataset	It deals with only headlines mentioned in the image.

Choudhary[10]	Syntax+ Semantics+ Readability+ Correlation	Feature based sequential neural networks	BuzzFeed	The memory and temporal data should do taken into consideration for efficient implementation
Zeng[11]	Multiple Components Encoder with Soft Attention	LSTM	1.Twitter 2.Weibo	Usage of GANS can improve the Size and Quality of the dataset.
Kaliyar[12]	Tensor Formation	XGBOOST+CNN	1.Buzzfeed 2.PolitiFact	The activation function selected suffers from a zero gradient problem.
M. Umer[13]	PCA	CNN+LSTM	FNC	PCA is the traditional approach. Advanced techniques like variational autoencoders can extract high-quality and related features.
S. Schwarz[14]	Signal transformations	EMET	Twitter	Data augmentation techniques did fail in the case of an image with side annotations.
Kaur[15]	TF-IDF, Count, and hashing vectors	Multi-level voting system	1. NewsTrends 2.Kaggle 3. Reuters	Further research should control Impulsive data at an early stage.
Goldani[16]	Embedding Layer	CNN	LIAR	The embedded layer used is static.
Kaliyar[17]	Glove+Pre-trained word embeddings	Deeper CNN: HCNN+HLSTM	Kaggle	Instead of traditional neural networks, residual or transfer learning neural networks can handle complex relations and associations**.**
Hamdi T[18]	Twitter API	Node2vec selection classifier	CREDBANK	It has to generate synthetic works to handle the abnormal relations findings.
Masciari E[19]	Word Embedding	Google BERT	1. LIAR 2.PolitiFact	Diffusion mechanisms integrated with BERT improve the space complexity of the environment.
A. Agarwal[20]	N-Grams	Different Algorithms in ML & DL	LIAR+Kaggle	Hyper turned LSTM with enhanced activation functions is needed.

TABLE III. METRICS COMPARISON TABLE

Authors	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
[1]	PolitiFact	0.874	0.889	0.903	0.896
	GossipCop	0.838	0.857	0.937	0.895
[2]	Weibo	0.891	0.873	0.822	0.932
	Twitter	0.798	0.791	0.833	0.760
[3]	PolitiFact	0.925	0.911	0.911	0.911
	GossipCop	0.829	0.815	0.815	0.815
	MediaVal	0.622	0.885	0.885	0.885
[4]	Twitter	0.91	0.909	0.913	0.910
[5]	Weibo	0.85	0.869	0.836	0.852
	Weibo Filtered	0.865	0.850	0.920	0.88
[7]	Kaggle	0.91	0.97	0.925	0.946
[9]	Cross Domain	0.916	0.964	0.918	0.929
[10]	BuzzFeed	0.841	0.77	0.84	0.812
[11]	Twitter	0.772	0.813	0.741	0.775
	Weibo	0.839	0.867	0.784	0.828
[12]	BuzzFeed	0.856	0.833	0.869	0.851
	PolitiFact	0.886	0.821	0.846	0.84
[13]	FNC	0.978	0.974	0.982	0.978
[14]	Twitter	0.940	0.913	0.912	0.916
[15]	NewsTrends	0.93	0.958	0.916	0.937
	Kaggle	0.98	0.988	0.98	0.983
	Reuters	0.961	0.968	0.95	0.958
[16]	LIAR	0.99	0.967	0.945	0.95
[17]	Kaggle	0.983	0.994	0.96	0.98
[18]	CredBank	0.98	0.98	0.98	0.98
[19]	LIAR	0.588	0.565	0.449	0.528
	PolitiFact	0.448	0.443	0.495	0.473
[20]	LIAR+Kaggle	0.97	0.965	0.951	0.958

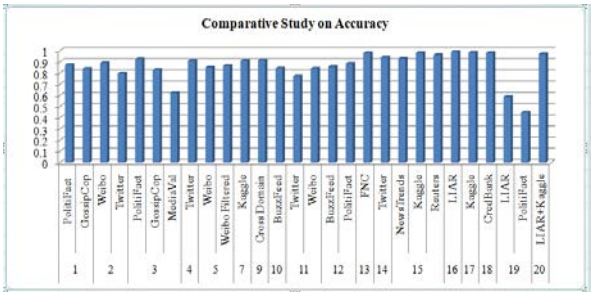


Fig. 5. Comparative Study on Accuracy Metric.

Fig. 5 describes the comparative study on accuracy metric, and from the figure, it has been found that author [16] has achieved nearly 99.9% accuracy on the Liar dataset, which is highest in terms of both similar and non-similar datasets [20][21]. The author has implemented a simple static word embedding technique to extract the features. Still, in further research, one can try to implement the non-static and redundant dimensions to improve the learning rate and other adaptability parameters because accuracy alone cannot justify the model's reliability.

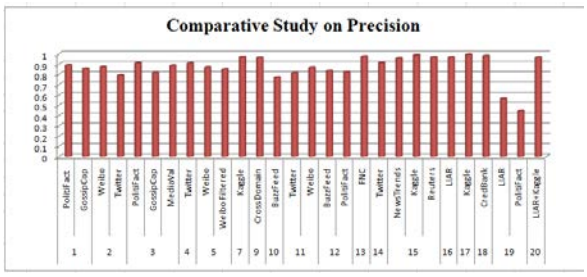


Fig. 6. Comparative Study on Precision Metric.

Fig. 6 clearly states that the author [17] [22-24] has achieved the highest precision rate, “99.4%,” by working on the Kaggle dataset. The model has successfully implemented hyper-turning of estimators inclemently, which involves checking of 2**n combinations, increasing the execution and training time. So, in further research, the selection of combinations can be reduced by creating a transfer learning environment.

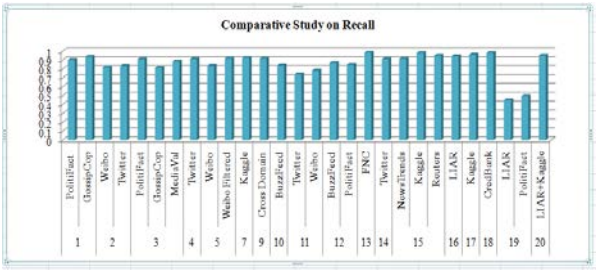


Fig. 7. Comparative Study on Recall Metric.

Fig. 7 states that the author [13] has achieved the highest recall rate, “98.2%”, by working on the FNC dataset. Still, it has applied traditional PCA algorithms for feature reduction and extraction, which further improves by defining the ensemble or variational autoencoders, which has proved its capability in extracting the visual features from multimedia content.

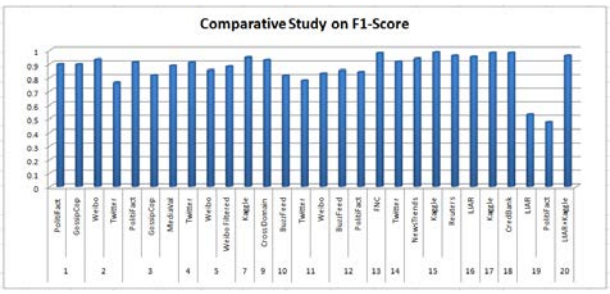


Fig. 8. Comparative Study on F1-Score Metric.

Fig. 8 showed that the author [15] had obtained an F1-score of “98.3%” on the Kaggle dataset by experimenting with three types of feature vectors and designed its classification algorithm, but it suffered when the data, either images or text are in impulsive nature so, in the early stages of model designing, one has to take care of these type of persistent relations.

IV. Observations

Table II inferences that [19], [3c], and [11] are inferior in performance in terms of all metrics. So this problem can be addressed by neural style transfer learning GANS because they can handle both images and text very effectively. The reason for selecting GANS does explain in two situations, one where the pre-processing step fails to separate the visual and text and the second where the number of images is significantly less to design a model. Next, the proposed system from the above study identified that [16] and [17] are the top two systems in identifying phony news. From these limitations, the system identified the second research gap has to take care of feature extraction with the latest deep learning techniques like autoencoders, by hyper turning and by integrating an enhanced activation function at its output layer, which is dense as fully connected layers. So that the high complex features propagate to the next layers as a simple linear input, finally, to improve the overall capability of the system, the latest transfer learning approach implements by incorporating residual neural networks with pre-trained models to reduce the time for training and to define the correct lambda values to predict the correct output labels associated with the annotated images.

V. Conclusion

This extensive literature survey helps the system to identify the drawbacks associated with the previous methodologies. From this study, the system has identified that Natural Language Processing tools like N-grams, SVD decomposition, Lemmatization’s and other simple processing techniques are efficient to work on the text extracted from the images. To work with the images, instead of simple CNN techniques, it is better to design modified auto encoders which can efficiently solve the problem of bottle neck in linear amount of time. The feature extraction is the crucial step in fake news detection because the text objects in the image play a vital role in deciding whether it is real or fake news. For the classification of news, it is better to implement ensemble algorithms or Meta classifiers, which achieves good accuracy and true positive rates. The ensemble algorithms have proved

their efficiency in terms of AUC and Kappa Statistics measurement. In future, to detect fake news detection, the proposed research first implements GANs to increment the size of data then it extracts the features using AEDECNN and then it performs classification using ensemble mechanism.

GLOSSARY

S.No	Term	Full Form
1	CNN	Convolution Neural Networks
2	GAN	Generative Adversial Networks
3	LSTM	Long Short Term Memory
4	RBFNN	Radial Bias Function Neural Network
5	MLP	Multi Layer Perceptron
6	SVD	Single Value Decomposition
7	AUC	Area Under Characteristics

REFERENCES

[1] Zhou, Xinyi, et al. “SAFE: Similarity-Aware Multi-Modal Fake News Detection.” Advances in Knowledge Discovery and Data Mining, edited by Hady W. Lauw et al., vol. 12085, Springer International Publishing, 2020, pp. 354–67. DOI.org (Crossref), doi:10.1007/978-3-030-47436-2_27.

[2] Shah, P., &Kobti, Z. (2020). Multimodal fake news detection using a Cultural Algorithm with situational and normative knowledge. 2020 IEEE Congress on Evolutionary Computation (CEC). doi:10.1109/cec48606.2020.9185643.

[3] Giachanou A., Zhang G., Rosso P. (2020) Multimodal Fake News Detection with Textual, Visual and Semantic Information. In: Sojka P., Kopeček I., Pala K., Horák A. (eds) Text, Speech, and Dialogue. TSD 2020. Lecture Notes in Computer Science, vol 12284. Springer, Cham. https://doi.org/10.1007/978-3-030-58323-1_3.

[4] Mangal, D., & Sharma, D. K. (2020). Fake News Detection with Integration of Embedded Text Cues and Image Features. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). doi:10.1109/icrito48877.2020.9197817.

[5] Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., & Cui, L. (2020). BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. 2020 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn48605.2020.9206973.

[6] Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing, 100, 106983. https://doi.org/10.1016/j.asoc.2020.106983.

[7] Agarwal, A., Mittal, M., Pathak, A. et al. Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning. SN COMPUT. SCI. 1, 143 (2020). https://doi.org/10.1007/s42979-020-00165-4.

[8] Cao J., Qi P., Sheng Q., Yang T., Guo J., Li J. (2020) Exploring the Role of Visual Content in Fake News Detection. In: Shu K., Wang S., Lee D., Liu H. (eds) Disinformation, Misinformation, and Fake News in Social Media. Lecture Notes in Social Networks. Springer, Cham. https://doi.org/10.1007/978-3-030-42699-6_8.

[9] Yin-Fu Huang, Po-Hong Chen, Fake news detection using an ensemble learning model based on Self-Adaptive Harmony Search algorithms, Expert Systems with Applications, Volume 159,2020, 113584, ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2020.113584.

[10] Choudhary, A., & Arora, A. (2021). Linguistic feature based learning model for fake news detection and classification. Expert Systems with Applications, 169, 114171. https://doi.org/10.1016/j.eswa.2020.114171'.

[11] Zeng, J., Zhang, Y., & Ma, X. (2021). Fake news detection for epidemic emergencies via deep correlations between text and images. Sustainable Cities and Society, 66, 102652. https://doi.org/10.1016/j.scs.2020.102652.

[12] Kaliyar, R.K., Goswami, A. & Narang, P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J Supercomput 77, 1015–1037 (2021). https://doi.org/10.1007/s11227-020-03294-y.

[13] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. -W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," in IEEE Access, vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[14] S. Schwarz, A. Theóphilo and A. Rocha, "EMET: Embeddings from Multilingual-Encoder Transformer for Fake News Detection," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2777-2781, doi: 10.1109/ICASSP40776.2020.9054673.

[15] Kaur, S., Kumar, P. &Kumaraguru, P. Automating fake news detection system using multi-level voting model. Soft Comput 24, 9049–9069 (2020). https://doi.org/10.1007/s00500-019-04436-y.

[16] Goldani, M. H., Safabakhsh, R., &Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 58(1), 102418. https://doi.org/10.1016/j.ipm.2020.102418.

[17] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet – A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32–44. https://doi.org/10.1016/j.cogsys.2019.12.005.

[18] Hamdi T., Slimi H., Bounhas I., Slimani Y. (2020) A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding. In: Hung D., D´Souza M. (eds) Distributed Computing and Internet Technology. ICDCIT 2020. Lecture Notes in Computer Science, vol 11969. Springer, Cham. https://doi.org/10.1007/978-3-030-36987-3_17.

[19] Masciari E., Moscato V., Picariello A., Sperli G. (2020) A Deep Learning Approach to Fake News Detection. In: Helic D., Leitner G., Stettinger M., Felfernig A., Raś Z.W. (eds) Foundations of Intelligent Systems. ISMIS 2020. Lecture Notes in Computer Science, vol 12117. Springer, Cham. https://doi.org/10.1007/978-3-030-59491-6_11.

[20] laiahKavati, A. Mallikarjuna Reddy, E. Suresh Babu, K. Sudheer Reddy, RamalingaSwamyCheruku,Design of a fingerprint template protection scheme using elliptical structures,ICT Express,Volume 7, Issue 4,2021,Pages 497-500,ISSN 2405-9595,https://doi.org/10.1016/j.icte.2021.04.001.

[21] Ayaluri MR, K. SR, Konda SR, Chidirala SR. 2021. Efficient steganalysis using convolutional auto encoder network to ensure original image quality. PeerJ Computer Science 7:e356 https://doi.org/10.7717/peerj-cs.356.

[22] A. M. Reddy, V. V. Krishna, L. Sumalatha and S. K. Niranjan, "Facial recognition based on straight angle fuzzy texture unit matrix," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, 2017, pp. 366-372, doi: 10.1109/ICBDACI.2017.8070865. (C 5).

[23] M. Reddy, K. SubbaReddy and V. V. Krishna, "Classification of child and adulthood using GLCM based on diagonal LBP," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015, pp. 857-861, doi: 10.1109/ICATCCT.2015.7457003.

[24] R. T, G. Sirisha and A. M. Reddy, "Smart Healthcare Analysis and Therapy for Voice Disorder using Cloud and Edge Computing," 2018 4th International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Mangalore, India, 2018, pp. 103-106, doi: 10.1109/iCATccT44854.2018.9001280.

DoItRight: An Arabic Gamified Mobile Application to Raise Awareness about the Effect of Littering among Children

Ayman Alfahid[1], Hind Bitar[2], Mayda Alrige[3], Hend Abeeri[4], Eman Sulami[5]
Computer Science, Majmaah University, Majmaah, Saudi Arabia[1]
Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia[2, 3, 4, 5]

***Abstract*—Littering contributes significantly to environmental pollution. Previous studies have noted that children are more likely to litter than adults. This target age group can be easily reached through mobile applications and games. Therefore, this study aims to investigate the effect of a gamified application in raising awareness on the effect of littering in the environment. We developed a gamified app, called DoItRight to promote an environment friendly behavior and improve the littering behavior of children. The DoItRight app is in Arabic language and targets children between 5 and 13 years old. It is a gamified application that enables kids to learn the importance of picking up litters and dropping it in trash cans. The app was evaluated using the System Usability Scale (SUS) standardized instrument which was administered on the target audience. The results of the evaluation showed that the DoItRight app has an SUS score of 93.25 which represents an A+ grade and a percentile range of 96 to 100. This indicates that the DoItRight app is technically usable and can potentially serve the purpose of increasing kids' awareness about the downsides of littering on the environment.**

***Keywords*—Littering; mobile application; gamification; children intention; raise awareness; behavior change Saudi Arabia**

I. Introduction

Over the years, human practices have caused several environmental hazards which can inhibit future growth and wellbeing. The collective impact of human behaviors on the environment further threatens our ecosystem [1] – [3]. One of such behaviors is littering which can either be accidental or intentional. Therefore, identifying factors responsible for high rate of littering is critical to design an efficient littering-reduction intervention. A study carried out in 2020 used the Motivation, Opportunity, Ability and Behavior (MOAB) framework to examine the individual littering behaviors of Saudi citizens. It conducted twenty-five semi-structured interviews on individuals between 20 and 40 years in Saudi Arabia. The outcomes of the study showed that the lack of knowledge, at the individual level, influences littering behavior. Other factors found include social norms and the built environment. Therefore, any plan to reduce littering must include the implementation of an individual-level strategy [4].

Additional individual and environmental factors influencing littering behavior in Saudi Arabia have also been unraveled. Recent research observed 362 individuals and their surrounding environments for twelve days in Saudi Arabia [5]. Findings revealed that littering rates were higher in environments that were less attractive and where trash cans were far from individuals. Also, the study found out that younger people are more likely to litter than older people.

Meanwhile, the global increase in the use of smartphones has led to the emergence of mobile applications being developed and deployed for education and enlightenment on several issues including environment and health (for instance, mhealth systems) [6]. Also, researchers have acknowledged the effectiveness of mobile applications in educating children. Therefore, the use of a mobile app to raise children's awareness of the threats posed by littering can prove to be an effective litter-reduction strategy [7]. Therefore, this research designs, and evaluate an Arabic gamified mobile application following user-centered design approach, called "DoItRight" to increase children's awareness and intent towards littering. It targets children from 5 to 13 years.

A. User-Centered Approach

As the name implies, a user-centered design (UCD) approach puts the user at the center of the design process. That is, the designer of a product or application considers the needs, limitations, and interests of the end users and designs a product that meets those needs. Having an understanding of who we design for, what they need, and their environmental circumstances is an effective way to ensure that a product or design is successful. It is also a good way of preventing a bad design that can potentially frustrate users [8]. The UCD approach involves understanding user context, defining requirements, developing solutions, and evaluating the outcomes with respect to the users' requirements and contexts [9]. In this study, we adopted a user-centered design approach to develop the DoItRight application.

The rest of this paper is organized as follows: Section 2 explores the related work in the existing literature. Section 3 presents the methodology adopted in conducting the research. Results are presented and discussed in Section 4 while Section 5 concludes the paper and sets the direction for future work.

II. Related Work

A. Improving Children's Littering Behavior

Education in different forms can be used to improve children's behavior towards the environment. Reference [10] reported outcomes of the "We Love Reading Program" that

leveraged the reading of social stories to address littering and environmental issues across different communities in Jordan. The results indicated that the program improved children's knowledge of littering issues and created a positive change in the behavior of the kids.

The author in [11] explored 5th grade children's solutions to littering and environmental pollution using the kids' drawings. Forty children at the age of 10 and 11 years, including 25 girls and 15 boys, completed drawings to provide solutions and future plans to littering and environmental pollution. The study utilized three themes to analyze the drawings; these include persuasion, physical action, and political action. However, most of the children found it difficult to visually express their thoughts on persuasion and political action. The major solutions common in the drawings include collecting litters, dropping litters in the trash can, and planting saplings [11]. The children in the research knew littering is a great environmental challenge that needs to be addressed with more than one strategy.

In a recent publication [12], researchers echoed the fact that a proper upbringing of a child potentially represents a great beginning towards developing a good littering behavior. A survey administered by the study recorded responses from 2,349 individuals. The outcome of the survey showed that low-income households need support in terms of quality education and disposal facilities while high-income households need support in terms raising awareness on the problems of littering. Though self-initiative and parental guidance constitute a good approach towards improving the littering behavior of children, they are not sufficient [12]. The researchers suggested that future research should explore the effectiveness of religious education on the littering behavior of children.

To find a lasting solution to the littering problem, it is important to understand the factors that motivate people to dump litters in the environment or in the bin [13]. Researchers [13] explored these motivating factors by conducting a survey that asked people to indicate what motivates them to drop litters in the bin. An analysis of the survey responses divided the motivators into intrinsic and extrinsic motivators. The results showed that sense of morals, ethics and upbringing were the highest-level intrinsic motivators while the highest-level extrinsic factors were the presence of kids, being in a clean place, and recycling programs. Therefore, to create a successful anti-littering campaign in developing countries, stakeholders should combine intrinsic and extrinsic motivators. The study [13] noted that, for the intrinsic motivators, authorities should remind people of their core values, morals and ethics. For the extrinsic factors, authorities should create a convenient infrastructure, recycling programs, rewards, and penalties.

B. Use of Digital Technologies to Address Littering Issues

Some existing research works have explored the possibility of leveraging digital technologies and mobile applications to improve the littering behavior of children. Researchers [14] concluded that it is extremely important to raise children's awareness towards making sustainable choices when buying, using or dismissing products. One of the solutions proposed was the development of Contact from the future, a digital game focusing on plastic pollution education of children. After defining the requirements and objectives of the game, the study designed and developed the game application. The ultimate goal was to raise awareness and stimulate pro-environmental behaviors in children.

Another study [15] identified inadequate environmental education or awareness at the early age as a critical cause of littering and environmental problems. To address the issue, the study developed an Android-based sorting waste game to teach children the different kinds of waste, (i.e., organic and inorganic) and the appropriate litter box to drop each waste. The Waterfall Development Model was adopted meaning that the development of the sorting waste game followed a 4-step process including analysis, design, implementation, and testing. The functionality of the game was tested and the outcome indicated that the game is user friendly, runs smoothly on Android, and can serve as an effective medium of gaining environmental education.

Also, an exploratory study [16] was conducted to understand how children's use of digital technology, especially iPads, can impact outdoor environmental education programs. In particular, iPads were integrated into water quality education for 5th grade children. The qualitative observations obtained from the study were analyzed and seven major themes emerged. They include children's reaction to mobile devices, digital natives versus immigrants, group interactions, mobile devices in the hands of kids, instruments for learning, nature prevails, and introduction of mobile devices. These themes offered new insights to understand best practices for technology integration into littering and environmental education.

The author in [17] examined the effectiveness of web-based animation videos in the environmental education of elementary school pupils. The research [17] developed web-based animation video pages for climate change, waste recycling, mangrove forest, ozone depletion, and biodiversity. The impact of the intervention was measured via pupils' level of natural curiosity, improvement in environmental awareness, and the pupils' level of pro-conservation values. The outcome of the evaluation revealed that the use of web-based animation videos has positive effect on pupils' littering and environmental education.

Moreover, other researchers [18] have equally emphasized the urgent need to consider the use of virtual and augmented realities, videoconferencing, and mobile apps to engage primary school children with the ultimate goal of preventing littering and restoring the environment. The research [18] indicated that these technologies can capture children's interest while enabling them to learn important practices in preventing littering and protecting the environment. Quick Response (QR) codes are another example of a technology that can be integrated into mobile learning technology in littering education [19]. When attached to an object, the QR codes can add a layer of digital functionality thereby empowering users of mobile devices to access information without any restriction.

III. Methodology

A. DoItRight App Intervention Design and Development

We followed four main stages in the design and development of the DoItRight App:

1) Understanding user context

2) Obtaining requirements from parents and families of children who are target audience of the proposed application.

3) Creating a prototype of the app features based on the requirements collected in stage 2 as well as a review of similar mobile applications in the market.

4) Deploy and test the DoItRight app for evaluation; create a formative evaluation protocol to get feedback that can be used to further refine the app.

In stage 1, we determined the target audience in terms of who they are, their access to smartphones, age group, level of education etc. In stage 2, we administered a survey filled by parents and families to enable us to identify and define user requirements while also exploring the operational feasibility of the proposed DoItRight app.

In stage 3, the team reviewed existing mobile applications in the market with similar features and target audience to DoItRight app. The team also developed all necessary features to meet the requirements obtained in stage 2 and fill the gaps in the existing similar mobile applications. In stage 4, the team prepared and administered a survey on children between 5 and 13 years (target audience) to evaluate the usability, reliability, and response time of the DoItRight app. The design and development stages of the DoItRight app can be seen in Fig. 1.

B. Requirements: Increasing Children Awareness and Intention about Littering

In stage 1, preliminary studies and review of the existing literature indicate that kids are more likely to litter than adults. These children also have little awareness on the consequences of littering. Therefore, the target audience of this study is determined to be 5 to 13 years. In stage 2, we defined a survey to obtain user requirements of the target audience. The questionnaire was administered using Google Forms, distributed via WhatsApp and filled by parents and families of the target audience. In total, 161 responses were recorded.

- Demography of Participants: 87.6% of participants are female while 12.4% are male. 20.5% of them are less than 20 years, 51.6% are between 20-40 years, and 28% are over 40 years. Also, 46% of the participants have children with the age of 5 years.
- Results: 90.7% of participants reported facing the problem of littering in different public places, thereby confirming the significance of the littering problem that the DoItRight app aims to address. Also, 16.8% of the respondents reported that their children have bad littering behavior and throw wastes on the floor rather than the trash can, even at home. When asked about the causes of littering, 39.1% chose "lack of education", 54.5% chose "no responsibility" and 6.4% chose "others". This indicates that the lack of adequate awareness (education) contributes significantly to the menace of littering. In addition, 98.8% of participants are willing to allow their children use a mobile application that can increase their level of awareness and intent on littering. Also, 98.8% will appreciate a solution that is highly interactive, can improve the children's motivation towards dumping litters correctly and can help them apply whatever they learn.

C. DoItRight App Description

In stage 3, a prototype of the DoItRight app was developed while ensuring that there the features address the user requirements identified in stage 2. The mapping of the requirements to the features and the anticipated techniques to increase children's awareness and intention is provided in Table I. The DoItRight app is an education gaming mobile application targeting Saudi children between the age of 5 and 13 years. To access the application, a child needs to create an account which includes providing a username and creating a password of not less than 8 characters. The game consists of 3 levels; each level has specific tasks.

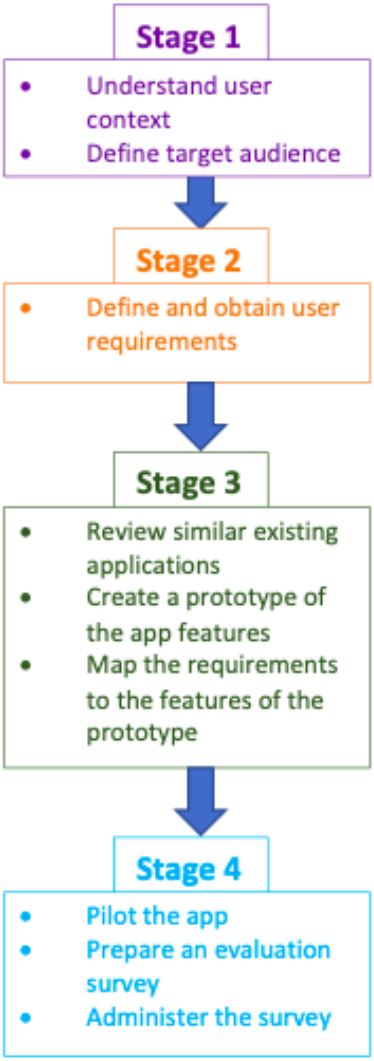


Fig. 1. Shows the Major Highlights of all the Four Stages in the Design and Development of the DoItRight App.

TABLE I. MAPPING USER REQUIREMENTS TO DOITRIGHT APP FEATURES AND AWARENESS/INTENTION TECHNIQUES

User requirements	App features	Increasing awareness and intention techniques
A means of raising awareness on littering	Levels 1 to 3 of the game, food items in the reward and evaluation stage of the game	Instruction on how to perform the behavior
Highly interactive	Use of visuals (videos and pictures), sounds, texts	Information provision
Motivation for children to drop litters appropriately	Food reward after level 3 (candy, banana, juice, and chips), stars, accumulation of points	Rewarding
Ability to apply what is learnt	Evaluation task on the empty candy bag, banana peel, empty can and empty chips bag on the floor	Feedback on the behavior; prompts
Understanding the importance of good littering behavior	Videos displaying positive and negative impacts of varying actions on the environment	Feedback on the behavior

1) First level (the beach)

- The game starts with a child carrying a bag in the beach.
- The child must collect 50% of the litter at the beach.
- In this first level, the child's speed will be medium.

2) Second level (the park)

- The second level starts with a child carrying a bag in a park.
- The child must collect 50% of the litter in the park.
- Here, the child's speed will be faster than the first level's speed.

3) Third level (the city streetss)

- The third level starts with a child carrying a bag in the city streets.
- The child must collect 50% of the litter in the streets.
- The child's speed will be faster than the speed in the second level.
- Throwing the dirtbag into the waste bin completes level 3 while the game proceeds to the reward and evaluation stage.

For completing all three levels successfully, the child is rewarded with food which includes a candy, banana, a can of juice, and a bag of crispy chips. While the food items serve as a reward for a job well done in levels 1 to 3, they are equally serving as a means to assess the child's level of awareness about littering having played the game. Therefore, after eating the food, the empty candy bag, banana peel, empty can, and empty chip bag are all thrown on the floor. Afterwards, a creative character will appear on the screen to ask the child a question; "what are you going to do now?" If the child picks up the litters and drop in the waste bin, he has passed the test. A congratulatory message will appear. However, if he did not drop the litters in the waste bin after 10 seconds, he will lose and repeat the game.

D. Key Highlights of the DoItRight App

1) Methods of collecting points: In each level, the child gets the maximum points obtainable if he collects the highest possible amount of litter that comes his way in the course of the game and successfully drops them all in the waste bin before the end of the level.

2) Difficulty: The child's walking speed will increase after each level, potentially allowing him to collect all the litters before reaching the waste bin.

3) Motivation: Aside from the food reward given to the child at the end of level 3, encouraging sound and objects will appear during the child's walk e.g., stars or clapping sound.

4) Videos at the end of each level: The first video congratulates the child who gets the required number of points and scales through the evaluation test. The video will show the positive impact of the child's act on the environment. Then, it also shows the negative effects of leaving litters on the floor.

The second video shows the child who did not get the required number of points or who did not pass the evaluation test. It shows the negative effects of the litter on the environment. Then, it ends with the positive impact of collecting litters and throwing them in the appropriate waste bin. This helps the child to visualize the positive change his efforts could have brought to the environment if he had dropped the litters in the waste bin.

5) Activity diagram: An activity diagram shows the flow of the interaction between a system and the user in a sequential order. The activity diagram for the DoItRight app is shown in Fig. 2.

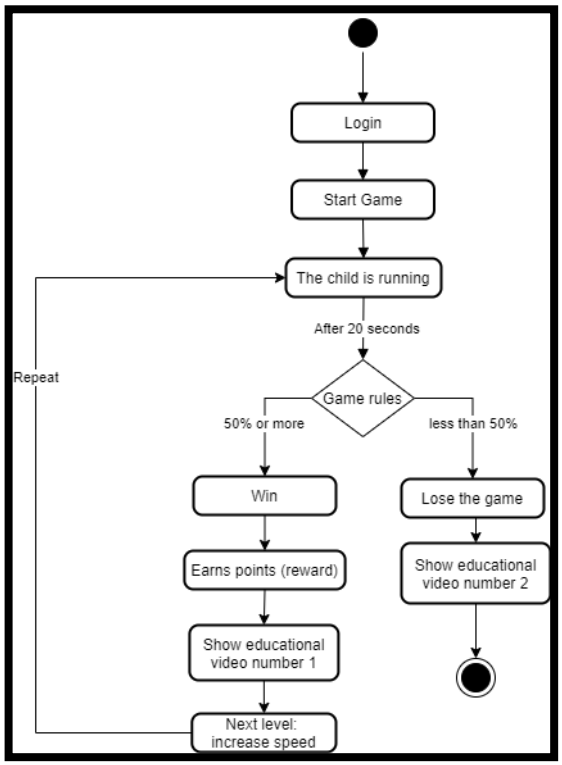


Fig. 2. The Activity Diagram of the DoItRight App Showing the Flow of Interaction between the System and the user.

6) User interface: The Graphics User Interface of the DoItRight app was designed using JustInMind. Some of the snapshots of the interface like welcome page, education video page, and game page are shown in Fig. 3-5.

Fig. 3. A user Interface that shows the Welcome Page of DoItRight App.

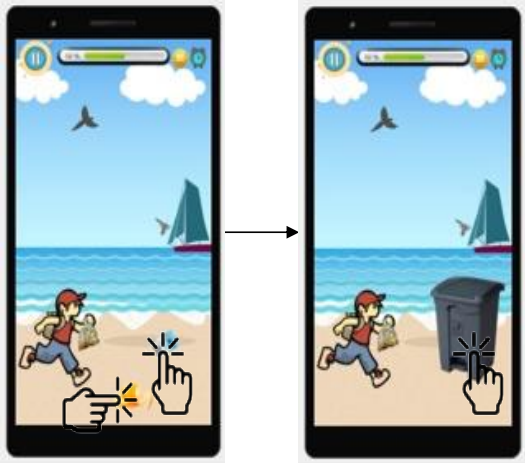


Fig. 4. Two user Interfaces with the First One showing 50% Progress made in the Level 1 of the Game and the Second Interface Showing a Child Throwing the Dirtbag into the Waste bin.

Fig. 5. The Interface Shows an Educational Video on the Positive Environmental Impact of Dropping Litters in the Waste Bin.

E. Testing and Evaluation

To test and evaluate (stage 4 of UCD) the effectiveness of the DoItRight app, three techniques were implemented namely unit testing, integration testing, and usability testing. These steps help to test the application for quality assurance, both technically and in terms of user satisfaction.

1) Unit testing: Unit testing is a software testing method by which each part of an application (known as a unit) is tested to ensure it is fit for use and meets the desired goal. Part of the objective is also to identify and solve any problems that may arise [20]. In the DoItRight app, we performed unit testing for several units of the app, such as the signup function and the login functions. In the login unit testing, for instance, we tested several cases like leaving the username and password fields blank, filling the password field alone, filling username field alone, and filling correct username and password in their respective fields.

2) Integration testing: While unit testing involves testing a part of the application, integration testing involves testing a combination of units. It is carried out to check if specific units of an application work and interact together as expected. It is usually carried out after the unit testing [21]. For the DoItRight app, we implemented seven different integrated testing objectives. The results are shown in section 4.0.

3) Usability testing: Here, we adopted the Arabic version of System Usability Scale (SUS) which is known as the most popular and widely used standardized questionnaire for assessing perceived usability [22]. SUS consists of a 10-item questionnaire; each question has five response options which include strongly agree, agree, neutral, disagree, and strongly disagree [23]. This evaluation was carried out to determine whether the DoItRight app meets the needs and expectations of the major stakeholders who are the target audience. The questionnaire was deployed using Google Forms. The ten questions include the following:

i. I think that I would like to use this system frequently.
ii. I found the system unnecessarily complex.
iii. I thought the system was easy to use.
iv. I think that I would need the support of a technical person to be able to use this system.
v. I found the various functions in this system were well integrated.
vi. I thought there was too much inconsistency in this system.
vii. I would imagine that most people would learn to use this system very quickly.
viii. I found the system very cumbersome to use.
ix. I felt very confident using the system.
x. I needed to learn a lot of things before I could get going with this system.

Ten children between the age of 5 and 13 were selected as participants in this survey. The full consent of the parents was secured before selecting a child or administering the questionnaire. As per the demography of study participants, we recorded 7 boys with age ranging from 5 to13 years and 3 girls

with age ranging from 8 to 13. Age 8, 11, and 13 recorded 2 children each while age 5, 6, 7, and 10 recorded one child each.

IV. Results and Discussion

The outcome of the integrated testing implemented on the DoItRight app is shown in Table II. As it can be seen in the table, all the test cases gave positive results.

After analyzing the responses recorded for the System Usability Scale questions, the following results were obtained:

i. 30% strongly agreed that they would like to use the DoItRight app frequently, 50% agreed while 20% were neutral.
ii. 80% of respondents strongly agreed that the system was easy to use while 20% agreed.
iii. 30% strongly agreed that they found the various functions in the system well integrated, 60% agreed and 10% were neutral.
iv. 70% strongly agreed that they would imagine that most people would learn to use the system very quickly; 30% agreed.
v. 60% strongly agreed that they felt confident using the system while 40% agreed.
vi. All respondents strongly disagreed that the system was unnecessarily complex or that there was too much inconsistency in it or that the system was too cumbersome to use or that they needed to learn a lot of things before they could get going with the system. They all strongly disagreed!
vii. 90% strongly disagreed that they would need the support of a technical person to be able to use this system while 10% agreed.

Also, based on the responses, we computed the SUS score for the DoItRight app. The SUS score of the app was found to be 93.25. This falls between 84.1 and 100 on the Sauro-Lewis grading scale for SUS scores [24]. Therefore, the DoItRight app falls into the grade A+ category and a percentile range of 96 to 100 on the Sauro-Lewis. As a result, it can be concluded that the DoItRight app is highly acceptable among users.

There are some existing apps that have some similarities with the DoItRight app. Such applications include Watten Games, Garbage Truck, Trash Stash and Littering Game. A comparative analysis between the DoItRight app and the four identified similar apps revealed that the DoItRight app offers better and superior features to address the challenge of littering than any of the other four applications. It supports Arabic language, has several levels and scenarios, rewards players with points, and provides realistic images of characters. This puts the DoItRight application in a position to change the behavior and intent of children towards littering in Saudi Arabia, much better than any other application.

TABLE II. Results from the Integrated Testing Carried Out on Doitright App

Test Case ID	Test Case Objectives	Test Case Description	Expected Results	Remark
1	To check the interface link between the first page and the signup page.	The user clicks the " أنشأ حسابك الان" button.	DoItRight directs the user to the signup page.	Positive test
2	To check the interface link between the first page and the login page.	The user clicks the "سجل دخولك" button.	DoItRight directs the user to the login page.	Positive test
3	To check the interface link between the signup page and the start page.	The user enters name and password, confirms it, and clicks the "أنشأ حسابك" button.	DoItRight application creates a user account and directs the user to the start page.	Positive test
4	To check the interface link between the login page and the start page	The user enters name and password and clicks the " سجل دخولك" button.	DoItRight application allows the user to access the app and directs user to the start page.	Positive test
5	To check the interface link between the start page and the first page	The user clicks the " تسجيل الخروج" button.	DoItRight directs the user to the first page.	Positive test
6	To check the interface link between the start page and game page	The user clicks the "ابدأ اللعب" button.	DoItRight directs the user to the game page.	Positive test
7	To check the interface link between the stop page and the start page	The user clicks the "انهاء اللعب" button.	DoItRight directs the user to the start page.	Positive test

V. Conclusion and Future Work

To the best of our knowledge, the DoItRight app is the first mobile application aimed at increasing awareness level and intention of Saudi children about littering. The app features meet all user requirements identified and defined during the UCD process such as the potential to educate a child on the proper way to dump litters, motivation, and the application of knowledge gained in real-life. A survey conducted to determine the acceptability of the DoItRight shows that the app has a high acceptability rate among the target audience having recorded a System Usability Scale score of 93.25%.

In the future, we would like to explore the following areas:

i. Conduct a more comprehensive evaluation of the DoItRight app by getting feedback from more children. To get this done, we intend to get approval from the Institutional Review Board (IRB) in Saudi Arabia.
ii. Add more levels to the DoItRight game with varying degrees of difficulty and duration.
iii. Connect the application with the Internet and GPS such that the stages can appear on the map and give a child the feeling of cleaning his current living environment.

iv. Create a dashboard where users can track their performance with respect to other players using the DoItRight app in their areas.
v. Add multiplayer stages to allow users compete with nearby players and friends.

The DoItRight app can have a significant positive impact on Saudi's public environment if widely adopted among children across the country.

REFERENCES

[1] Milfont, T. L., & Schultz, P. W. (2015). Culture and the Natural Environment. Current Opinion in Psychology.

[2] Veiga, J. M., Vlachogianni, T., Pahl, S., Thompson, R. C., Kopke, K., Doyle, T. K., . . . Alampei, I. (2016). Enhancing public awareness and promoting co-responsibility for marine litter in Europe: The challenge of MARLISCO. Marine Pollution Bulletin, 102(2), 309-315. doi:10.1016/j.marpolbul.2016.01.031.

[3] Weaver, R. (2015). Littering in context (s): Using a quasi-natural experiment to explore geographic influences on antisocial behavior. Applied Geography, 57, 142-153.

[4] Yara A. (2017). Extending Understanding of Middle Eastern Littering Behaviour Beyond the Individual: A Formative Research. Griffith Business School. https://doi.org/10.25904/1912/1998.

[5] Yara Almosa, Joy Parkinson & Sharyn Rundle-Thiele (2020) Preventing Littering: It's Not All about Sticks!, Journal of Nonprofit & Public Sector Marketing, DOI: 10.1080/10495142.2020.1865236.

[6] Zapata, B., Fernández-Alemán, J., Idri, A. *et al.* (2015). Empirical Studies on Usability of mHealth Apps: A Systematic Literature Review. *J Med Syst* **39,** 1. https://doi.org/10.1007/s10916-014-0182-2.

[7] Yanghee Kim & Diantha Smith (2017) Pedagogical and technological augmentation of mobile learning for young children interactive learning environments, Interactive Learning Environments, 25:1, 4-16, DOI: 10.1080/10494820.2015.1087411.

[8] Still, B., & Crane, K. (2017). Fundamentals of user-centered design: A practical approach. CRC press.

[9] Dopp, A., Parisi, K., Munson, S., and Lyon, A. (2019). A glossary of user-centered design strategies for implementation experts. *Translational behavioral medicine*, *9*(6), 1057-1064.

[10] Mahasneh, R., Romanowski, M. and Basem, R. (2017). Reading social stories in the community: A promising intervention for promoting children's environmental knowledge and behavior in Jordan, The Journal of Environmental Education, 48:5, 334-346, DOI: 10.1080/00958964 .2017.1319789.

[11] Sağlam, M. (2016). Exploring fifth-grade Turkish children's solutions and future plans for environmental pollution through their drawings. Asia-Pacific Forum on Science Learning and Teaching. 17.

[12] Herdiansyah, H, Brotosusilo, A., Negoro, H., Sari, R., and Zakianis, Z. (2021). "Parental Education and Good Child Habits to Encourage Sustainable Littering Behavior" *Sustainability* 13, no. 15: 8645. https://doi.org/10.3390/su13158645.

[13] Moqbel, S., El-tah, Z., and Haddad, A. (2020). Anti-littering in developing countries: Motivating the people of Jordan. Waste Management & Research, 38(7), 726–733. https://doi.org/10.1177/ 0734242X19900654.

[14] Panagiotopoulou, L., Cía Gayarre, N., Scurati, G. W., Etzi, R., Massetti, G., Gallace, A., and Ferrise, F. (2021). "Design of a Serious Game for Children to Raise Awareness on Plastic Pollution and Promoting Pro-Environmental Behaviors." ASME. *J. Comput. Inf. Sci. Eng*. December 2021; 21(6): 064502. https://doi.org/10.1115/1.4050291.

[15] Rahmayanti, H, Oktaviani, V. and Syani, Y. (2020). Development of sorting waste game android based for early childhood in environmental education. Journal of Physics: Conference Series. 1434. 012029. 10.1088/1742-6596/1434/1/012029.

[16] Kacoroski, J., Liddicoat, K., and Kerlin. S. (2016) Children's use of iPads in outdoor environmental education programs, Applied Environmental Education & Communication, 15:4, 301-311, DOI: 10.1080/1533015X.2016.1237903.

[17] Safitri, D. and Ika, L., Maksum, A., Nurzengky, I., Marini, A., Zahari, M. and Iskandar, R. (2021). Web-Based Animation Video for Student Environmental Education at Elementary Schools. International Journal of Interactive Mobile Technologies (iJIM). 15. 66. 10.3991/ijim.v1 5i11.22023.

[18] Buchanan, J., Pressick-Kilborn, K., and Maher, D. (2019). Promoting Environmental Education for Primary School-aged Students Using Digital Technologies. *Eurasia Journal of Mathematics, Science and Technology Education*, *15*(2), em1661. https://doi.org/10.29333/ ejmste/100639.

[19] Kalogiannakis, M. and Papadakis, S. (2017). Combining mobile technologies in environmental education: a Greek case study. International Journal of Mobile Learning and Organisation. 11. 108-130. 10.1504/IJMLO.2017.10005249.

[20] Yu, J., Zhang, J., Chen, Y., Wu, N., Mei, Y., Zhang, D., ... & Sheng, Y. (2021). A Test Method for Instructional Software of Evaluation and Exercise Based on Mobile Platform. In *Journal of Physics: Conference Series*. IOP Publishing.

[21] Hendradjaya, B. (2018). A Proposal for New Software Testing Technique for Component Based Software System. *International Journal on Electrical Engineering & Informatics*, *10*(1).

[22] James R. Lewis (2018) The System Usability Scale: Past, Present, and Future, International Journal of Human–Computer Interaction, 34:7, 577-590, DOI: 10.1080/10447318.2018.1455307.

[23] Kaya, A., Ozturk, R., & Gumussoy, C. A. (2019). Usability measurement of mobile applications with system usability scale (SUS). In *Industrial engineering in the big data era* (pp. 389-400). Springer, Cham.

[24] Lewis, J. and Sauro, J. (2019). Item Benchmarks for the System Usability Scale. Journal of Usability Studies.

Noise Cancellation in Computed Tomography Images through Adaptive Multi-Stage Noise Removal Paradigm

Jenita Subash, Dr.Kalaivani S*
School of Electronics Engineering (SENSE)
Vellore Institute of Technology, Vellore, Tamil Nadu, India

***Abstract*—Image de-noising is a noise removal approach, which is utilized to remove noise from the noisy image and is utilized to protect the significant features of images namely, corners, edges, textures, and sharp structures. For medical diagnosis Computer tomography (CT) images are mainly utilized. Due to acquisition and transmission in CT imaging, the noise that appears leads to poor image quality. To overcome this problem, an efficient Noise cancellation in computed tomography images using adaptive multi-stage noise removal paradigm is proposed. The proposed approach consists of three phases namely, Optimal Discrete Wavelet Transform, first stage noise removal using Block Matching, and 3D filtering (BM3D) filter and second stage noise removal using the bilateral filter (BF). Initially, Discrete Wavelet Transform (DWT) is applied to the input image to diminish noise in CT images. In this method, co-efficient ranges are optimally selected with the help of Crow Search Optimization (CSO) algorithm. Secondly, to remove the noise present in the bands, BM3D algorithm is applied. Finally, bilateral filter is applied to the BM3D output image to further enhance the image. The performance of the proposed methodology is analyzed in terms of Peak signal-to-noise ratio (PSNR), Root Mean Square Error (RMSE), and Structural Similarity Index (SSIM). Furthermore, the multi-stage noise removal model obtained gives the best PSNR values compared to other techniques.**

Keywords*—*De-noising; computer tomography; discrete wavelet transform; crow search optimization; bilateral filter

I. Introduction

An image is a collection of dimensions in two dimensional (2-D) or three dimensional (3-D) spaces [1]. Computer tomography (CT) images usually have noise due to faults in image holding methods. Noise will be removed from images so that the analysis of image elements (e.g., blood vessels, inner folding, or tumors in the human brain) can be completely observed and the upcoming image researches are trustworthy. The image restoration presented appears to be the sharpest possible among the multi-scale image smoothing methods by preserving uniqueness and stability [2]. The medical imaging technology is fetching a valuable section of a huge amount of purpose namely research, diagnosis, and treatment. It has enabled doctors to construct images of patient body for medical objectives [3]. Basically medical images namely X-Ray, CT, MRI, and PET contain the information of

Heart, brain, and nerves but these images are suffered from huge shortcomings, which include the acquisition of noise [4].

The noise is irregular fluctuations that accompany a transmitted signal that tend to obscure the signal that has to make the data to slow down or reduce the clarity or accuracy of the data. Medical images may be clear, sharp, noisy and vague. Usually computed tomography (CT) images are distorted by Gaussian noise and salt and pepper noise [5]. The Gaussian noise, which increases due to acquisition and it can be reduced by using spatial filters. Salt and pepper noise, which rarely occurs in form of white and black pixels, can be effectively eliminated by the morphological filter. Two approaches Empirical Mode Decomposition (EMD) and Dual-Tree Complex Wavelet Packets (DTCWP) are used for de-noising the CT-images. All noisy algorithms are based on the local or global noise model and the generalized image softness model [6]. In modern hospitals, X-RAY and CT images are mostly used because these have several importance, but it may lead to potential radiation hazard to patient because x-rays could cause hereditary harm and actuate malignant growth in likelihood identified with radiation portion [7].

A lot of filtering methods for example median filter, mean filter, bilateral filter, Gaussian filter, linear filters, non-linear filters, spatial filters and transform domain filters are used for remove the noise present in the CT images. Moreover, edge-preserving approaches are utilized for reducing undesirable effects on images [18]. In [9] nonlocal means filtering based CT image de-noising is explained in [8] 3D collaborative filtering based de-noising. A lot of methods are available for de-noising even though an efficient de-noising method is urgently needed.

The important goal of this paper is to eliminate noise existing in CT image with the help of multi-stage noise removal paradigm model. The proposed multi-stage noise removal paradigm model consists of two stages of the noise removal process. The first stage noise removal is done with the help of the BM3D filtering algorithm and second stage noise removal is done with the help of a bilateral filter. These two stages have improved the quality of the image. The contribution of the research work is listed below:

- DWT is applied to the input image to convert the spatial domain image into a transform domain. In this, coefficient ranges are optimally selected with the help of CSO algorithm.

*Corresponding Author.

- To evacuate the noise present in the input image BM3D channel is connected to the input image. The BM3D channel accomplishes the dependable PSNR and resolves difficulties of related methodologies when tending to the distinctive level of noise.
- To further progress the input image quality, a bilateral filter is accomplished to output of the BM3D image.
- The performance of proposed technology is scrutinized in conditions of various metrics and performance is in comparison with a different algorithm.

The remainder paper sorted out as pursues; the background of the proposed method is analyzed in Section II and the proposed image de-noising methodology is analyzed in Section III. The performance analysis is discussed in Section IV. Finally the article is concluded in Section V.

II. Background

A lot of researchers had elucidated the image de-noising technique. Among them some of the methods are analyzed here; Elhoseny et al. [10] had proposed a medical image de-noising using optimal bilateral filter (OBF) and convolution neural network (CNN). For the noise removal process optimized bilateral filter (OBF) is utilized. In this filter, Gaussian and spatial weights are the parameter used in the OBF. To increase the characteristic of the de-noised image, the parameters are excellently selected with the help of a combination of dragonfly (DF) and modified fruit fly algorithm. After the filtering process, the normal and abnormal images are classified using CNN classifier. Manduca et al. [11] proposed a novel locally adaptive projection space denoising algorithm for a low dose CT image. Similarly, Katsuhiro et al. [12] have developed an edge preserving based noise reduction using three-dimensional cross-directional bilateral filter (3D-CDBF) in CT images. The filtering process is mainly used for noise removal and edge preserving process. The bilateral filter is a mixture of two types of filters namely spatial and Gaussian filter. Finally, the noise spectrum is calculated for all the de-noised image and performance are analyzed.

In [13], Wojciech and Ewa have explained a medical image noise cancellation and edge preserving based on a granular filter. Here, two different methods namely, crisp and fuzzy are developed. For experimentation, CT and US breast images are utilized. The granular filter performance is compared with different filters namely, relating to space balancing and median, bilateral filter, anisotropic diffusion. Moreover, Hsuan and Chieh [14] have explained a kernel-based image de-noising technique for developing parametric image creation. To eliminate the noise in input image, general-threshold filtering method is combined with a whole variation and this method was investigated. The mathematical explanation of improved intravoxel incoherent motion (IVIM) method based de-noising is proposed. The suggested method was effective than IVIN method.

Manoj and Pardeep, [15] have explained a CT image de-noising using the bilateral method with the concept of Bayes Shrinkage rule in the wavelet domain. Initially, the image is filtered with the help of the bilateral filter. After the noise removal process, wavelet packet based thresholding is applied. Then, to attain the efficient de-noised image, the threshold output image is added with the bilateral filter. The performance of the presented technology is analyzed in conditions of the PSNR and similarity measures. Similarly, in [16], Manoj et al., have explained CT image de-noising in the curvelet domain. In high frequency coefficients, inter- and intra-scale responsibilities are used in side by side. From the high frequency coefficients, correlation values are obtained. Then, both the high frequency coefficients, aggregation are performed. After aggregation, the inverse curvelet transform is applied to get a de-noised image. Moreover, Bing et al. [17] have explained a Coupling de-noising methods depends on individual wavelet transform and modified median filter for medical image. The method consists of four phases namely, image acquisition, image storage, image processing, and image reconstruction. Initially, the image is captured from the patient that contained the noise. Then the collected images are stored on the cloud. In the third phase, the medical image is breakdown into four modules namely, LL, LH, HL, and HH. Then, for further processing, high frequency co-efficient are utilized. Then the changed median filter is applied to three high frequency sub bands. Finally, they have obtained the de-noised image. The performance is analyzed in terms of PSNR measures. Jenita Subash et. al. [19], Shyna.A et.al.[20], Devinder Singh et. al. [21] introduced improved fuzzy based approach for noise removal.

III. Adaptive Multistage Noise Removal Methodology

In this work, a novel multi-stage noise removal paradigm to deal with noise is proposed. The overall architecture of proposed method is shown in Fig. 1. For de-noising process, at first, images are decomposed with the help of DWT. To enhance the sensitive regions with higher visual quality, initially, DWT is employed to input image; while optimal coefficients are selected using the Crow Search Optimization algorithm. After decomposition, BM3D filtering algorithm is applied to high frequency sub bands of DWT output. Then at the subsequent stage, the bilateral filter is used to take out the noise cleanly and it retains the uncorrupted information. The overall concept is compressed into three phases:

A. Crow Search Optimization Algorithm

The most important goal of this section is to segregate the input image into four sub divisions, that is, LL, LH, HL, and HH. To increase the image quality in terms of PSNR, this methodology optimally selects the wavelet coefficient. For wavelet co-efficient optimization, CSO algorithm is utilized. CSO is a recently developed metaheuristic algorithm and also developed based on crow's behaviour.

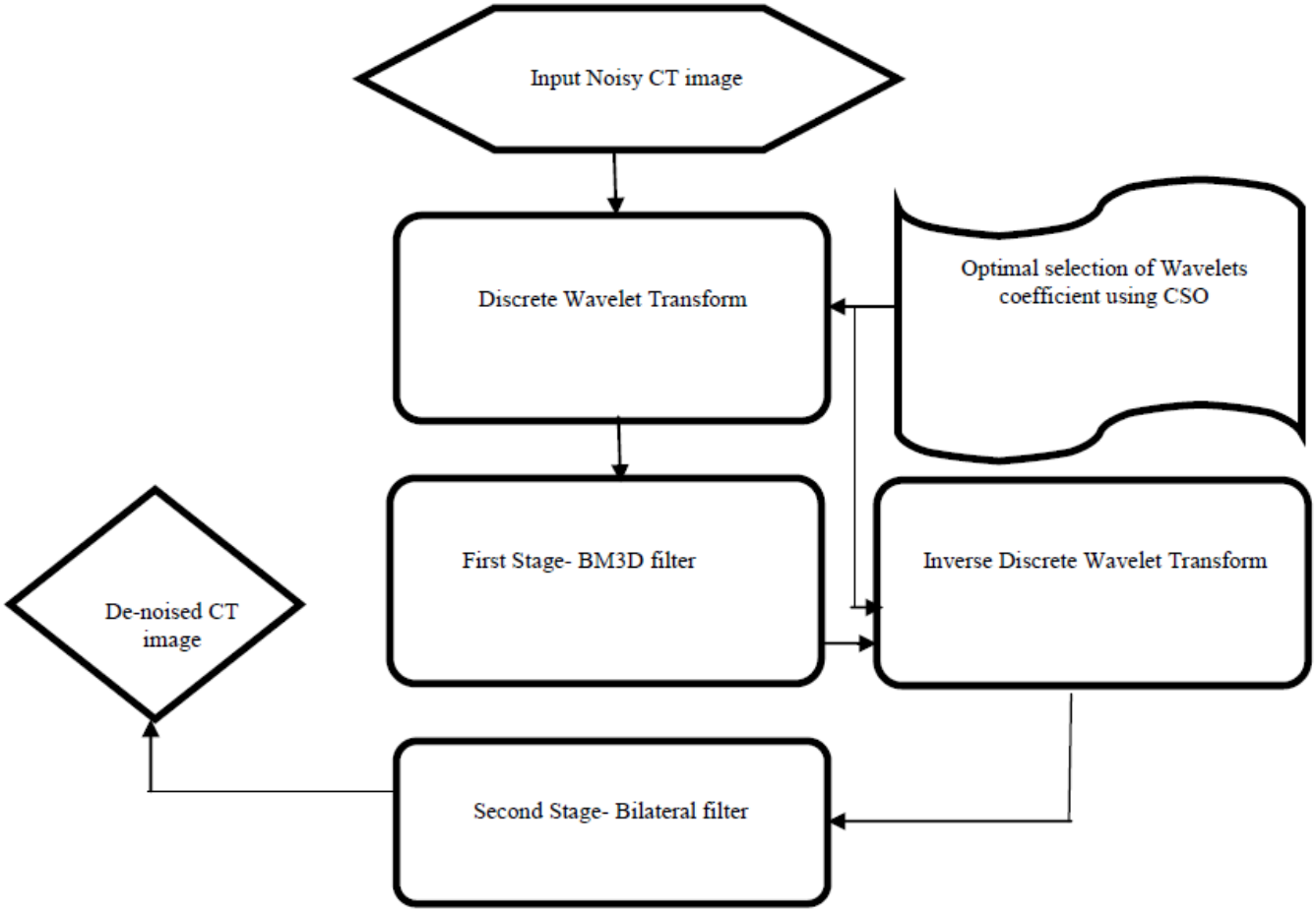


Fig. 1. Overall Architecture of Proposed Method.

B. *First Stage Noise Removal using BM3D Filtering Algorithm*

After image de-composition, BM3D procedure is used to eliminate the noisy contents present in image. BM3D filtering algorithm is mainly applied to LH, HL and HH frequency subdivision bands of DWT image outputs. The BM3D algorithm is divided into two fundamental step ladders. In the initial stage the main focus is on producing essential image estimation and it's widely less noise than the noisy image. In the second stage the fundamental estimate is utilized as a block matching base for pragmatic wiener filtering. The second step could be empirically confirmed to get better quality of an image compared with the initial stage of output.

The principal step is termed as basic estimation, which emphases on eliminating noise present in the image. This step consists of three phases namely, Block-matching (BM), Collaborative filtering (CF), and Aggregation:

1) Block Matching (BM): Initially, the image I is converted into a number of blocks size of $a \times a$. Consider the reference block A. Blocks have been a high similarity with the reference block is formed as a group. Then, the blocks are converted into the 3D array. The similarity of the reference block A and other block O are calculated using equation distance function which is given in equation (1).

$$S(A_i, O_j) = \frac{\|A_i - O_j\|}{N_1^2} \quad (1)$$

Where A_i represent the 2D array of reference block A, O_j represent the 2D array of similar blocks. After the similarity calculation, blocks are grouped.

2) Collaborative filtering: At present, the most equivalent blocks of reference blocks are recognized and might be gathered to frame a 3D Array S_1 size is $a \times a \times |S_1|$ .After that, every 3D array is converted to a frequency domain using a 1D and 2D intra-block transform. From transferred data the valuable data is placed in the utmost huge coefficients. Therefore it is conceivable towards decreasing the noise through disposing of little coefficients. The procedure of CF can be represented as:

$$C_F = K_{3D}^{-1}\left(R\left(K_{3D}(S_1)\right)\right) \quad (2)$$

Where C_F denotes the collaborative fitter output, K_{3D} is the 3D linear transform is usually actualized by isolated 2D

and 1D linear transformation, R represent the hard thresholding operator, which is used to regulate the transform coefficient. The threshold is calculated using equation 3.

$$R(x)=\begin{cases}0, & |x|\le\sigma.\chi_{3D}\\ x, & |x|>\sigma.\chi_{3D}\end{cases} \quad (3)$$

Where

χ_{3D} →Hard-thresholding parameter

σ →Standard deviation

Meanwhile, noise is found in the little coefficients and it is feasible to lessen noise and safeguard best subtleties in the images over the hard-thresholding administrator. Simultaneously CF understands the separation of image signal and noise deprived of expending energy.

3) Aggregation: After filtering, overlapped blocks of the image are to be recovered to their unique positions. Then weighted average (WA) measure based images are estimated. In common, the WA is apprehended by conveying suitable weighting components to the groups of blocks and the weight W_ρ can be displayed as:

$$W_\rho=\begin{cases}\dfrac{1}{N_c}, & N_c\ge 1\\ x, & N_c<1\end{cases} \quad (4)$$

Where

N_c →Quantity of the persisted non-zero coefficient

We obtain the small number of non-zero co-efficient implies the number of noise detached using collaborative filter since its weight of the block is superior. The basic estimation of the de-noised block D_{basic} is given in equation (5).

$$D_{basic}(i)=\frac{\Sigma_{Op}\,\Sigma_{Q\in O_p}W_i.R_{PQ}}{\Sigma_{Op}\,\Sigma_{Q\in O_p}W_i.x_Q}\quad \forall i\in I \quad (5)$$

Where

Q → Similar block of the present operational block comprises the pixel i

O_p →Set of entirely available blocks

R_{pQ} → Estimation of the block Q

$$R_{pQ}=\begin{cases}R_{PQ}, & i\in Q\\ 0, & i\notin Q\end{cases} \quad (6)$$

The characteristic function x_Q is symbolized as follows;

$$x_Q=\begin{cases}1, & i\in Q\\ 0, & i\notin Q\end{cases} \quad (7)$$

*4) Final estimation:*In this stage, final estimation is done with the help of a wiener filter which is used to improve de-noising performance. Initially, the basic estimation output is given to the input of the block BM. In the BM, two groups arrive which is from a noisy image (TP1) and other one from the basic Estimation (TP2). The attained basic estimation output is the same as the actual image. The final weight is calculated using equation (8).

$$W_{final}=\frac{|K_{3D}(T_{P2})|^2}{|K_{3D}(T_{P2})|^2+\sigma^2} \quad (8)$$

The final estimation F_{final} is attained using equation (9).

$$E_{final}(i)=\frac{\Sigma_{Op}\,\Sigma_{Q\in Op}W_{final}.R_{PQ}}{\Sigma_{Op}\Sigma_{Q\in Op}W_{final}.x_Q}\quad \forall i\in I \quad (9)$$

Compared with the mutual modest hard-thresholding process, Wiener filtering is more effective and the outcome is more precise.

C. Second Level Noise Removal using the Bilateral Filter

After the initial stage of noise removal, bilateral filter (BF) is applied to the first stage output image to improve the image quality. The BF has established its potential for image de-noises with edge protection related to additional spatial domain filtering. It is a very simple and non-iterative filter. BF is based on domain filtering as well as range filtering.

Let f(x) be the input image the low-pass domain filter is applied to the input as well as the output is given by the condition 10-13,

$$f(z)=\ k_d^{-1}(z)\iint f\ (\alpha).e(\alpha,z)d\alpha \quad (10)$$

where, e (α, z) = geometric distance among centre x also close by point α .likewise,range filtering is given in the subsequent equation 11 as,

$$\text{h (z) } k_r^{-1}(z)\iint f(\alpha).s\big(f(\alpha)f(z)\big)d\alpha \quad (11)$$

Whereas s(f (α)f(z)) = “photometric distance” in the middle of center x and close by points ε.

While the bilateral filter is a mixture of both domains filter as well as range filtering, its production could be characterized as (12).

$$h(z)=\ k^{-1}(z)\iint f(\alpha).e(\alpha,z).s\big(f(\alpha)f(z)\big)d\alpha \quad (12)$$

At this point, k (z) = normalization constant is defined by the subsequent equation,

$$K(z) = \iint e(\alpha, z)s\big(f(\alpha)f(\alpha)f(z)\big)d\alpha \quad (13)$$

Bilateral filters are mostly utilized to remove noise exactly or precisely. Here first we prefer the window size then we pre-compute the distance weight. Later the bilateral filter is applied to remove noise accurately that is the first step in local region extraction and the next step is intensity value calculation and then the subsequent step calculates the bilateral filter response. Later the de-noised image is obtained.

IV. SIMULATION RESULT

The performance of proposed image de-noising method is analyzed in this section. For experimentation CT images are utilized. The adaptive multi-stage noise removal methodology is implemented in the platform of MATLAB.

A. Data Set Description

The CT lungs images are efficiently occupied in the innovative image segmentation and classification technique that is attained from therapeutic facility just as web sources. The corresponding gathered image dataset have 1000 CT lungs images. Here 750 lungs images acquired from TNMSC kudangulam CT scan center, Tamil Nadu and Marthandam MRI and CT scan center, Tamil Nadu. The remaining images are collected from web resources. Datasets are collected during June 2018- July 2019. It has CT images of females, males and an infant. Fig. 2, 3, and 4 provide some input images.

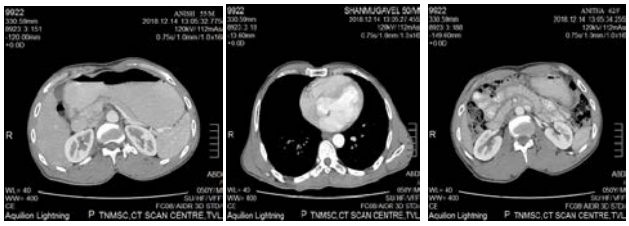


Fig. 2. Sample Images Collected from TNMSC CT Scan Center Tamil Nadu.

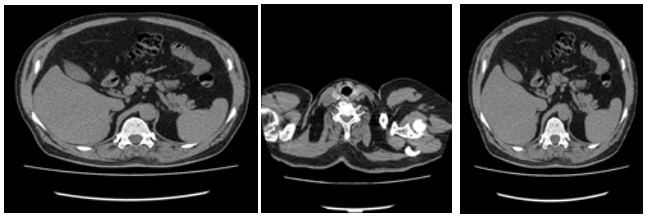


Fig. 3. Sample Images Collected from Marthandam MRI and CT Scan Center Tamil Nadu.

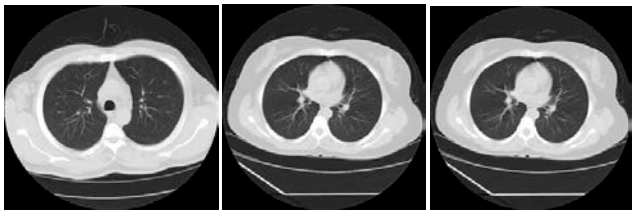


Fig. 4. Experimentally used Sample Images Collected from a Web Source.

B. Evaluation Metrics

The performance of adaptive multi-stage noise removal method is analyzed in terms of various metrics namely peak signal for the noise ratio (PSNR) and Root-Mean-Square error (RMSE), and structural similarity index are explained as below:

1) PSNR: This measure is utilized to measure de-noised image quality. The PSNR is the ratio among the input image and the noise image. Higher PSNR value is given a good quality image.

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \quad (14)$$

$$MSE = \frac{1}{M * N}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[I(i,j) - I'(i,j)\right]^2 \quad (15)$$

Anywhere;

$I(x,y)$ → Input image.

$I'(x,y)$ → De-noised image.

2) RMSE: RMSE minimizes the error rates. It serves to summative the magnitudes of the errors in predictions for a variety of times into a solitary measure of predictive power. RMSD is the square root of the mean of the square error. *RMSE Formula* is given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (16)$$

3) SSIM: The structural similarity index which is used to measure the comparison among several images. It is a sensitivity based model, that considers the image,

$$\text{SSM}(x,y) = \frac{(2\mu_x 2\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1)(\mu_x^2 + \sigma_y^2 + c_2)} \quad (17)$$

Where,

μ_x →Specifies the middling of x,

μ_y →Specifies the middling of y,

μ_x^2 →Specifies the variation of x, σ_y^2which specifies the variation of y

Experimental results are attained from the adaptive multi-stage noise removal methodology. The following Fig. 5 to Fig. 7 displays the comparative outcomes of existing methodology as well as adaptive multi-stage noise removal methodology for de-noising CT medicinal images.

The principal goal of this article is to perform image de-noising utilizing a blend of BM3M and bilateral filter. Initially the images are decomposed using DWT. To improve the final de-noising performance the coefficient range of DWT is optimally selected with the help of Crow Search Optimisation Algorithm.

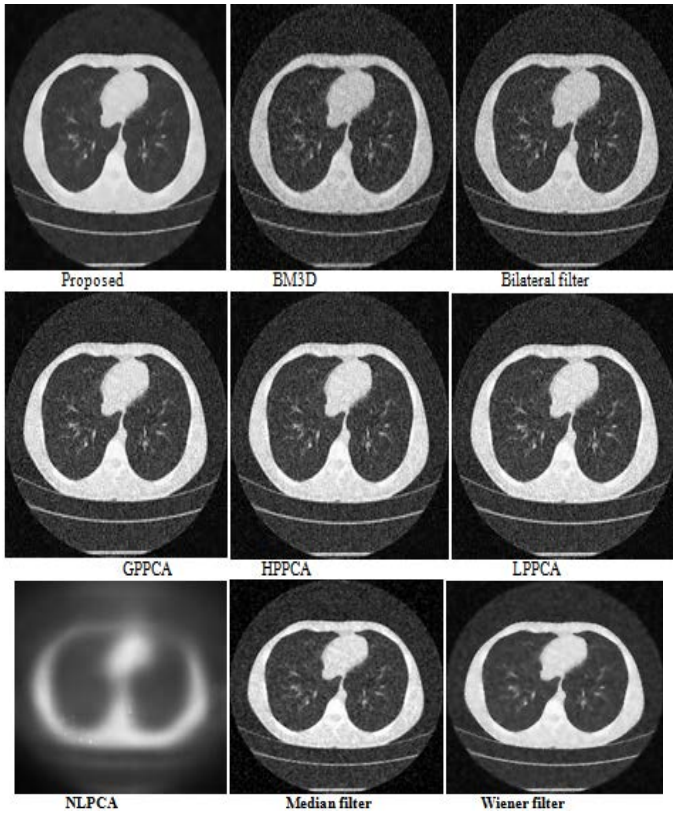


Fig. 5. De-noising Result of Images Collected from Internet Source.

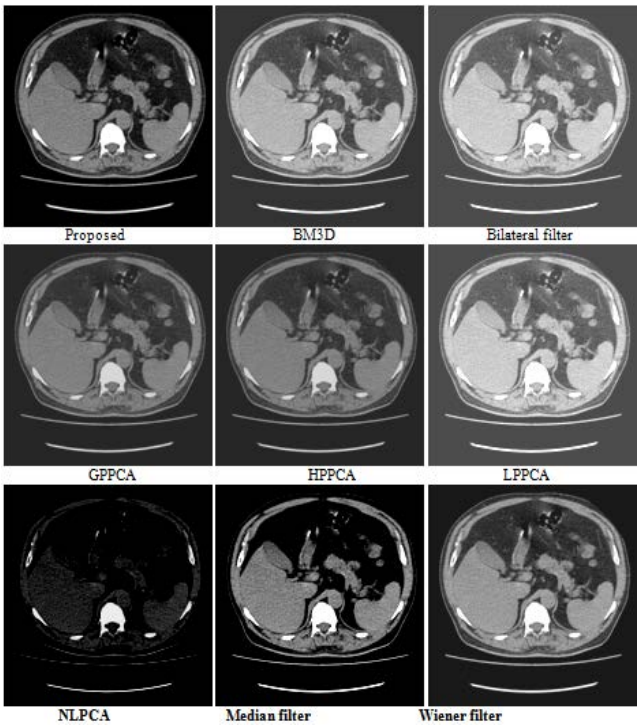


Fig. 6. De-noising Result of Images Collected from Marthandam MRI and CT Scan Center.

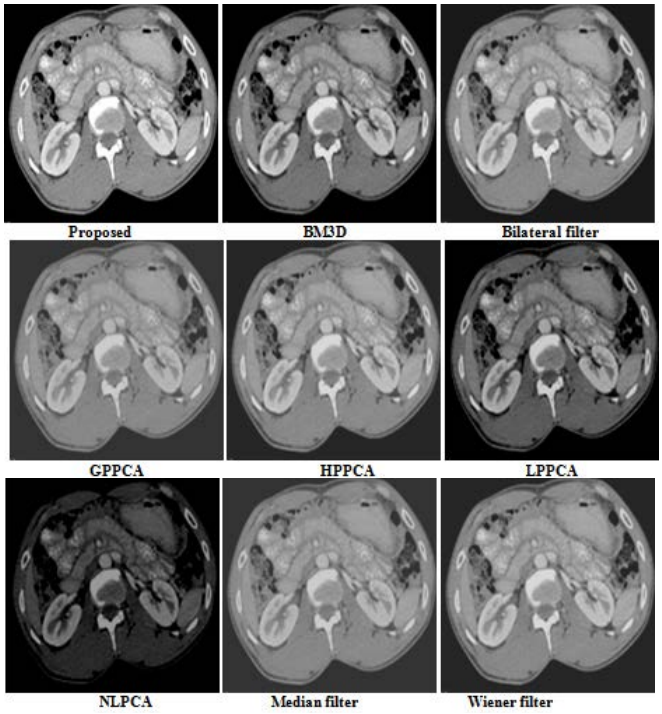


Fig. 7. De-noising Result of Images Collected from the TNMSC CT Scan Center.

Once the optimal coefficients are selected, the BM3D algorithm is applied to the LH, HL and HH frequency bands of DWT image output, and then the bilateral filter is applied to eliminate the noise clearly and to retain the uncorrupted information. In this section, the above Fig. 5 to Fig. 7 shows the comparison results of existing and proposed de-noising CT medicinal images. To evaluate the performance of our adaptive multi-stage noise removal methodology approach, we examine our results with other existing approaches. In this we compare our work with the following existing techniques which is named as BM3D, Bilateral filter, Global Patch Principle Component Analysis (GPPCA), Hierarchical Patch Principle Component Analysis (HPPCA), Local Patch Principle Component Analysis (LPPCA), Non-Linear Principal Component Analysis (NLPCA), Median filter and Wiener filter based image de-noising approaches. Compared to the existing approach our proposed adaptive multi-stage noise removal methodology achieve a better result because of our proposed strategy using the de-noising process in two steps applying BM3D filter the noise is almost removed clearly in the first step and for more exactness and precision bilateral filter is applied, which removes the noise clean and clearly in CT medical images.

Fig. 8 demonstrates the performance analysis of the proposed method using PSNR measure. The maximum PSNR is considered as a good quality image. To demonstrate the efficiency of proposed adaptive multi-stage noise removal methodology based de-noising approach, we compare our algorithm with different algorithm namely BM3D, Bilateral filter, Global Patch Principle Component Analysis (GPPCA), Hierarchical Patch Principle Component Analysis (HPPCA), Local Patch Principle Component Analysis (LPPCA), Non-

Linear Principal Component Analysis (NLPCA), Median filter and Wiener filter based image de-noising.

When analyzing Fig. 8, our algorithm attains the average PSNR of 42.25 db, which is 41.25 db, 41.25, 21.53, 22.32db, 32.025 db, 25.50db, 25.47db and 25.44 db for using BMSD, Bilateral filter, Wiener filter, median filter, NLPCA, GPPCA, HPPCA, and LPPCA respectively. From the PSNR value, we precisely realize our proposed methodology is superior to existing methodology because our proposed scheme prefer optimal co-efficient at the same time it uses filters in two stages in 1st stage BM3D filters are used to remove noise among several CT images. Applying BM3D filter the noise is almost removed clearly in the first step and for more exactness and precision we use bilateral filter to remove noise clearly in CT medical images. Fig. 9 shows a comparative analysis based on RMSE measures. Of these, minimum value of RMSE gives better results of de-noising because the quality of a resultant image is being measured by using RMSE. Comparing to the existing techniques, RMSE measure minimize the error rates of our proposed methodology. When examining Fig. 9 our proposed adaptive multi-stage noise removal methodology achieves the minimum RMSE of 0.7157348629. Here the proposed approach attains maximum SSIM of 0.9110654649, which is highly compared to other existing algorithms. In this, the maximum value of SSIM measure gives better results because SSIM measures the similarity among several images when examining Fig. 10 our proposed adaptive multi-stage noise removal methodology accomplish the utmost SSIM measure of 0.9110654649. From the experimental outcomes, we obviously understand our proposed methodologically gives out enhanced results compared to other existing approaches and the performance comparison is shown in Fig. 11 using mean of PSNR(dB) values. Fig. 12 shows the intensity variation of noise and de-noised CT image. Table I shows the performance of proposed approach by altering noise level. Here, the performance is analyzed based on three level noises like as 0.02, 0.04 and 0.06. When analyzing Table I, after applying noise also our proposed approach attains the excellent PSNR value. This is because of two level filtering approaches. As a result, it is clear to us that our proposed method produced an excellent result compared to other methods. To prove the effectiveness of the proposed methodology, we compare our algorithm with different methods as shown in Table II. In this performance analyze, we compare our proposed method with already published literatures like Dual Tree Complex Wavelet (DTCWT) [22], Curvelet Transform (CT) [22], Harris and DWT [23], Harris Operator and Wavelet Domain Thresholding (RDTDWT) [24], SRTW [25]. When analyzing the above table our proposed method achieves a higher accuracy and higher PSNR of 42.05 because in our work multilevel denoising is performed as well as adaptive bilateral filter is used. Comparing these existing techniques our proposed method achieves a high quality results.

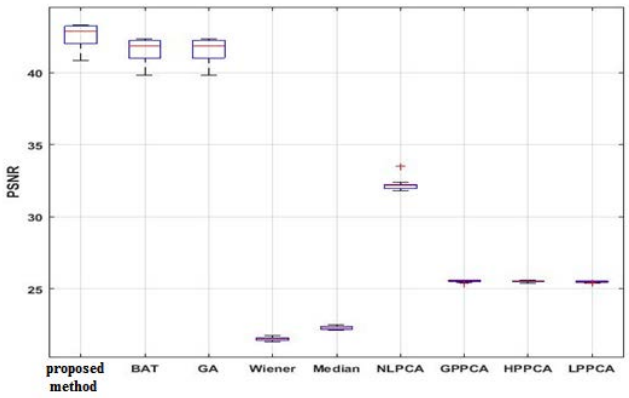


Fig. 8. Comparative Analysis based on PSNR Measure.

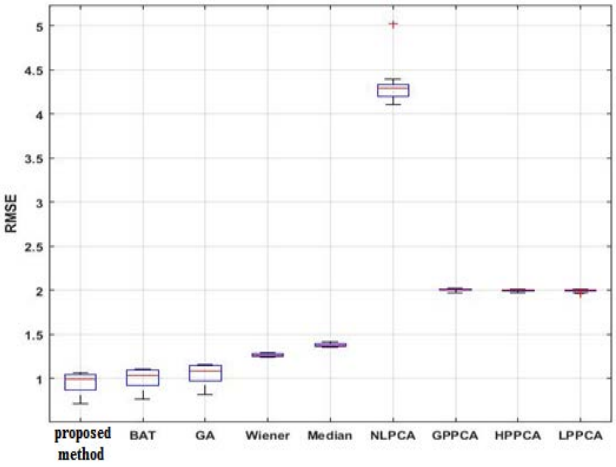


Fig. 9. Correlative Analysis based on RMSE Measures.

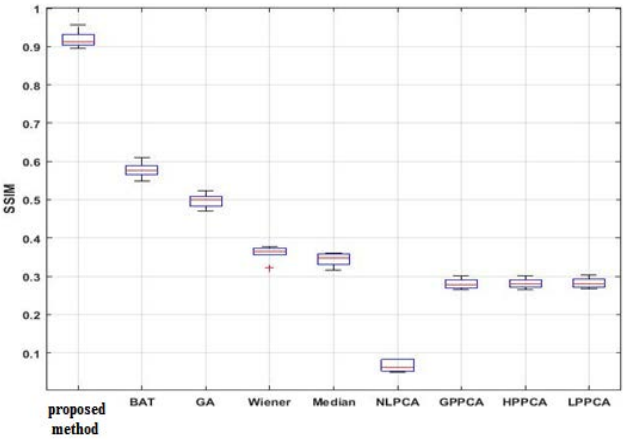


Fig. 10. Comparative Analysis based on SSIM Measures.

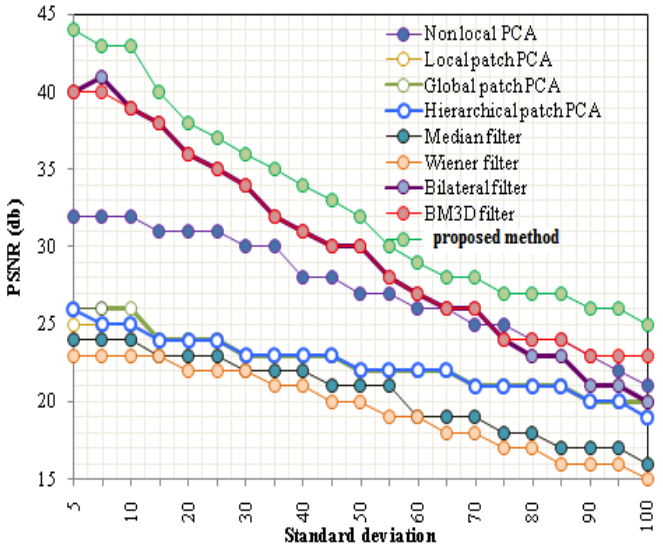


Fig. 11. Performance Comparison by using Mean of PSNR (dB) Values.

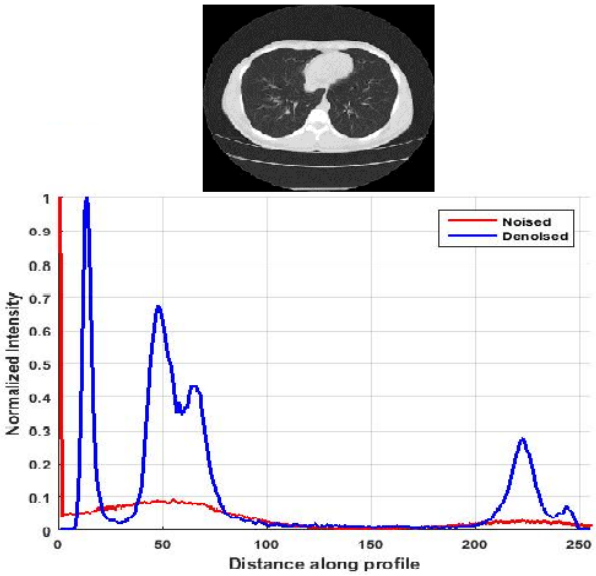


Fig. 12. Intensity Variation for Noised and de-noised Image.

TABLE I. PERFORMANCE EVALUATION USING PSNR MEASURE BY VARYING NOISE LEVEL

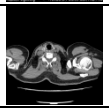
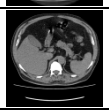
Images	Noise level	BM3D	Bilateral	Wiener	Median	NLPCA	GPPCA	HPPCA	LPPCA	Proposed
	0.02	38.83	38.83	20.55	21.26	30.77	24.58	24.55	24.12	39.83
	0.04	37.11	37.01	19.45	20.42	29.56	23.62	23.02	23.54	38.5
	0.06	35.56	35.23	18.6	19.24	27.10	21.56	22.52	22.32	37.2
	0.02	41.21	41.03	20.24	21.57	30.96	24.32	24.52	24.71	42.03
	0.04	39.67	39.21	19.54	20.63	28.68	23.78	23.02	23.83	41.45
	0.06	38.56	38.02	18.62	19.58	27.57	22.82	21.79	21.83	40.21
	0.02	41.05	39.32	20.48	22.27	31.12	25.41	24.39	25.36	42.05
	0.04	38.21	37.34	19.54	21.54	30.52	24.57	23.47	24.78	41.56
	0.06	37.21	36.33	18.45	20.63	29.47	23.67	22.10	23.02	40.11
	0.02	40.35	40.35	20.51	21.15	31.25	24.37	24.54	24.52	41.34
	0.04	39.47	39.32	19.46	20.64	30.68	23.79	23.89	23.68	40.45
	0.06	38.46	38.21	18.47	19.59	29.68	22.68	22.75	22.68	39.45
	0.02	40.67	40.63	20.61	21.45	30.59	24.69	24.30	24.36	41.21
	0.04	39.56	39.25	19.57	20.68	29.59	23.74	23.85	23.67	40.23
	0.06	38.02	37.95	18.35	19.57	28.51	22.64	22.67	22.65	39.46
	0.02	41.05	39.32	20.48	22.27	31.12	25.41	24.39	25.36	42.05
	0.04	38.21	37.34	19.54	21.54	30.52	24.57	23.47	24.78	41.56
	0.06	37.21	36.33	18.45	20.63	29.47	23.67	22.10	23.02	40.11
	0.02	40.35	40.35	20.51	21.15	31.25	24.37	24.54	24.52	41.34
	0.04	39.47	39.32	19.46	20.64	30.68	23.79	23.89	23.68	40.45
	0.06	38.46	38.21	18.47	19.59	29.68	22.68	22.75	22.68	40.45
	0.02	38.83	38.83	20.55	21.26	30.77	24.58	24.55	24.12	39.83
	0.04	37.11	37.01	19.45	20.42	29.56	23.62	23.02	23.54	38.5
	0.06	35.56	35.23	18.6	19.24	27.10	21.56	22.52	22.32	37.2
	0.02	41.21	41.03	20.24	21.57	30.96	24.32	24.52	24.71	42.03
	0.04	39.67	39.21	19.54	20.63	28.68	23.78	23.02	23.83	41.45
	0.06	38.56	38.02	18.62	19.58	27.57	22.82	21.79	21.83	40.21
	0.02	41.05	39.32	20.48	22.27	31.12	25.41	24.39	25.36	42.05
	0.04	38.21	37.34	19.54	21.54	30.52	24.57	23.47	24.78	41.56
	0.06	37.21	36.33	18.45	20.63	29.47	23.67	22.10	23.02	40.11

TABLE II. PERFORMANCE ANALYSIS IN COMPARISION WITH OTHER ALGORITHMS IN LITERATURE

Methods	**PSNR**
DTCWT	27.9
Curvelet	30.03
Harris&DWT	31.30
RDTDWT	34.45
SRTW	30.93
Proposed Method	42.05

V. CONCLUSION

In our paper, another new innovative image de-noising is proposed using optimal discrete wavelet transform, BM3D as well as the bilateral filter. In our work, the proposed strategy consists of two phases, which are named as discrete wavelet design and image denoising structure.

To improve the delicate regions with higher visual quality the DWT domain transform is applied; where the optimal coefficients are selected using the crow search optimization algorithm. Once the optimal coefficients are selected, the BM3D filtering algorithm is applied to the frequency sub-division bands of DWT image output. At the next stage, the bilateral filter is applied to take away the noise clearly as well as to keep the uncorrupted information well. Experimental results on CT medicinal images are obtainable to estimates presentation of a proposed filter. Our adaptive multi-stage noise removal methodology beat other methodologies based on PSNR, RMSE and SSIM measures.

REFERENCES

[1] DzungL.Phamy, ChenyangXu, Jerry L.Prince,"A Survey of Current Methods in Medical Image segmentation",Annual review of Biomedical Engineering2 (2000):315-37.

[2] Luis Alvarez, Pierre-Louis and Jean Michel, "Image selective and edge detection by non-linear diffusion", Society for industrial and applied mathematics 29.3(1992):845-866.

[3] Steven C.H.Hoi, Rong Jin, Jianke, Michael R, "Batch Mode Active Learning and Its Application to medical image classification",Proceeding of 23rd International conference on Machine Learning(2006):417-424.

[4] Avisha Sharma, Sanyam Anand, "An Efficient Technique of De-Noising Medical Images using Neural Network and Fuzzy-A Review", International Journal of Science and Modern Engineering (IJISME)1.4,(2013).

[5] Forest Agostinelli, Michael R., Anderson Honglak Lee, "Adaptive Multi-Column Deep Neural Networks with Application to Robust Image De-noising", Advances in neural information processing System 26(2013).

[6] A.BUADES, B.COLL, AND J.M. MOREL,"A Review of image denoising algorithms, with a new one", SIAM Journal on Multiscale Modeling and Simulation 4.2(2005).

[7] Yang, Qingsong, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K. Kalra, Yi Zhang, Ling Sun, and Ge Wang. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." IEEE transactions on medical imaging 37. 6 (2018): 1348-1357.

[8] Dabov, Kostadin, Alessandro Foi, and Karen Egiazarian. "Video denoising by sparse 3D transform-domain collaborative filtering." 15th European Signal Processing Conference, IEEE (2007) : 145-149.

[9] Li, Zhoubo, Lifeng Yu, Joshua D. Trzasko, David S. Lake, Daniel J. Blezek, Joel G. Fletcher, Cynthia H. McCollough, and Armando Manduca. "Adaptive nonlocal means filtering based on local noise level for CT denoising." Medical physics 41.1 (2014).

[10] Elhoseny, M., & Shankar, K. Optimal bilateral filter and Convolutional Neural Network based denoising method of medical image measurements. Measurement, 14.3(2019):125- 135.

[11] Manduca, Armando, Lifeng Yu, Joshua D. Trzasko, Natalia Khaylova, James M. Kofler, Cynthia M. McCollough, and Joel G. Fletcher. "Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT." Medical physics 36.11(2009): 4911-4919.

[12] Katsuhiro Ichikawa, Hiroki Kawashima, Masato Shimada, Toshiki Adachi and Tadanori Takata, "A three-dimensional cross-directional bilateral filter for edge-preserving noise reduction of low-dose computed tomography images", computer in biology and medicine11.1(2019).

[13] Wojciech Wieclawek and Ewa Pietka, "Granular filter in medical image noise suppression and edge preservation", Bio-cybernetics and Biomedical Engineering39.1(2019):1-16.

[14] Hsuan-Ming Huang and Chieh Lin, "A kernel-based image denoising method for improving parametric image generation", journal of medical image analysis, vol.55, pp.41-48, 2019.

[15] Manoj Diwakar and Pardeep Kumar, "Wavelet Packet Based CT Image Denoising Using Bilateral Method and Bayes Shrinkage Rule", Handbook of Multimedia Information Security: Techniques and Applications(2019) : 501-511.

[16] Manoj Diwakar, Arjun Verma, Sumita Lamba and Himanshu Gupta, "Inter- and Intra- scale Dependencies-Based CT Image Denoising in Curvelet Domain", Soft Computing: Theories and Applications(2018): 343-350.

[17] Bing-quan Chen, Jin-ge Cui, Qing Xu, Ting Shu and Hong-li Liu , "Coupling de-noising algorithm based on discrete wavelet transform and modified median filter for medical image", Journal of Central South University, 26.1 (2019): 120–131.

[18] Madhuri Khatri and Raunakraj Patel, "A Survey on Image Denoising methods",International Journal of Engineering Development and Research 4.4 (2016).

[19] Jenita Subash, Kalaivani.S," Image denoising using improved Fuzzy based approach" Indian Journal of Computer Science and Engineering (IJCSE) 12.5(2021):1325 -1333.

[20] Shyna A, Thomas Kurian,Jayakrishnan,Jidu Nandan, Mohammed Hazm Aneez ." A Fuzzy Based IMAGE Denoising Filter Using Non-Linear Fuzzy Membership Functions" Journal of Network Communications and Emerging Technologies 10.11(2020):1-8.

[21] Devinder Singh, Amandeep Kaur," Improved Fuzzy Based Non-Local Mean Filter to Denoise Rician Noise "Turkish Journal of Computer and Mathematics Education 12.7(2021):2116- 2121.

[22] V Vijay Kumar Raju, M Prema Kumar, "Denoising of MRI and XRay images using Dual Tree Complex Wavelet and Curve let Transforms", International Conference on Communication and Signal Processing-IEEE (2014).

[23] K.V.Thakur, Jitendra Kadam, A.M. Sapkal,' Poisson Noise Reduction from X-ray Medical Images Using Modified Harris Operator and Wavelet Domain Thresholding', International Conference of Instrumentation and Control (ICIC),IEEE conference (2015):568-572.

[24] K.V.Thakur, Pramod Ambhore, A.M. Sapkal,'A combined approach for Noise reduction in medical images using Dual tree Discrete Wavelet Transform and Rotated Dual tree Discrete Wavelet Transform", International Conference of Instrumentation and Control (ICIC),IEEE conference(2015):1003-1007.

[25] Sibin Wu Qingsong Zhu andYaoqinXiePh.D,"Evaluation of various specklereduction filters on medical ultrasoundimages", ,35th annual internationalconference of the IEEE EMBS Osaka, Japan(2013) :3-7.

Smart Tourism Recommendation Model: A Systematic Literature Review

Choirul Huda, Arief Ramadhan*, Agung Trisetyarso, Edi Abdurachman, Yaya Heryadi
Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
Bina Nusantara University, Jakarta, Indonesia

***Abstract*—The tourism industry has become a potential sector to leverage economic growth. Many attractions are detected on several platforms. Machine learning and data mining are some potential technologies to improve the service of tourism by providing recommendations for a specific attraction for tourists according to their location and profile. This research applied for a systematic literature review on tourism, digital tourism, smart tourism, and recommender system in tourism. This research aims to evaluate the most relevant and accurate techniques in tourism that focused on recommendations or similar efforts. Several research questions were defined and translated into search strings. The result of this research was promoting 41 research that discussed tourism, digital tourism, smart tourism, and recommender systems. All of the literature was reviewed on some aspects, in example the problem addressed, methodology used, data used, strength, and the limitation that can be an opportunity for improvement in future research. This study proposed some references for further study based on reviewed papers regarding tourism management, tourist experience, tourist motivation, and tourist recommendation system. The opportunities for a further research study can be conducted with more data usage especially for a smart recommender system in tourism through many types of recommendation techniques such as content-based, collaborative filtering, demographic, knowledge-based, community-based, and hybrid recommender systems.**

Keywords*—*Systematic review; tourism; smart tourism; digital tourism; recommender system

I. Introduction

The tourism industry sector has become a potential sector for economic growth in various regions. The World Tourism Organization (UNWTO) reported in January 2020 that the international tourism arrival reached 1.5 billion with a growing percentage of 4% in 2019 [1]. This growth will certainly take place in line with the discovery of various new tourist destinations in various regions. In China, the tourism market growth of inbound and outbound tourism changed markedly in the past few years [2]. This potential tourism industry growth was also found in Thailand [3].

Tourism industries have faced a hard challenge since the coronavirus disease 2019 (COVID-19) pandemic. This pandemic has hurt the global economy, causing unprecedented global health and social emergencies [4]. Some studies have stated that this pandemic will end in the next few years so that it can help the growth of the tourism sector as in the previous year. Every country has a different level of rebound from the COVID-19 pandemic according to the numbers of daily confirmed COVID-19 cases [4]. Tourism managers should establish marketing strategies and improve their service [5].

As the tourism industry gets an opportunity to rebound, it is necessary to prepare adequate strategies and activities to be able to provide good service for every tourist and all stakeholders. This preparation needs to involve all stakeholders by paying attention to their respective roles. Intrinsic motivation and extrinsic motivation had a highly positive impact on perceived trust in the tourism industry [6]. The aim of this study is to conduct a literature review that supports the preparation of a tourism recommendation model that focuses on two aspects, namely tourist attractions, and tourist smart services. Along with the development of information technology, today we are faced with the flow of information that floods various media regarding searching for some tourism information. Information on various attractions in various locations can be obtained through various media. Some of the media or sources of information that we conventionally use to find information related to tourism include travel agents, brochures, and social media [7].

Human lifestyle in the tourism and hospitality areas has been influenced by Information and Communication Technologies (ICT) [8]. Today, ICT has become a strategic platform for business and other purposes. Digital technologies as a part of ICT have introduced important innovations in many human aspects of life such as factories, hospitals, hotels, cities, and territories [9]. By using a mobile device, as a popular digital technology, many people can get information easily. They can brow some web site, search engines, social media, and any other digital platform to get the needed information. In some cases, they get private messages or notifications for some uninteresting information. The flooding of information sometimes harms some people because of unsuitable information. For getting tourism information people can use some search engine manually. Of course, this searching is not efficient yet because they will get the flooding of information even though they have used keywords for searching. Frequently mobile devices have been used by contemporary travelers for making decisions in traveling and managing travel itineraries [8]. Furthermore, travel, tourism, and hospitality companies have started to adopt some digital technologies such as Artificial Intelligence (AI), robots, and service automation (RAISA) through chatbots, delivery robots, the concierge of a robot, conveyor restaurants, self-service kiosks, and many others [9].

Regarding the trend of smart ICT especially in intelligent systems and location detection, the tourism industry has the

*Corresponding Author.

opportunity to improve better services for stakeholders. Through smart ICT Tourist as a user of tourism industry needs more smart system and dynamics recommendation for specific recommendation in tourism in order to improve tourism experience and tourism service. Smart tourism is various changes driven by the application of some new technologies in tourism and it is a modernization of providing tourism service, innovating tourism management, improving tourism experience, and optimizing the tourism resources usage [10]. It would be important for managers of destinations to make greater development in the tourism destination resources, in order to provide a competitive advantage and enhance experiences [11]. Interactive solution through online media is needed for potential tourist for improvement of their tourism experience [12]. The smart technology solution in tourism should consider some effectiveness and healthy of tourism experience related to digital well-being experienced [13]. Search strings as the form of research questions translation were used in this Systematic Literature Review (SLR) method regarding the need for appropriate search results.

II. METHOD

For achieving the objective, this SLR has three main phases: Planning the Review, Conducting the Review, Reporting the Review [14]. Additional explanation of these main phases and the sub-phases is showed in Fig. 1.

All the stages listed in Fig. 1 may appear to be sequential, but practically some of the stages can be repeated for further review. This review is conducted to summarize the existing literature related to the relevant and accurate techniques for recommender systems in tourism through the defined problem, method, data usage, strength, and limitation.

This review protocol is built to reduce bias from the researcher in the selection of individual studies that may be influenced by researcher intentions [14]. The SLR activities also adopt several approaches in [15]. The review protocol is shown in Fig. 2. The steps of the review protocol are slightly modified to achieve the aim of this research.

There are ten steps of the review protocol according to Fig. 2. The activities are arranged in sequence as a guideline before the review of papers conducted. The first step of the review protocol is defining or formulating the research question (RQ) as the most important activity during protocol [14]. The research questions are showed in Fig. 3.

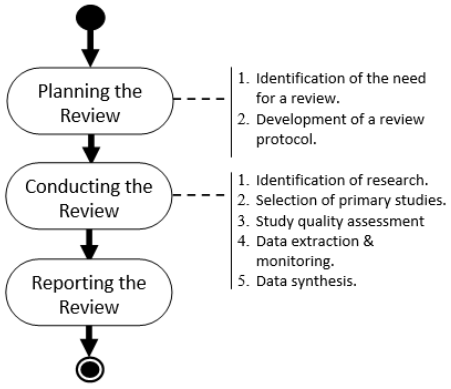


Fig. 1. The Phases of SLR.

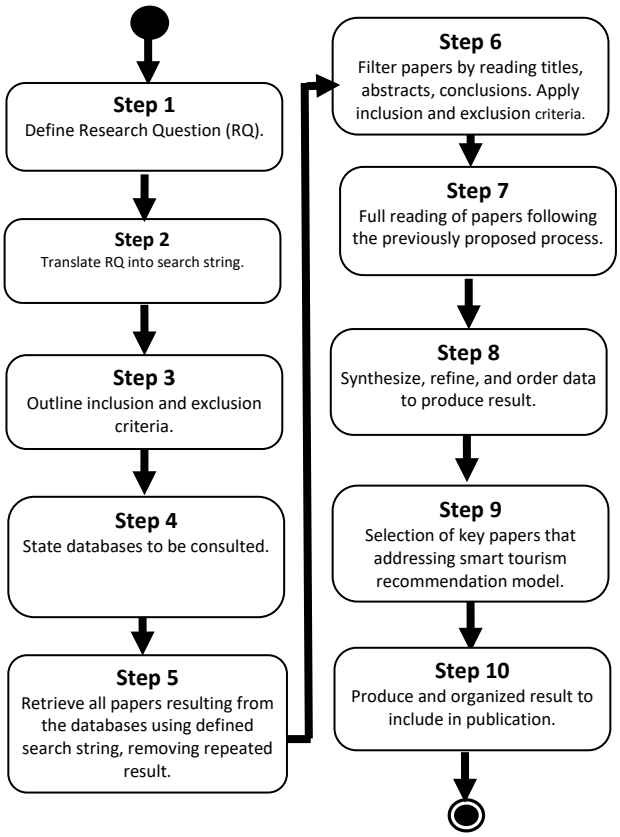


Fig. 2. Steps of the Review Protocol.

RQ1	What is the problem to be solved?
RQ2	What kind of data is being used?
RQ3	What kind of methodology is being used?

Fig. 3. Defined Research Questions.

The second step is translating the RQ into search strings such as: “Tourism”, “Digital Tourism”, “Smart Tourism”, “Recommender System”. The complete search strings can be seen in Fig. 4. The third step is to outline the inclusion and exclusion criteria to specifying the selected paper according to the inclusion aspect and exclusion aspect. The inclusion is intuition aspect that is considered for further process, while the exclusion is the aspect that should be consideration of rejection. The further explanation of the inclusion aspects, exclusion aspects, and justifications are showed in Table I.

String1	“Tourism”
String2	“Digital Tourism”
String3	“Smart Tourism”
String4	(“Recommender System” AND “Tourism”)
String5	(“Recommendation System” AND “Tourism”)

Fig. 4. Search Strings.

TABLE I. CRITERIA (INCLUSION, EXCLUSION), AND RELATED JUSTIFICATION

Inclusion criteria	**Justification**
Published papers in 2016 to 2021 in journal or conference proceedings.	Use the most recent findings only.
Papers present tourism study regarding smart tourism recommendation model.	The opportunity of smart tourism recommendation model development in the future.
Exclusion criteria	**Justification**
Not in English written papers.	Standardization of international language in English.
Paper is a secondary (review) or tertiary study.	Focus on primary studies.

The fourth step is stating the reputable databases to be consulted for the source of papers. Several databases were classified into four groups according to authors' experiences and authors' decisions. The first group consists of IEEE, Elsevier, Science Direct, Springer, ProQuest, Emerald, Wiley, and UNWTO. The second groups consist of Routledge, IOP, MDPI, ICCT, and IJTC. The third group consists of ACIS, AITM, E3S, ICEB, and SAGE. The fourth group consists of CBUNI and Pertanika.

The review process was started by the first group and followed by the second group, third group, and finally fourth group. The searching process was conducted through the "Publish or Perish" application developed by Prof. Anne-Will Harzing [16], and Mendeley software applications. The fifth step is the retrieval of all papers resulting from the searching process and also removing repeated results. The sixth step is filtering papers by reading part of papers starting by reading the titles, abstracts, conclusions and then applying the inclusion and exclusion criteria. The seventh step is conducted by a full reading of papers following the reviewing part of papers, i.e. introduction, methodology, result and discussion, and any other part of papers. This step is conducted manually by reading every statement from all parts of the paper to find more implicit and explicit ideas and the relationship between every section in existing research. The eighth step is to synthesize, refine, and order extracted data to produce the result in some figures and tables. The result of this study is presented based on some aspects, and be analyzed based on the process, the database used, topics, and years of publication or conference. The ninth step is the selection of key papers that addressing the smart tourism recommendation model. The tenth step is producing and organizing the review result to be submitted in international publication.

III. RESULTS

The results in some topics seemed on a small number because of the specific requirement on searching proses and the result of the review protocol. Some papers were retrieved but many of them were not providing the recommender system in tourism through computer science research.

Steps of filtering papers were conducted through searching from the databases according to the defined search strings. The summary of this process is presented in Table II. Because of similarity in content, the result from searching by using ("Recommender System" AND "Tourism") and ("Recommendation System" AND "Tourism") as search strings are summarized become one-row label: "Recommender System" at Table II.

Summary of publication period and related databases used on searching based on steps above are presented in Table III.

The locations of selected research were distributed into many regions or continentals with the majority of research were established in Asia (China, Indonesia, Japan, India, and Hong Kong). Further information on distributed research by region is displayed in Table IV. This discussion consists of some topics according to the topics in the searching process. The discussion was sequentially be arranged through Tourism, Digital Tourism, Smart Tourism, and Recommender Systems. The questions of what, why, how, when, where, and who were considered on every discussion based on the available information on each paper.

TABLE II. SEARCH RESULT

TOPIC	**REVIEW**	**DEEP REVIEW**	**RELEVANT**	**SELECTED**
Tourism	36	17	15	13
Digital Tourism	15	11	11	10
Smart Tourism	25	15	13	13
Recommender System	6	6	6	5
Total	**82**	**49**	**45**	**41**

TABLE III. SELECTED PAPERS BY DATABASE USED AND YEAR OF PUBLICATIONS

DATABASE	**YEARS**						**TOTAL**
	2021	**2020**	**2019**	**2018**	**2017**	**2016**	
Elsevier	3	1	0	0	0	1	**5**
IEEE	1	2	1	2	1	0	**7**
IOP	1	2	1	1	0	0	**5**
MDPI	0	5	0	0	0	0	**5**
ProQuest	0	1	0	0	0	0	**1**
Routledge	0	3	4	0	0	0	**7**
SAGE	0	2	1	0	0	0	**3**
Science Direct	0	2	2	0	1	0	**5**
Springer	1	0	1	0	0	0	**2**
UNWTO	0	1	0	0	0	0	**1**
TOTAL	**6**	**19**	**10**	**3**	**2**	**1**	**41**

TABLE IV. SELECTED PAPERS BY DATABASE USED AND REGIONS

REGIONS	**TOPICS**				**TOTAL**
	Digital Tourism	**Recommender System**	**Smart Tourism**	**Tourism**	
Africa	0	0	1	0	**1**
Asia	3	2	9	6	**20**
Europe	4	0	1	3	**8**
South America	1	1	0	1	**3**
Global	2	2	2	3	**9**
TOTAL	**10**	**5**	**13**	**13**	**41**

IV. Discussion

A. Tourism

Thirteen papers were selected for this topic discussion below and introduction above. The authors in [1] and [4] provide the contribution, opportunity, and challenge of the tourism industry. How is the pandemic will rebound was stated in [4]. The first paper being analyzed is the paper produced by The World Tourism Organization (UNWTO) for the global tourism industry based on January 2020 report. Tourism industries face a hard challenge on Pandemic COVID-19, but it will be rebound for the next period according to research from [4] with various recovery levels for each country.

This sub-topic was discussed in [17] and [18]. The author in [17] used Exploratory Factor Analysis (EFA), T-test, and ANOVA analysis to explore tourists' motivations in China, and involvement in adventure tourism activities, and also the kinds of personality and location affect their motivation and involvement in 2019. This research provided the reasons for preferred adventure activities and key requirements of adventure activity but using a relatively small sample size and only focus on adventure tourism in Chengdu and Xiamen, China.

Lindberg et. al. in [18] describe the integrated analysis of individual characteristics, such as attitudes and demographic factors, and situational characteristics, such as interpretive center features through discrete choice models use random utility theory in Norway in 2019. This paper gave an excellent aspect of analysis based on respondent characteristics according to basic demographic data (gender, education, age, and income) but has limitations on the generalization of the demographic factor for other locations like Asia. The author in [19] explains tourist experience regarding the judgment of tourists' willingness to pay to increase the public managers' income to produce policies through binary logistic regression and decision. This paper used travel characteristics and variables of sociodemographic of tourists visiting Andalusia (Spain) in 2021.

Based on the research question and discussion in tourism, the selected papers have some various problems to be solved. All of them can be seen in Table V.

Based on the research question and discussion in tourism, the selected papers have some various data used, and methodologies used. All of them can be seen in Table VI.

Based on the research question and discussion in tourism, the selected papers have some various methodologies used. All of them can be seen in Table VII.

TABLE V. Problem Types in Tourism Topics

Problem	∑ Papers	Selected Papers
Tourism Management	3	[1], [2], [4]
Tourist Experience	4	[12], [13], [19], [20]
Tourist Motivation	6	[17], [18], [11], [3], [6], [5]
TOTAL	**13**	

TABLE VI. Data used in Tourism Topics

Data Used	∑ Papers	Selected Papers
Questionnaire	7	[17], [18], [19], [11], [12], [6], [5]
Paper/ Report	6	[1], [4], [2], [3], [13], [20]
TOTAL	**13**	

TABLE VII. Method Types in Tourism Topics

Methods	∑ Papers	Selected Papers
Review	4	[1], [18], [13], [20]
Exploratory factor analysis (EFA), T-test, and ANOVA	1	[17]
Exploratory Factor Analysis through Kaiser–Meyer–Olkin (KMO)	1	[3]
Exploratory Factor Analysis through K-means, ANOVA	1	[5]
Statistical Model	6	[4], [19], [11], [2], [12], [6]
TOTAL	**7**	

B. Digital Tourism

The next topic discussion is digital tourism that is a fundamental platform for smart tourism recommendation models through digital technology. Digital tourism is a part of digital transformation in the tourism industry that used digital technology as a strategic platform for transformation from the traditional approach to digital. The digital platform offers a better solution for tourism rather than the traditional approach through a more effective and efficient service for stakeholders in tourism. Some various approach was conducted at the existing research in different regions. Ten papers were selected at final review: [8], [9], [10], [21], [22], [23], [24], [25], [26], [27].

In [8], it was addressed the problem of overuse in technology-related addiction issues and mental health regarding tourism context has been explored through literature review and exploratory study about the perception of 17 participants on 2020. On the other hand, the impact of the digital revolution on tourism, the different and common work between tourism 4.0 and smart tourism through a conceptual approach is addressed in [9].

The author in [10] used blockchain technology to develop a tourism information intelligent service platform for tourism enterprises, tourists, government and promoted tourism management and service coordination, and economic development in China. In [21], it was developed a system to automatically recommend tourists to visit a particular tourist destination through an ontological approach. This approach included current situation detection such as location, existing fixed schedule, time, means of transportation, place accessibility to visit Japan in 2019.

Veloso et. al. in [22] identified the main challenges within tourism crowdsourcing platforms through a comparison of existing models. They detect future research trends to ensure the quality and authenticity of tourism-related crowdsourced data by using published papers about tourism crowdsourcing

platforms. The author in [23] developed a software architecture that can be specified in a web application based on the proposal of a management model of the NTS-TS 002 standard through conceptual software design in Columbia. Paper [24] conducted IoT-based research on tourism industrial clusters and information platforms through niche principles and its theoretical framework in China.

Briciu in [25] use Virtual Reality and mobile application for city development in terms of cultural tourism solutions and evaluation on cultural heritage sites through descriptive analysis in Romania. The author in [26] conducted comparative and content analysis for tourism zone development in Russia's economic space through tourism technology platforms implementation. Paper [27] conducted a Blackwell-Miniard-Engel model to develop a customer journey map creation in Russia.

Based on the research question and discussion in digital tourism, the selected papers have some various problems to be solved. All of them can be seen in Table VIII.

TABLE VIII. PROBLEM TYPES IN DIGITAL TOURISM

Problem	∑ Papers	Selected Papers
Tourism Management	2	[23], [24]
Tourist Experience	7	[22], [21], [10], [9], [25], [26], [27]
Tourist Motivation	1	[8]
TOTAL	**10**	

Based on the research question and discussion in digital tourism, the selected papers have some various data used. All of them can be seen in Table IX.

TABLE IX. DATA USED IN DIGITAL TOURISM

Data Used	∑ Papers	Selected Papers
RDB	1	[10]
Web Site	3	[10], [25], [26]
Questionnaire	1	[8]
Google Map	1	[10]
Paper/ Report	8	[22], [21], [9], [23], [25], [24], [26], [27]
TOTAL	**14**	

Based on the research question and discussion in digital tourism, the selected papers have some various methodologies used. All of them can be seen in Table X.

TABLE X. METHOD TYPES IN DIGITAL TOURISM

Methods	∑ Papers	Selected Papers
Review	5	[8], [22], [25], [26], [27]
Conceptual Model	5	[21], [10], [9], [23], [24]
TOTAL	**10**	

C. Smart Tourism

Further topic discussion is smart tourism that used various approaches in research and model development. Some papers used an intelligent approach while others used some different approaches. It was explored in [28], a tourist preference methodological approach of Smart Tourism Attractions (STA), and comprehensive review for the weaknesses and strengths of an STA through related study site, literature on Hongshan Zo in China, data collection by questionnaire for collecting the travel experience, and FCEM-AHP evaluation in 2016. The authors in [29] have addressed a definition comparison between traditional and smart tourism information services based on published papers.

Ramadhani in [30] explained the creation of the automatically rating system of tourism destination based on a travel blog through an automatic approach by a new semantic analysis algorithm in Indonesia in 2017. However, in [31], it was found the important factors for the tourists' preferences to visit the Vredeburg museum in Yogyakarta, Indonesia in 2018. This research provided the preferences of tourists towards smart tourism and having challenges for a more general purpose.

The author in [32] provides a better user travel experience of smart tourism in China in 2021. On the other hand, Dey in [33] focused on a literature study on Indian tourism sectors providing online services and talking about the current Artificial Intelligent used in India in 2020. It briefed technology usages in tourism and needs more introductions for implementation of recommender system in tourism.

A system is developed in [34] to arrange professional game rangers for a visitor in a particular tourist destination using machine learning through Google's TensorFlow based on animal mages in South Africa in 2018. The author in [35] introduced available datasets for the public in Europe regarding the area of tourism demand prediction for future comparisons and experiments using Pearson correlation analysis for features extracted from the environmental, social media, and official datasets.

Li in [36] extracted useful search query data and construct relevant econometric models. However, in [37] it was proposed the incorporation of reliable traditional methodologies with text analytics and machine learning to facilitate a deeper understanding of concepts and theory building through a step-by-step methodological and analytical framework based on an analysis of online reviews of existing research.

The author in [38] provided an understanding of the aspirations of people with visual impairments in terms of tourism and explore how smart tourism destinations could potentially enhance the tourism experience they offer. Qualitative research was adopted at this research through in-depth expert interviews and multisensory observation and implemented the PERMA model as a framework for designing the app in Hong Kong. The author in [39] provides definitional clarity and a comprehensive approach to the smart tourism city anatomy regarding smart tourism and smart city through a conceptual approach. The author in [40] provided the relationship among smart tourism technology attributes, travel satisfaction, happiness, and revisit intention regarding travel experience satisfaction through the structural equation method in China. The questionnaire was conducted for Chinese tourists' travel satisfaction analysis.

Based on the research question and discussion in smart tourism, the selected papers have some various problems to be solved. All of them can be seen in Table XI.

TABLE XI. PROBLEM TYPES IN SMART TOURISM

Problem	∑ Papers	Selected Papers
Tourism Management	2	[39], [37]
Tourist Experience	9	[30], [34], [32] [33], [29] [35], [36], [38], [40]
Tourist Preference	2	[28], [31]
TOTAL	**13**	

Based on the research question and discussion in smart tourism, the selected papers have some various data used. All of them can be seen in Table XII.

TABLE XII. DATA USED IN SMART TOURISM

Data Used	∑ Papers	Selected Papers
RDB	1	[32]
Web Site	3	[30], [34], [35]
Social Media	1	[35]
Questionnaire	4	[30], [31], [38], [40]
Search Engine	1	[36]
Paper/ Report	5	[28], [33], [29], [37], [39]
TOTAL	**15**	

Based on the research question and discussion in smart tourism, the selected papers have some various methodologies used. All of them can be seen in Table XIII.

TABLE XIII. METHOD TYPES IN SMART TOURISM

Methods	∑ Papers	Selected Papers
Review	6	[30], [28], [33], [29], [38], [39]
Statistical Model	3	[31] [35], [40]
Conceptual Model	1	[37]
Intelligent Model	3	[34], [32], [36]
TOTAL	**13**	

D. Recommender Systems

The final topic discussion is about Recommender Systems according to the selected papers: [7], [41], [42], [43], and [44]. In [7], it is offered a travel recommender system for the effects of automating Word-of-Mouth (WOM) and established personalized travel-planning services to tourists through Collaborative Filtering (CF)-based recommender using WOM communication. This research used tourists' preference ratings on some destinations through interpersonal communication in South Korea in 2019. The author in [41] provide a framework of a travel recommender system by combining knowledge-based filtering and hybrid recommendation methods with decision-making theory in China in 2020.

Sagar et al. in [42] developed a hotel recommender system based on collaborative filtering and regression. Whereas Roy and Dietz in [43] develop a travel recommender system based on the content-based recommender system method. This research used user information and preferences such as home region, destination region, traveler type, and maximum travel duration and fondness for different types of venues in a city through Twitter social media in 2021.

The author in [44] proposed a solution for detection of tourist implicit preferences based on photos from social media on Facebook, Instagram, and Google Plus, and recommend a set of emerging techniques for tourism such as Convolutional Neural Network (CNN) and fuzzy logic to classify tourists and provide the recommendation at Brazil on 2018. This research provided two recommendation types: item-based and user-based.

Based on the research question and discussion in recommender system in tourism, the selected papers have one type of problem to be solved. All of them can be seen in Table XIV.

TABLE XIV. PROBLEM TYPES IN RECOMMENDER SYSTEM.

Problem	∑ Papers	Selected Papers
Tourist Recommendation	5	[7], [44], [41], [43], [42]
TOTAL	**5**	

Based on the research question and discussion in recommender system in tourism, the selected papers have various data used. All of them can be seen in Table XV.

TABLE XV. DATA USED IN RECOMMENDER SYSTEM

Data Used	∑ Papers	Selected Papers
RDB	1	[7]
Web Site	2	[7], [42]
Social Media	2	[44], [43]
Paper/ Report	1	[41]
TOTAL	**6**	

Based on the research question and discussion in recommender system in tourism, the selected papers have various methodologies used. All of them can be seen in Table XVI.

TABLE XVI. METHOD TYPES IN RECOMMENDER SYSTEM

Methods	∑ Papers	Selected Papers
Intelligent Model (AI, ML, CNN, FL, CF)	5	[7], [44], [41], [43], [42]
TOTAL	**5**	

This research found that motivation and preference in tourism should become a concern for a better tourist experience. Digital technology offers some efficiency and flexibility for the tourist experience. Implementation of digital tourism offers better integration for stakeholders for maintaining, searching, and decision making of tourism factors such as facility, attraction, location. Some approaches for digital tourism can be adopted from [10] and [9]. For further smart experience, a smart tourism solution should be conducted. Smart tourism development through any type of

data format and intelligent models that will improve the tourist experience and management can be adopted from [34], [35], and [40]. A personal recommendation for a better tourist experience can be provided by a recommender system in tourism through various approaches. Recommender system improvement through Collaborative filtering and Content-based Recommender systems can be adopted from [43], and [7] with further personalization and generalization for other attractions.

V. Implication and Conclusion

This study provided 41 selected papers from various data sources, regions, and years that were reviewed and promoted as references for the next study on tourism, digital tourism, smart tourism, and recommender systems. This research found opportunities for further research in tourism based on the review and discussion of existing research through their addressed problems, methodologies, data usages, research times and locations, strengths, and limitations. The next study can be conducted on a different aspect of what, why, where, and how to solve the problem based on previous studies and the enhancement in the future. Due to the conceptual model provided from this research, some steps or development and result were provided on some figures that need further discussion.

The result of this study will help decision-makers in the tourism industry to improve their service for stakeholders in the tourism industry. Findings in tourism, digital tourism, smart tourism, and recommender system will help different approaches to decisions. Providing public policy in tourism and excellent facilities by local government will contribute to tourist experience and willingness to pay. The available personal recommendation will help the tourist to make a good decision in traveling. This study found some guidelines for tourism management to government through existing study. The opportunity of the tourism industry rebound after a pandemic should be prepared with adequate strategy although every country has a different scale of time to rebound.

Limitations of the current research regarding the scalability of data usage and quality of data usage and their methodologies can be enhanced to the next research. This study proposed some references for further study based on reviewed papers regarding tourism management, tourist experience, tourist motivation, and tourist recommendation system. The opportunities for a further research study can be conducted with more data usage especially for a smart recommender system in tourism through many types of recommendation techniques such as content-based, collaborative filtering, demographic, knowledge-based, community-based, and hybrid recommender systems. The tourist recommendation system in the future can be started with a study in link and match of personal demographics with available attractions on dynamic location through machine learning model, and location detection technology regarding available data in social media.

References

[1] W. Tourism. and UNWTO, “UNWTO World Tourism Barometer and Statistical Annex, May 2020”, UNWTO World Tourism Barometer, 18(2), 2020, pp. 1–48. doi: 10.18111/wtobarometereng.2020.18.1.2.

[2] H. Huang, ‘”The spatial distribution, influencing factors, and development path of inbound tourism in China-An empirical analysis of market segments based on travel motivation”, Sustainability (Switzerland), 12(6), 2020. doi: 10.3390/su12062508.

[3] P. Fakfare, “A scale development and validation on domestic tourists’ motivation: the case of second-tier tourism destinations”, Asia Pacific Journal of Tourism Research, 25(5), 2020, pp. 489–504. doi: 10.1080/10941665.2020.1745855.

[4] H. Zhanga, H. Song, L. Wen, and C. Liu, “Forecasting tourism recovery amid COVID-19”, Annals of Tourism Research, 87, 2021, p. 103149. doi: 10.1016/j.annals.2021.103149.

[5] M. Carvache-Franco, “Segmentation and motivations in eco-tourism: The case of a coastal national park”, Ocean and Coastal Management, 178, 2019. doi: 10.1016/j.ocecoaman.2019.05.014.

[6] M. Kim, “The effects of motivation, deterrents, trust, and risk on tourism crowdfunding behavior”, Asia Pacific Journal of Tourism Research, 25(3), 2020, pp. 244–260. doi: 10.1080/10941665.2019.1687533.

[7] I. Choi, “A recommender system based on personal constraints for smart tourism city *”, Asia Pacific Journal of Tourism Research, 2019. doi: 10.1080/10941665.2019.1592765.

[8] I. Egger, “Digital free tourism – An exploratory study of tourist motivations”, Tourism Management, 79, 2020. doi: 10.1016/j.tourman.2020.104098.

[9] T. Pencarelli, “The digital revolution in the travel and tourism industry”, Information Technology and Tourism, 22(3), 2020, pp. 455–476. doi: 10.1007/s40558-019-00160-3.

[10] C. Wei, “Research on Construction of a Cloud Platform for Tourism Information Intelligent Service Based on Blockchain Technology”, Wireless Communications and Mobile Computing, 2020. doi: 10.1155/2020/8877625.

[11] M. H. Pestana, “Motivations, emotions and satisfaction: The keys to a tourism destination choice”, Journal of Destination Marketing and Management, 16, 2020. doi: 10.1016/j.jdmm.2018.12.006.

[12] D. Suhartanto, “Tourist loyalty in creative tourism: the role of experience quality, value, satisfaction, and motivation”, Current Issues in Tourism, 23(7), 2020, pp. 867–879. doi: 10.1080/13683500.2019.1568400.

[13] U. Stankov, “Digital well-being in the tourism domain: mapping new roles and responsibilities”, Information Technology and Tourism, 2021. doi: 10.1007/s40558-021-00197-3.

[14] B. Kitchenham, “Procedures for performing systematic reviews”, Keele, UK, Keele University, 2004.

[15] A. Ramadhan, D. I. Sensuse, Muladno, and A. M. Arymurthy, ”Synthesizing success factors for e-government initiative”, Research Journal of Applied Sciences, Engineering and Technology, Vo. 6, No. 9, 2013, pp. 1685-1702. doi: 10.19026/rjaset.6.3891.

[16] A. W. Harzing, “Publish or Perish”, 2020, https://harzing.com/resources/publish-or-perish.

[17] X. Jin, “Motivation and involvement in adventure tourism activities: a Chinese tourists’ perspective”, Asia Pacific Journal of Tourism Research, 24(11), 2019, pp. 1066–1078. doi: 10.1080/10941665.2019.1666152.

[18] K. Lindberg, K. Veisten, and A. H. Halse, “Analyzing the deeper motivations for nature-based tourism facility demand: a hybrid choice model of preferences for a reindeer visitor center”, Scandinavian Journal of Hospitality and Tourism, 19(2), 2019, pp. 157–174. doi: 10.1080/15022250.2018.1482565.

[19] J. L. Durán-Román, P. J. Cárdenas-García, and J. I. Pulido-Fernández, “Tourists’ willingness to pay to improve sustainability and experience at destination”, Journal of Destination Marketing and Management, 19, 2021. doi: 10.1016/j.jdmm.2020.100540.

[20] J. Stienmetz, J. Kim, Z. Xiang, and D. R. Fesenmaier, “Managing the structure of tourism experiences: Foundations for tourism design”, Journal of Destination Marketing and Management, 19, 2021. doi: 10.1016/j.jdmm.2019.100408.

[21] Z. Tang, and E. Pyshkin, “Ontological approach to personalized situational planning: Concept and scenarios”, Proceedings - 2019 IEEE International Conferences on Ubiquitous Computing and

Communications and Data Science and Computational Intelligence and Smart Computing, Networking and Services, IUCC/DSCI/SmartCNS 2019, 2019, pp. 561–564. doi: 10.1109/IUCC/DSCI/SmartCNS.2019.00118.

[22] B. Veloso, F. Leal, B. Malheiro, and F. Moreira, "Distributed trust & reputation models using blockchain technologies for tourism crowdsourcing platforms", Procedia Computer Science, 160, 2019, pp. 457–460. doi: 10.1016/j.procs.2019.11.065.

[23] F. G. Ramirez, "Architecture of a Technology Platform for sustainable tourism management under the NTS-TS 002 standard", IOP Conference Series: Materials Science and Engineering, 2019. doi: 10.1088/1757-899X/519/1/012029.

[24] X. Li, "Research on tourism industrial cluster and information platform based on Internet of things technology", International Journal of Distributed Sensor Networks, 15(7), 2019. doi: 10.1177/1550147719858840.

[25] A. Briciu, "Evaluating how "smart" Brasov, Romania can be virtually via a mobile application for cultural tourism", Sustainability (Switzerland), 12(13), 2020. doi: 10.3390/su12135324.

[26] V. N. Sharafutdinov, "Tourism Technology Platforms as a Tool for Supporting Competitiveness of Regional Tourism Products", Regional Research of Russia, 10(1), 2020, pp. 48–55. doi: 10.1134/S2079970520010104.

[27] T. Maslova, "Transformation of consumer behavior in the tourism industry in the conditions of digital economy", IOP Conference Series: Materials Science and Engineering, 2020. doi: 10.1088/1757-899X/940/1/012070.

[28] X. Wang, X. Li, F. Zhen, and J. H. Zhang, "How smart is your tourist attraction?: Measuring tourist preferences of smart tourism attractions via a FCEM-AHP and IPA approach", Tourism Management, 54, 2016, pp. 309–320. doi: 10.1016/j.tourman.2015.12.003.

[29] Y. Li, C. Hu, C. Huang, and L. Duan, "The concept of smart tourism in the context of tourism information services", Tourism Management, 58, 2017, pp. 293–300. doi: 10.1016/j.tourman.2016.03.014.

[30] D. Ramadhani, "Tourism destination rating system based on social media analysis (proposal and dataset development in Indonesian language)", Proceedings - 2017 International Conference on Sustainable Information Engineering and Technology, SIET 2017, 2018, pp. 41–46. doi: 10.1109/SIET.2017.8304106.

[31] R. Amanda, "Analysis of Tourists Preferences on Smart Tourism in Yogyakarta (Case: Vredeburg Fort Museum)", Journal of Physics: Conference Series, 2018. doi: 10.1088/1742-6596/1007/1/012040.

[32] Y. Pei, and Y. Zhang, "A Study on the Integrated Development of Artificial Intelligence and Tourism from the Perspective of Smart Tourism", Journal of Physics: Conference Series, 1852(3), 2021. doi: 10.1088/1742-6596/1852/3/032016.

[33] S. Dey, "Analytical study on use of AI techniques in tourism sector for smarter customer experience management", 2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020, 2020. doi: 10.1109/ICCSEA49143.2020.9132925.

[34] L. Butgereit, "On Safari with TensorFlow: Assisting Tourism in Rural Southern Africa Using Machine Learning", 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD 2018, 2018. doi: 10.1109/ICABCD.2018.8465441.

[35] A. Khatibi, "FISETIO: A FIne-grained, Structured and Enriched Tourism Dataset for Indoor and Outdoor attractions', Data in Brief, 28, 2020. doi: 10.1016/j.dib.2019.104906.

[36] X. Li, "Machine Learning in Internet Search Query Selection for Tourism Forecasting", Journal of Travel Research, 2020. doi: 10.1177/0047287520934871.

[37] T. H. Le, "Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality", Current Issues in Tourism, 2020. doi: 10.1080/13683500.2020.1829568.

[38] L. Huang, "Enhancing the smart tourism experience for people with visual impairments by gamified application approach through needs analysis in Hong Kong", Sustainability (Switzerland), 12(15), 2020. doi: 10.3390/su12156213.

[39] P. Lee, "Smart tourism city: Developments and transformations', Sustainability (Switzerland), 12(10), 2020. doi: 10.3390/SU12103958.

[40] C. K. Pai, "The role of perceived smart tourism technology experience for tourist satisfaction, happiness and revisit intention", Sustainability (Switzerland), 12(16), 2020. doi: 10.3390/su12166592.

[41] X. Chen, Q. Liu, and X. Qiao, "Approaching Another Tourism Recommender", Proceedings - Companion of the 2020 IEEE 20th International Conference on Software Quality, Reliability, and Security, QRS-C 2020, 2020, pp. 556–562. doi: 10.1109/QRS-C51114.2020.00097.

[42] K. V. D. Sagar, P. S. G. Arunasri, S. Sakamuri, J. Kavitha, and D. B. K. Kamesh, "Collaborative Filtering and Regression Techniques based location Travel Recommender System based on social media reviews data due to the COVID-19 Pandemic", IOP Conference Series: Materials Science and Engineering, 981(2), 2020. doi: 10.1088/1757-899X/981/2/022009.

[43] R. Roy, and L. W. Dietz, "TripRec - A recommender system for planning composite city trips based on travel mobility analysis", CEUR Workshop Proceedings, 2855, 2021, pp. 8–12.

[44] M. Figueredo, "From photos to travel itinerary: A tourism recommender system for smart tourism destination", Proceedings - IEEE 4th International Conference on Big Data Computing Service and Applications, BigDataService 2018, 2018, pp. 85–92. doi: 10.1109/BigDataService.2018.00021.

Predicting Aesthetic Preferences: Does the Big-Five Matters?

Carolyn Salimun[1], Esmadi Abu bin Abu Seman[2], Wan Nooraishya binti Wan Ahmad[3], Zaidatol Haslinda binti Abdullah Sani[4]
Faculty of Computing and Informatics
Universiti Malaysia Sabah, Jalan Sungai Pagar
87000 F.T. Labuan, Malaysia

Saman Shishehchi[5]
Imam Khomeini International University
Buin Zahra Higher Education Center of Engineering and Technology, Buin Zahra, Qazvin
3451745346, Iran

***Abstract*—User experience is imperative for the success of interactive products. User experience is notably affected by user preferences; the higher the preference, the better the user experience. The way users develop their preferences are closely related to personality traits. However, there is a void in understanding the association between personality traits and aesthetic dimensions that may potentially explain how users develop their preferences. This paper examines the relationship between the Big-Five personality traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) and the two dimensions of aesthetics (classical aesthetics, expressive aesthetics). Two hundred twenty participants completed the Big-Five questionnaire and rated their preference for each of the ten images of web pages on a 7-point Likert scale. Results show Openness to Experience, Conscientiousness, Extraversion, and Neuroticism were not significantly correlated with aesthetic dimensions. Only Agreeableness showed a significant correlation (although weakly) with both classical and expressive aesthetics. The finding conforms to literature that personality traits have influence on the preference of individual design features in lieu of aesthetic dimensions. In other words, personality traits are inapt predictor of aesthetic dimension. Therefore, more studies are needed to explore other factors that potentially help to predict aesthetic dimensions.**

Keywords*—User experience; aesthetic dimensions; personality traits; big-five

I. Introduction

User preferences play a significant role in improving the quality of user experience; the higher the preference, the better the user experience [1][2][3]. Many factors can influence user preferences, and among the most important factors is interface aesthetics [4][5][6]. The author in [7] suggested in their influential work on the dimensionality of aesthetics that aesthetics in interface design consisted of two dimensions: classical aesthetics and expressive aesthetics. Classical aesthetics emphasise clarity, orderly, symmetrical and clean design and is closely related to many of the design rules advocated by usability experts. Conversely, expressive aesthetics accentuate creativity, originality, exquisiteness, and the ability to go against design conventions. The aesthetic dimensions of classical and expressive aesthetics have been discussed thoroughly in the literature. However, little is known about which of these two aesthetic dimensions highly correspond with user preferences.

According to [8][9][10][11], user preference is influenced by overall aesthetics during the initiation stage and followed by individual design features on the latter. Wood and Keller [11] explained that preference is the result of how the visual system organises and groups the incoming information. This conclusion corroborates [9]'s finding, where only 50 milliseconds are required to react to overall aesthetics. This spontaneous reaction time was found to be consistent throughout the system utilisation by [10], who studied the consistency of immediate aesthetic perceptions. Furthermore, [8]'s work on web usability using eye-tracking concludes that users looked at the overall design first, followed by specific design features afterwards. Therefore, aesthetics is arguably the by-product of the overall design effect rather than the details comprising it.

Despite the notion that aesthetics is a result of the overall design, several works report that aesthetic preferences are closely related to personality traits through the users' mental model [12]. The mental model explains how users simplify the complexity and details of external reality into a proportional representation applicable for decision making. This process of representation reduces individual details to fewer relevant entities with relationships between them, useful for decision making at hand. Since personality is integrated within the mental model, personality traits are likely to affect users' preferences.

User preferences over interface design can at some point be predicted by personality traits, where certain personality traits trigger preferences for specific design features (e.g. buttons, font, icons, information density, menu structure, navigation, theme, etc.) [13][14][15][16][17][18][19][20]. Personality traits as classified by the big-five also known as Five-Factor Inventory (FFI), categorised human characteristic patterns into five broad dimensions, represented by the acronym OCEAN or CANOE: Openness to Experience (O) or sometimes abbreviated as Openness, Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N) [21]. Openness to experience is defined as the tendency to actively seek and appreciate new experiences and tolerate novel situations. This personality trait is manifested by curiosity, imagination, and being untraditional [22][23][24].

Conscientiousness is defined as the tendency to be cautious, consistent, and organised; manifested by self-regulation and adherence to norms [25] [26][27]. Extraversion is defined as the tendency to be extrovert, manifested by assertiveness, talkativeness, and optimism [28][29] [30]. Agreeableness is defined as altruistic, tolerant, and trustworthy. This personality trait exhibited gullible, meek, and selfless behaviours and other prosocial behaviours [31][32][33]. Finally, Neuroticism is defined by the tendency to experience negative emotions exhibited through anxious, depressive, and insecure behaviours [34][35].

While previous research findings confirmed the relationship between personality traits and users' preferences, there is limited work that studies the association between personality traits and aesthetic dimensions. Most of the research studies focused on specific design features, neglecting the aesthetic dimensions. This research gap raises a question of whether personality traits affect user preference on aesthetic dimensions as it does with specific design features. Addressing this gap will present an opportunity to explore the predictive role of personality traits in determining which dimension of aesthetics has a strong influence on user preference. Therefore, this study was undertaken to fill the gap by analysing the correlation between the big-five personality traits and users' preferences over screenshots of web pages (homepage) representing classical and expressive aesthetics.

The rest of this paper is organized as follows. Section 2 discusses related works associated with the big-five personality traits and aesthetic preferences. Section 3 describes the methodology used in this study. Sections 4 and 5 present the finding and discussion, respectively. The conclusions are given in Section 6.

II. Related Work

This section provides some references to previous work related to the big-five personality traits and aesthetic preferences.

According to [36], visual aesthetic sensitivity is independent of personality whilst others indicated that people high in Conscientiousness, Extraversion, and Agreeableness preferred common elements of design, such as the contrast between background and text, straightforward information display with the typical organisation of menu bar, scroll bar, and buttons [13][37][38]. These design elements are literally the execution of affordance and design convention. Affordance solicits users' actions without mental effort, whilst design convention influences users' expectations in the absence of affordance. Deviating from such intuition-centric would cause confusion and frustration for people high in Conscientiousness. People high in Conscientiousness further demonstrated positive inclinations toward conventional, clean, and orderly interface design features [39][40] and detest intense design style; complex and unconventional designs with irregular shapes of features [41].

Despite the fact that Openness to Experience has been widely examined, only a few empirical studies address its locus in design preference [12]. The author in [42] reported that people with high Openness to Experience tend to focus and adapt more inquisitively to an unconventional style of interface. Their preference for imaginative, untraditional, and personalised features allows them to quickly adapt to a new interface style. They relatively are not concerned with convention and are eager to explore new design features [43]. The consistent findings on unconventional design elements may explain the disposition towards expressive design in people with high Openness to Experience. However, unlike people high in Extraversion, people with high Openness to Experience are generally deterred by persuasive designs. Fear and stress are the reasons that discourage them from persuasive design [44].

Unlike Openness to Experience, which desires uniqueness, Extraversion desires extravagance, that is, the extreme continuum of design [45]. According to [46], Extraversion correlates with the desire for status, leading to a preference for extravagant designs. This is induced by the assertiveness trait in people high in Extraversion. People high in Extraversion prefer high colour contrasts, saturated hues, and bold or sharp-edged shapes of the graphical interface [13][20]. Concerning utilisation of the adaptive interface, [47] revealed that the approach does not benefit Extraversion. In other words, people high in Extraversion do not respond well to monotonous interface set up by adaptation design.

On the other hand, people high in Agreeableness are relatively easy to accept any design presented to them [48]. They tend to be more receptive to design that generates a sense of certitude in relation to their personality traits defined by altruistic, tolerant, and trustworthiness. According to [49], Agreeableness is generally considered as an adaptive trait and correlates with the preference of authority figures. This is evident in a study to explore how personality features affect compliance towards recommendations, where they found that people high in Agreeableness follow the editor's suggestions. In other words, people high in Agreeableness cope with individuals who have reliable and verifiable qualities in related disciplines [50][51]. This might explain why they are easily more satisfied aesthetically compared to other traits [48]. Nevertheless, [52] revealed that people high in Agreeableness have a high preference for classical, representational art and less preference for abstract art.

Compared to the rest of the traits, Neuroticism is strongly associated with sensitive, obsessive, and anxious characters. Neurotic individuals exhibit an inclination towards asynchronous media as it enables higher situational control and avoids direct interactions with others [53][54]. They are concerned with the visual clarity and readability of interface design [13][15]. A study by [44] reported that people high in Neuroticism have a low preference toward designs that utilise persuasive strategies (i.e., competition, simulation, self-monitoring and feedback, goal setting and suggestion, customisation, reward, social comparison, cooperation, punishment, and personalisation). These findings imply that people high in Neuroticism respond poorly to unconventional design features. Considering the immense magnitude of paranoia embedded in these individuals, it is very unlikely that they will advocate for expressive aesthetics.

In general, the five personality traits of the big-five; Openness to experience, Cautiousness, Extraversion, Agreeableness, and Neuroticism, have unique characteristics that can influence their preferences towards classical or expressive aesthetics of interface design. Both Openness to experience and Extraversion individuals have shown novelty-seeking behaviour that corresponds to unconventional design, seemingly reflecting preference over expressive aesthetics. On the contrary, Conscientiousness and Neuroticism individuals have exhibited non-receptive behaviour towards novel-centric design concurrently with a positive response to design convention. These inclinations are apparently in agreement with an exposition of classical aesthetics. In comparison to other personality traits, Agreeableness individuals inhabit disparate behaviour that is highly flexible to both conventional and non-conventional design. The absence of preference over particular design categories implicates that Agreeableness individuals are receptive to both classical and expressive aesthetics.

The unique characteristics of each personality trait and their preference for specific design features, as discussed above, are useful to predict aesthetic dimensions. However, empirical evidence to support the accuracy of the prediction has been lacking. This study aims to provide an empirical evidence to confirm the predictive role of personality traits in predicting preference for aesthetic dimensions.

III. Method

The study consisted of two phases. Phase 1 was designed to classify website homepages into classical and expressive aesthetics. The selected web pages in phase 1 were used as stimuli in phase 2 to examine the relationship between aesthetic preferences and personality traits.

A. Phase 1

1) Participants: 128 undergraduate students of Universiti Malaysia Sabah voluntarily participated in the study. Of the 128 students, 98 (77%) were students of the Faculty of Computing and Informatics, and 30 (23%) were students of the Faculty of International Business and Finance. 86 (67%) of the participants were female, and 42 (33%) were male, with ages ranging from 19 to 26 years (Mean=22.69, SD=1.515). 46 (36%) participants identified themselves as Chinese, 34 (27%) as Native Borneo, 31 (24%) as Malay, 16 (12%) as Indian, and one (1%) as others. The majority of the participants (65 (51%)) spent more than 8 hours per day on computer/ internet, while 43 (34%) spent 5-8 hours per day, 17 (13%) spent 2-5 hours per day, and the remaining 3 (2%) spent less than two hours per day.

2) Instrument: An online questionnaire administered via Google form and advertised through Facebook group pages and WhatsApp served as the main instrument of the experiment in phase 1. The questionnaire was composed of three sections. Section 1 consisted of the consent form and instruction of the experiment. Section 2 was the demographic questions on age, gender, ethnicity, faculty, and duration of computer or internet use per day. Section 3 begins with definitions of classical and expressive aesthetics, followed by 30 screenshots of web pages, each accompanied by a 7-point Likert scale (1, classical aesthetics; 7, expressive aesthetics). The 30 screenshots were arranged vertically, one above another and presented in a different random order for each participant to avoid the order effect.

3) Task and procedure: The participants were required to complete all three sections of the questionnaire, starting from Sections 1 to 3. In Section 1, the participants were required to read the instruction and give their consent to participate. If the participants consented, they can click on the “Next” button to proceed to the next section. Otherwise, they can click on a “Cancel” button to withdraw from the experiment. In Section 2, the participants were required to answer five demographic questions related to age, gender, ethnicity, faculty, and the duration of computer/internet use per day. Upon completing Section 2, the participants proceeded to Section 3 by clicking the “Next” button. In Section 3, the participants were required to rate their perceived aesthetic dimension on each of the 30 screenshots of web pages on a 7-point Likert scale (1-classical aesthetics, 7-expressive aesthetics). A definition of classical and expressive aesthetics was provided to ensure that all participants understood the term and rated the screenshots based on the same definition. The experiment ended when the participants click the “Submit” button.

4) Measure: The ratings of each web page were summed and ordered from lowest to highest score. Five web pages with the lowest scores and five with the highest scores were selected, and the remaining 20 were discarded to create a proper gap between the lowest and the highest. The five web pages with the lowest scores and five with the highest scores were classified as classical and expressive aesthetics, respectively, and used as stimuli in phase 2.

B. Phase 2

1) Participants: A total of 220 undergraduate students of Universiti Malaysia Sabah voluntarily participated in the study. 153 of the total participants were students of the Faculty of Computing and Informatics, and 67 were International Business and Finance students. In terms of gender distribution, 70% (146) were females, and 30% (71) were males with ages ranging from 20 to 27 (Mean=22.43, SD=1.41). Ethnicity wise, 41% (91) participants identified themselves as Native Borneo, 30% (66) as Chinese, 20% (44) as Malay, 7% (16) as Indian, and 1% (3) as others. In terms the duration of computer/ internet use per day, 53% (116) of the participants reported that they spent more than 8 hours a day, 23% (50) spent 5 – 8 hours, 16% (36) spent 3 – 5 hours, 7% (16) spent 2 - 3 hours, and 1% (2) spent less than an hour.

2) Instrument: An online questionnaire administered via Google form and advertised through Facebook group pages and WhatsApp served as the main instrument of this study. The questionnaire was composed of four sections. At the bottom page of Sections 1, 2 and 3, there was a “Next” button to move to the next section, whereas in Section 4, there was a

"Submit" button to finish the questionnaire. The first two sections were identical to the first two sections in phase 1. The third section consisted of 10 items of the Big Five Inventory (BFI-10) [51] (Table I). Each item of the BFI-10 was accompanied by a 5-point scale ranging from 1 (disagree strongly) to 5 (agree strongly).

TABLE I. BFI-10 [55]

I see myself as someone who ...	Disagree strongly	Disagree a little	Neither disagree or agree	Agree a little	Agree strongly
1)... is reserved	1	2	3	4	5
2)... is generally trusting	1	2	3	4	5
3)... tends to be lazy	1	2	3	4	5
4)... is relaxed, handles stress well	1	2	3	4	5
5)... has few artistic interests	1	2	3	4	5
6)... is outgoing, sociable	1	2	3	4	5
7)... tends to find fault with others	1	2	3	4	5
8)... does a thorough job	1	2	3	4	5
9)... gets nervous easily	1	2	3	4	5
10)... has an active imagination	1	2	3	4	5

The fourth section contained 10 screenshots of web pages (see Appendix) derived from phase 1, five screenshots each for classical and expressive aesthetics. Each screenshot was presented with a 7-point Likert scale, anchored by 1 (I don't like it) and 7 (I like it a lot). The ten screenshots were arranged vertically, one above another and presented in a different random order for each participant to avoid the order effect.

3) Task and procedure: The participants were required to complete all four sections of the questionnaire, starting from Sections 1 to 4. Upon completing Section 1, the participants can move to the next section by clicking on the "Next" button. This process continued until all sections were completed. The experiment ended when the participants clicked on the "submit" button in Section 4.

The task and procedure of sections 1 and 2 were similar to that of Sections 1 and 2 in phase 1. In Section 3, the participants were asked to answer all 10 items of BFI-10 by giving a rating from 1 (disagree strongly) to 5 (agree strongly) to each item. In Section 4, the participants were asked to indicate the extent of their preference from 1 (I don't like it) to 7 (I like it a lot) for each of the 10 screenshots of web pages.

4) Measures: Personality traits were measured using BFI-10 [55]. BFI-10 is a shorter version of the well-established BFI-44 [56]. Compared to BFI-44, which consisted of 44 items with 8 to 10 items per trait, BFI-10 consisted of only 10 items with two items per trait (Table I). Although short, this shorter version of BFI-44 claimed to predict almost 70% of the variance of the scales [57]. Each item of BFI-10 was rated from 1 (disagree strongly) to 5 (agree strongly). Scale score was obtained by the average rating of 2 items where 1 item was reversed-scored (Extraversion: 1R, 6; Agreeableness: 2, 7R; Conscientiousness: 3R, 8; Neuroticism: 4R, 9; Openness to experience: 5R, 10, R=item is reversed-scored). The internal consistency reliability of the scales was not reported in this study. A small number of items always yielded inadequate internal consistency reliability; thus, internal consistency is not an adequate measure for BFI-10 [58]. Preference for classical and expressive aesthetics was measured using a 7-point Likert scale; 1 reflects the least preference, and 7 reflects the greatest preference. Fig. 1 shows the hypothetical model of this study.

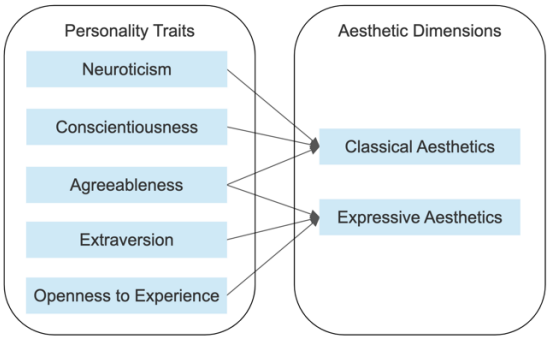


Fig. 1. Hypothetical Path Model.

IV. Result

This study used SPSS (version 26) for Windows to perform the statistical analysis of the questionnaire data. The statistical analysis used, including Pearson's correlation and multiple regression analysis with the stepwise method, to identify a possible association between aesthetic preferences and personality traits. Table II shows the overall results of this study.

TABLE II. Correlation Matrix

Personality / Dimension	μ	σ	Classical aesthetics		Expressive aesthetics	
			r	α	*r*	α
O	3.1000	.5564	-.029	.664	-.118	.081
C	3.0068	.7207	-.010	.883	.075	.265
E	2.8250	.6908	-.041	.549	.075	.271
A	3.5614	.7132	**.161***	.017	**.133***	.049
N	3.1318	.8335	-.042	.537	-.056	.407
Classical aesthetics	3.9291	1.0128				
Expressive aesthetics	5.2327	.7938				

*Correlation is significant at the 0.05 level (2-tailed).

The result of the Pearson product-moment correlation showed that Agreeableness was significantly correlated, albeit weak, with classical aesthetics (r=.161, p=.017, n=220) and expressive aesthetics (r=.133, p=.049, n=220) (Fig. 2). The

strength of correlation of Agreeableness on classical aesthetics was slightly higher than on expressive aesthetics; but clearly the numbers are too small to make the differences meaningful. Other personality traits (i.e., Openness to Experience, Conscientiousness, Extraversion, Neuroticism), however, were found not significantly correlated with either classical or expressive aesthetics (Fig. 2).

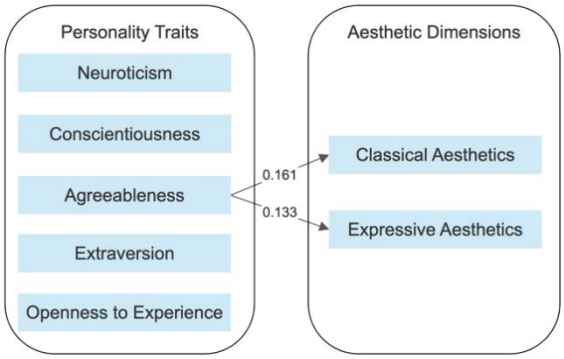


Fig. 2. Data-driven Path Model.

The result of a series of stepwise linear regression analyses with preferences as criteria showed that Agreeableness was the only personality trait that survived the stepwise procedure for classical and expressive aesthetics. For classical aesthetics, a significant regression equation was found ($F(1, 218) = 5.794$, $p<0.05$) with an R^2 of .026. Participants' predicted classical aesthetics is equal to 3.115 + .228 (Agreeableness) when Agreeableness is measured on a Likert scale. Classical aesthetics increased .228 for each point of Agreeableness. For expressive aesthetics, a significant regression equation was found ($F(1, 218) = 3.909$, $p<0.05$) with an R^2 of .018. Participants' predicted expressive aesthetics is equal to 4.707 + .148 (Agreeableness) when Agreeableness is measured on a Likert scale. Expressive aesthetics increased .148 for each point of Agreeableness.

The personality traits' contribution to being able to predict preferences for aesthetics of different dimensions was relatively low, with all predictors together accounting for only 2.8% of the variance for the classical aesthetics and 4.2% of the variance for the expressive aesthetics. As in the case of aesthetic preference, Agreeableness was shown as the most important characteristic of the big-five personality traits. Agreeableness was predictive of classical and expressive aesthetics whereas the others were not.

V. Discussion

The study demonstrates that among the big-five personality traits, Agreeableness was the only trait that significantly, albeit weakly, correlated with classical and expressive aesthetics. A closer look at the correlation of Agreeableness on classical and expressive aesthetics revealed that the strength of correlation between Agreeableness and classical aesthetics (r =.161) was slightly higher than between Agreeableness and expressive aesthetics (r =.133). This result can be interpreted in two ways. First, people with Agreeableness prefer classical aesthetics over expressive aesthetics. This interpretation is in line with [52], who conducted a study in visual art and found that people with Agreeableness prefer representational arts over abstract arts. Second, people with Agreeableness prefer both classical and expressive aesthetics. This interpretation is based on the trivial difference in the correlation coefficient between classical and expressive aesthetics (i.e., 0.028) and is aligned with our hypothesis that people with Agreeableness are receptive to any form of aesthetics presented to them.

Conversely, the absence of a significant correlation of Openness to Experience, Conscientiousness, Extraversion, and Neuroticism on both classical and expressive aesthetics implies that these four personality traits have no influence over user preference on aesthetic dimensions. This finding correlated well with [36], who claimed that visual aesthetic sensitivity was independent of personality. Our work extends their study by employing a website interface with all adult subjects (21-27 years) in comparison to using artwork as stimuli with mostly children (10-15 years) in assessing the independence of visual aesthetic sensitivity from personality traits. This finding adds to the literature suggesting that aesthetic preference is independent from personality across computer or non-computer interface and age groups (i.e. children and young adults).

Apart from the immediate results, the findings of this study should also be considered within the context of its limitations. As in most empirical studies, the study conducted here was limited by the measure used to examine user preferences. Since user preferences were measured using static screenshots of web pages, the results may only apply to non-interactive interfaces, whereas the use of interactive interfaces as stimuli may potentially alter the results. This is an interesting aspect to be carried out in prospective work.

In addition, the subjects in this study were undergraduate university students. This restricted sample of subjects may hinder the generalizability of the findings across the populations. The opportunity to evaluate other populations may provide insight into interpopulation perspectives of user preference. Replicating findings in different contexts and with different populations may possibly benefit the knowledge-building process, theoretical refinement, and applicability in other situations [59][60].

VI. Conclusion

This study examined the relationship between personality traits and aesthetic dimensions. The results showed a weak relationship between personality traits and aesthetic dimensions, thus suggesting that personality traits are not a good predictor of aesthetic dimension preferences. Previous literature and our findings collectively illuminate the fact that personality measures appeared more useful in predicting aesthetic preference of individual design features than the aesthetic dimension of interface. In the future, other than using different stimuli and populations, further study should consider inductive research design to gain a qualitative understanding of underlying reasons and motivations for aesthetic preferences.

References

[1] Shirole, A. Chowdhury, and D. Dhar, "Identification of Aesthetically Favourable Interface Attributes for Better User Experience of Social

Networking Application," in Ergonomics in Caring for People, Springer, 2018, pp. 251–259.

[2] M. Simões-Marques, A. Correia, M. F. Teodoro, and I. L. Nunes, "Empirical studies in user experience of an emergency management system," in International Conference on Applied Human Factors and Ergonomics, 2017, pp. 97–108.

[3] S. L. T. Hui and S. L. See, "Enhancing user experience through customisation of UI design," Procedia Manuf., vol. 3, pp. 1932–1937, 2015.

[4] S. Liu, T. Liang, S. Shao, and J. Kong, "Evaluating localized MOOCs: The role of culture on interface design and user experience," IEEE Access, vol. 8, pp. 107927–107940, 2020.

[5] S. Lee and R. J. Koubek, "Understanding user preferences based on usability and aesthetics before and after actual use," Interact. Comput., vol. 22, no. 6, pp. 530–543, 2010.

[6] D. A. Norman, "Emotion & design: attractive things work better," interactions, vol. 9, no. 4, pp. 36–42, 2002.

[7] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," Int. J. Hum. Comput. Stud., vol. 60, no. 3, pp. 269–298, 2004.

[8] J. Nielsen and K. Pernice, Eyetracking web usability. New Riders, 2010.

[9] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!," Behav. Inf. Technol., vol. 25, no. 2, pp. 115–126, 2006.

[10] N. Tractinsky, A. Cokhavi, M. Kirschenbaum, and T. Sharfi, "Evaluating the consistency of immediate aesthetic perceptions of web pages," Int. J. Hum. Comput. Stud., vol. 64, no. 11, pp. 1071–1083, 2006.

[11] C. H. Wood and C. P. Keller, Cartographic design: Theoretical and practical perspectives. Wiley Chichester, UK, 1996.

[12] T. Alves, J. Natálio, J. Henriques-Calado, and S. Gama, "Incorporating personality in user interface design: A review," Pers. Individ. Dif., vol. 155, p. 109709, 2020.

[13] S. M. Sarsam and H. Al-Samarraie, "Towards incorporating personality into the design of an interface: a method for facilitating users' interaction with the display," User Model. User-adapt. Interact., vol. 28, no. 1, pp. 75–96, 2018.

[14] R. Aktivia, T. Djatna, and Y. Nurhadryani, "Visual usability design for mobile application based on user personality," in 2014 International Conference on Advanced Computer Science and Information System, 2014, pp. 177–182.

[15] L. Arockiam and J. C. Selvaraj, "User interface design for effective e-learning based on personality traits," Int. J. Comput. Appl., vol. 61, no. 14, 2013.

[16] J. Kim, A. Lee, and H. Ryu, "Personality and its effects on learning performance: Design guidelines for an adaptive e-learning system based on a user model," Int. J. Ind. Ergon., vol. 43, no. 5, pp. 450–461, 2013.

[17] A. D. Shaikh, B. S. Chaparro, and D. Fox, "Perception of fonts: Perceived personality traits and uses," Usability news, vol. 8, no. 1, pp. 1–6, 2006.

[18] B. Saati, M. Salem, and W.-P. Brinkman, "Towards customized user interface skins: investigating user personality and skin colour," Proc. HCI 2005, vol. 2, pp. 89–93, 2005.

[19] E. Abrahamian, J. Weinberg, M. Grady, and C. M. Stanton, "The effect of personality-aware computer-human interfaces on learning," J. Univers. Comput. Sci., vol. 10, no. 1, pp. 27–37, 2004.

[20] A. Karsvall, "Personality preferences in graphical interface design," in ACM International Conference Proceeding Series, 2002, vol. 31, pp. 217–218, doi: 10.1145/572020.572049.

[21] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure.," J. Pers. Soc. Psychol., vol. 59, no. 6, p. 1216, 1990.

[22] P. J. Silvia and A. P. Christensen, "Looking up at the curious personality: Individual differences in curiosity and Openness to Experience," Curr. Opin. Behav. Sci., vol. 35, pp. 1–6, 2020.

[23] M. Ahmad and Z. Maochun, "Personality traits and investor decisions," Asian J. Econ. Financ. Manag., pp. 19–34, 2019.

[24] R. R. McCrae and P. T. Costa Jr, "Openness to experience," Perspect. Personal., vol. 1, pp. 145–172, 1985.

[25] L. Zhang and H. Liu, "Effects of Conscientiousness on Users' Eye-Movement Behaviour with Recommender Interfaces," in International Conference on Human-Computer Interaction, 2019, pp. 367–376.

[26] U. J. Wiersma and R. Kappe, "Selecting for extroversion but rewarding for conscientiousness," Eur. J. Work Organ. Psychol., vol. 26, no. 2, pp. 314–323, 2017.

[27] H. S. Friedman and M. L. Kern, "Personality, well-being, and health," Annu. Rev. Psychol., vol. 65, pp. 719–742, 2014.

[28] A. Spark and P. J. O'Connor, "State extraversion and emergent leadership: Do introverts emerge as leaders when they act like extraverts?," Leadersh. Q., vol. 32, no. 3, p. 101474, 2021.

[29] L. Terrier, S. Kim, and S. Fernandez, "Who are the good organizational citizens for the environment? An examination of the predictive validity of personality traits," J. Environ. Psychol., vol. 48, pp. 185–190, 2016.

[30] R. R. McCrae and P. T. Costa Jr, "Discriminant validity of NEO-PIR facet scales," Educ. Psychol. Meas., vol. 52, no. 1, pp. 229–237, 1992.

[31] M. T. Argan and M. Argan, "Do altruistic values of an individual reflect personality traits," Int. J. Recent Adv. Organ. Behav. Decis. Sci. An Online Int. Res. J., vol. 3, no. 1, pp. 858–871, 2017.

[32] B. Osatuyi, "Personality traits and information privacy concern on social media platforms," J. Comput. Inf. Syst., vol. 55, no. 4, pp. 11–19, 2015.

[33] N. Capuano, G. D'Aniello, A. Gaeta, and S. Miranda, "A personality based adaptive approach for information systems," Comput. Human Behav., vol. 44, pp. 156–165, 2015.

[34] S. Sauer-Zavala, J. G. Wilner, and D. H. Barlow, "Addressing neuroticism in psychological treatment.," Personal. Disord. Theory, Res. Treat., vol. 8, no. 3, p. 191, 2017.

[35] C. Arslan, "Interpersonal problem solving, self-compassion and personality traits in university students," Educ. Res. Rev., vol. 11, no. 7, pp. 474–481, 2016.

[36] J. P. Fróis and H. J. Eysenck, "The Visual Aesthetic Sensitivity Test applied to Portuguese children and fine arts students," Creat. Res. J., vol. 8, no. 3, pp. 277–284, 1995.

[37] C. M. R. Leung, "Human factors in website usability and aesthetics: theory and applications to hotels in Hong Kong," 2013.

[38] R. Leung, J. Rong, G. Li, and R. Law, "An analysis on human personality and hotel web design: A Kohonen network approach," in ENTER, 2011, pp. 573–585.

[39] J. Zhang, Y. Luximon, and Y. Song, "The role of consumers' perceived security, perceived control, interface design features, and conscientiousness in continuous use of mobile payment services," Sustainability, vol. 11, no. 23, p. 6843, 2019.

[40] E. Geisler-Brenstein and R. R. Schmeck, "The revised inventory of learning processes: A multifaceted perspective on individual differences in learning," in Alternatives in assessment of achievements, learning processes and prior knowledge, Springer, 1996, pp. 283–317.

[41] K. Cleridou and A. Furnham, "Personality correlates of aesthetic preferences for art, architecture, and music," Empir. Stud. Arts, vol. 32, no. 2, pp. 231–255, 2014.

[42] R. G. Saadé, D. Kira, F. Nebebe, and C. Otrakji, "Openness to Experience: An HCI Experiment.," Issues informing Sci. Inf. Technol., vol. 3, 2006.

[43] P. Kortum and F. L. Oswald, "The impact of personality on the subjective assessment of usability," Int. J. Human–Computer Interact., vol. 34, no. 2, pp. 177–186, 2018.

[44] R. Orji, L. E. Nacke, and C. Di Marco, "Towards personality-driven persuasive health games and gamified systems," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 1015–1027.

[45] D. R. Dunaetz, T. C. Lisk, and M. M. Shin, "Personality, gender, and age as predictors of media richness preference," Adv. Multimed., vol. 2015, 2015.

[46] D. Greenberg, E. Ehrensperger, M. Schulte-Mecklenbeck, W. D. Hoyer, Z. J. Zhang, and H. Krohmer, "The role of brand prominence and extravagance of product design in luxury brand building: What drives

consumers' preferences for loud versus quiet luxury?," J. Brand Manag., vol. 27, no. 2, pp. 195–210, 2020.

[47] K. Z. Gajos and K. Chauncey, "The influence of personality traits and cognitive load on the use of adaptive user interfaces," in Proceedings of the 22nd International Conference on Intelligent User Interfaces, 2017, pp. 301–306.

[48] K. Oyibo, R. Orji, and J. Vassileva, "The influence of personality on mobile web credibility," in Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, 2017, pp. 53–58.

[49] P. Unal, T. T. Temizel, and P. E. Eren, "What installed mobile applications tell about their owners and how they affect users' download behavior," Telemat. Informatics, vol. 34, no. 7, pp. 1153–1165, 2017.

[50] N. S. A. Karim, N. H. A. Zamzuri, and Y. M. Nor, "Exploring the relationship between Internet ethics in university students and the big five model of personality," Comput. Educ., vol. 53, no. 1, pp. 86–93, 2009.

[51] H. Zhao and S. E. Seibert, "The big five personality dimensions and entrepreneurial status: A meta-analytical review.," J. Appl. Psychol., vol. 91, no. 2, p. 259, 2006.

[52] A. Furnham and M. Avison, "Personality and preference for surreal paintings," Pers. Individ. Dif., vol. 23, no. 6, pp. 923–935, 1997.

[53] R. Leung, J. Rong, G. Li, and R. Law, "Personality differences and hotel web design study using targeted positive and negative association rule mining," J. Hosp. Mark. Manag., vol. 22, no. 7, pp. 701–727, 2013.

[54] G. Hertel, J. Schroer, B. Batinic, and S. Naumann, "Do shy people prefer to send e-mail? Personality effects on communication media preferences in threatening and nonthreatening situations," Soc. Psychol. (Gott)., vol. 39, no. 4, pp. 231–243, 2008.

[55] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," J. Res. Pers., vol. 41, no. 1, pp. 203–212, 2007.

[56] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," J. Pers. Soc. Psychol., 1991.

[57] B. A. Balgiu, "The psychometric properties of the Big Five inventory-10 (BFI-10) including correlations with subjective and psychological well-being," Glob. J. Psychol. Res. New Trends Issues, vol. 8, no. 2, pp. 61–69, 2018.

[58] B. Spinath, H. H. Freudenthaler, and A. C. Neubauer, "Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation," Pers. Individ. Dif., vol. 48, no. 4, pp. 481–486, 2010.

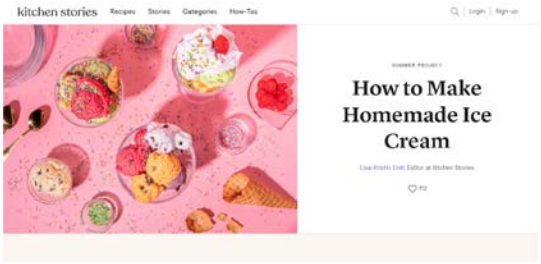
[59] C. J. Coulton, "The need for replication in social work research," in Social Work Research and Abstracts, 1982, vol. 18, no. 2, p. 2.

[60] R. A. Bettis, C. E. Helfat, and J. M. Shaver, "The necessity, logic, and forms of replication," Strateg. Manag. J., vol. 37, no. 11, pp. 2193–2203, 2016.

APPENDIX: SAMPLE OF WEB PAGES THAT WERE USED IN THE STUDY

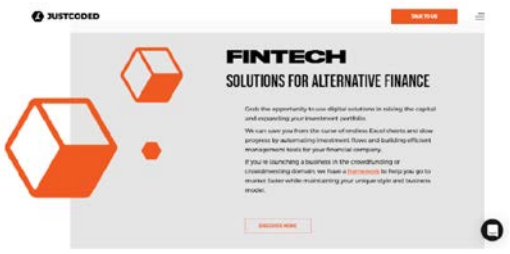
Source: http://www.meomi.com

Source: https://html5readiness.com



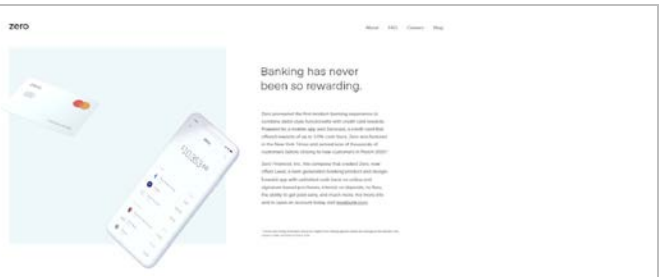
Source: https://www.kitchenstories.com

Source: https://temperrestaurant.com



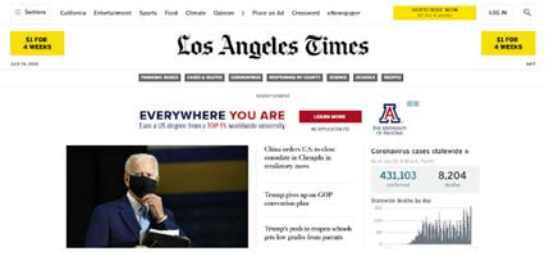
Source: https://justcoded.com

Source: https://www.grannyssecret.com/



Source: https://www.webdesignersacademy.com/inspiration/zero-app/

Source: http://www.fortherecord.simonfosterdesign.com/



Source: https://www.latimes.com

An Ontology-based Decision Support System for Multi-objective Prediction Tasks

Touria Hamim, Faouzia Benabbou, Nawal Sael
Laboratory of Modeling and Information Technology
Hassan II University, Faculty of Sciences Ben M'sik, Casablanca, Morocco

***Abstract*—Student profile modeling is a topic that continues to attract the interest of both academics and researchers because of its crucial role in the development of predictive or decision support systems. It provides platforms to build intelligent systems such as e-orientation, e-recruitment, recommendation, and prediction systems. The purpose of this research is to propose an ontology-based decision support system that can be used for multi-objective prediction tasks such as prediction of failure/abundance, orientation or decision-making. Two major contributions are proposed here: a new domain ontology that models the profile of a student and a system that is based on this ontology to perform multiple prediction tasks. The proposed approach relies on the efficiency of the ontology to ensure semantic interoperability and the benefits of machine learning techniques to build an intelligent system for a multipurpose decision support objectives. The proposed system uses Decision Tree algorithm (C5.0), but other machine learning models can be added if they prove to be more efficient. Furthermore, the performance of the developed method is computed using performance metrics and achieved 83.6% for accuracy and 81.9% for recall.**

***Keywords*—Profile modeling; student; ontology; machine learning; academic domain**

I. Introduction

In the educational field, student can be described over different information that changes over time and which constantly evolves. The profile model is the way to represent and cover the different dimensions describing accurately different features of the student such as personal, academic, social, psychological information and some others [1]. The reliability or quality of the profile description is very crucial to have efficient understanding of students. Profile modelling has the advantage of encompassing several aspects of the student that can be exploited for different purposes such as course recommendation, orientation, outcome prediction, recruitment, etc.

The education sector is a great field of promise for the uses of artificial intelligence. Artificial intelligence (AI) has the capacity to meet some of the greatest challenges facing the field of education today to develop innovative teaching and learning practices. Machine Learning (ML) can be used in many ways in the field of education, either for adaptive learning which, depending on the abilities and learning mode of each student, allows to choose personalized techniques and optimized techniques to the individual scale, or for improving student performance by identifying the cause of the problem and helping to remedy it and the institutions themselves can identify their weaknesses and find areas for improvement to maximize their students' results, or for predicting student success or failure. ML techniques have also shown their power to help students choose their path, where based on the data, the system can suggest a student to work in the industry or sector that best suits him, guidance can therefore be based on these results to better guide those who are struggling to find their way. Machine learning-based systems are used to design complex models and algorithms that lend themselves to prediction or decision support. These models allow researchers, scientists, engineers and analysts to "produce reliable and repeatable decisions and results" and to uncover "hidden information" by taking advantage of historical relationships and trends in data. To learn, these systems receive huge amounts of data, which they then use to learn how to perform a specific task. The quality and size of this dataset is important to building a system capable of performing the task assigned to it with precision. Researchers turned their attention to the possibilities of standardizing the representation of knowledge. Ontology is one of the approaches that allow concepts to be represented explicitly, it determines the concepts that exist or may exist in the area of interest [2-3].

In addition to representing knowledge, a machine learning-based system must encode the knowledge into a form that can be processed efficiently. The Ontologies are currently among the most talked models in Knowledge Engineering aiming to establish representations through which machines can manipulate the semantics of information, integrate new concepts according to the evolution of the system and use different data sources. Ontologies offer knowledge sharing facilitated by the use of a common conceptualization (vocabulary and semantics) and the adoption of a standard ontological language. The integration of machine learning (ML) techniques in ontologies makes it possible to enrich and broaden the context of use, from an ontology that offers a common conceptualization, to an ontological system of decision support. Several research studies combine ontologies with machine learning techniques for different purposes [4-5]. This combination can be used for the enrichment of the data available to a Machine Learning model, and this comes from the fact that ontology, offers in addition to raw data, a whole chain of associations and relations between data.

This paper presents two major contributions. The first one is the creation of a new Generic Student Modeling ontology (GSMonto) that aims to describe several student features in different levels while remaining extensible and scalable. The second contribution is the exploitation of GSMonto to build an

ontology-based decision support framework for multi-objective tasks. The objectives are numerous such as to offer adapted learning content, to predict student failure, success or dropout, to propose student orientation, or recommendation, etc.

This article is organized as follows. In Section 2, the authors present the background concept of ontology and machine learning techniques. Section 3 is a comparative study of the different researches dealing with student profile modelling using ontologies and dealing with the integration of Machine Learning with ontologies. The proposed generic student model ontology (GSMonto) is detailed in Section 4. Section 5 develops the proposed ontology-based machine learning system for student profile. Section 6 presents an experiment of the proposed system. At last, the authors give a discussion, conclusion and perspectives.

II. Background Concept

A. Ontology

The definition of ontologies is inherited from a philosophical tradition which is concerned with the science of "Being". Today, it means the "science of beings" that is to say the set of objects recognized as existing in a domain. It is a structured set of concepts that make sense of information [6]. Its primary objective is to model a body of knowledge in a given field.

The advantage of ontologies is the separation of knowledge. The ontological knowledge being separated, it can be reused in several applications, and these re-uses (total or partial) can form the basis of interoperability between different systems. For example, integrating an ontology into a ML based system therefore makes it possible to formally declare a certain amount of knowledge used to characterize the information managed by the system and to be based on these characterizations and the formalization of their meaning in order to automate data processing tasks. Ontologies are employed as a form of representation of knowledge in Artificial Intelligence, semantic web, software engineering, biomedical domain, and information architecture. Among the constraints of the use of ontologies is their creation difficulty as well as the visualization limits and the difficulty of finding ready-made ontologies to meet user needs.

There are several knowledge representation languages, as RDF (Resource Description Framework) [7] and RDF Schema [8] which have tried to solve the problem of the absence of the semantics of XML schemas by associating simple semantics with identifiers, RDF and RDFS were designed to be as generic as possible, this simplicity of language is also accompanied by an insufficient expressiveness for the description of complex situations. The OIL (Ontology Inference Layer) [9] and DAML (DARPA Agent Markup Language) languages were developed to fill the gaps of RDF, OIL allows defining classes and relations and a limited number of axioms and DAML intervenes to allow agents to share semantics. These two languages were then merged to give a DAML+OIL language that is based on the RDF and RDF Schema languages by enriching them with new primitives [10]. Based on the DAML + OIL language, OWL was defined based on the basic primitives defined by RDF schemas. However, far from being a simple extension of RDF, OWL provides all the semantics necessary for the description of knowledge especially for publishing and sharing resources on the semantic web, structuring them in an understandable and standardized way, and making them accessible by adding Meta information. For this, OWL is chosen to lead this study, given that OWL has more powerful means of expressing meaning and semantics than XML, RDF, and RDF-S. In addition, OWL allows information to be gathered from distributed sources, including allowing the import of information from other ontologies. OWL is developed as an extension of the RDF vocabulary and is derived from the DAML + OIL ontology language [11]. OWL has three increasingly expressive sub-languages. OWL Lite is a sub-language that supports users who mainly need a classification hierarchy and simple constraints, which makes the calculation time of inference processes limited. The advantage of OWL Lite is that is both easier to grasp (for users) and easier to implement (for tool builders). The disadvantage is its restricted expressivity. OWL DL, a sub-language of OWL that supports maximum expressiveness needs while guaranteeing the completeness of calculations and decidability necessary for reasoning systems. The advantage of OWL DL is that it permits efficient reasoning support but loose full compatibility. And OWL Full sub-language which gives the user maximum expressiveness, but there is no guarantee as to the completeness and completion of the procedures inference, the advantage of OWL full is its maximum expressiveness without sacrificing computational completeness and the inconvenient is that it is so powerful in expressiveness that it became undecidable. OWL DL is the language chosen to drive the proposed approach.

B. Machine Learning Techniques

Machine learning is an artificial intelligence (AI) field that enables systems to learn and improve automatically from the experience itself without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn on their own. In general, two main types of machine learning algorithms are used today: supervised learning and unsupervised learning. In supervised Learning all data is tagged and algorithms learn to predict the outcome of the input data. Unsupervised Learning uses an unlabeled data set, the machine is then asked to create its own responses. It thus offers answers from analyses and grouping of data.

In the proposed system of this paper, Decision Tree (DT) is used, which is an algorithm that estimates a target concept by a tree representation, where each internal node corresponds to an attribute, and each terminal node (or leaf) corresponds to a class. It is widely known and used in many fields to aid the decision-making. In the academic field, it was shown in author's previous study [12], that it is counted among the algorithms that give the best performance for the prediction of the academic performance of the students. There are several automatic algorithms for building decision trees like ID3, C4.5, C5.0 and CART, etc. ID3 (Iterative Dichotomiser 3) was developed by Ross Quinlan. It can be applied only on the nominal characteristics. It is used for ranking. Therefore, if the data contains continuous characteristics, then discretization should be applied. C4.5 is an extension of ID3 by Ross Quinlan

that can be applied on all types of features. Among the improvements of C4.5 regarding ID3, is the transformation of continuous (digital) characteristics into nominal characteristics dynamically, features without values are ignored when calculating entropy and information gain and pruning trees after creation. C5.0 is a commercial and vastly improved version of the C4.5 algorithm, which applies in large databases. Among the improvements that are added in the C5.0 algorithm is the addition of new data types such as dates and the improvements in efficiency, memory and data processing speed. And CART algorithm, which is similar to that of C4.5 with a few differences like supporting regression, representing the decision tree by a series of binary divisions leading to terminal nodes which can be described by a set of specific rules. The attractive presentation of the CART tree makes it easy to interpret.

III. Related Works Review

A. Ontology Construction in Educational Domain

In several research studies, ontologies were used to master the information resources of learners, and to facilitate their organization and exploitation. Ontologies is now one of the most important bases of the Semantic Web approaches in the educational field, whether for online or traditional learning. When building a model in adaptive e-leaning, several questions arise, such as which information will serve to represent a better the learner? Which formalism to choose for representing and managing the learner model? To answer these questions, authors in [13] proposed an ontology-based approach for the representation of learner profile and learning styles to simply use them for a personalized E-learning and to allow greater flexibility and reusability. The proposed approach collects personal information about the learners, their learning styles, prerequisites, preferences, objectives, online behaviors, etc.

In adaptive e-learning, the student information can be traced and used by systems to provide adaptive content, the authors in [14] addressed the involvement of digital library in the e-learning process and proposed a student model that adopts technologies, applications and standards from the Semantic Web by using OWL ontology language based on six main classes: Personal data, Background, Motivation state, Learning goal and Preferences. In adaptive learning, the content is adjusted to the learner's profile to respect their learning style. In [15], authors proposed to establish an ontological relationship between the formation of learner models with adaptive learning systems. As well, the authors in [16] presented an OWL learner ontology that supports personalization based on three learner style models: Felder-Silverman, Honey-Mumford and Kolb. While an ontology-based approach with six learning style model to adapted learning systems was presented in [17]. In [18], authors proposed a Framework for adaptive learning ontology to retrieve learning resources according to the learner style and knowledge level. The ontology captures information about learner's personal information, prior knowledge, and learning styles. In [19], proposed an ontology to model learners enrolled in distance learning. The proposed ontology arranges learner model characteristics into facets. The Learner class is the key concept of our hierarchy and it includes all specific details regarding learners. It's associated with the corresponding sub-classes through has Profile, has Education and has Personality.

One of the biggest challenges of adaptive e-learning systems is learner modelling. To create a model that meets the requirements, the authors in [20] proposed a novel adult learner's knowledge model using ontologies and rule reasoning. The proposed model takes into account different elements of the learner's knowledge. In [21] a ubiquitous lifelong user model ontology called LifeOnto is proposed, which meets the requirements of adaptive learning systems. Authors in [22] proposed a model based on OWL-DL ontology language that can provide support for recommended activities and personalization of educational context in Adaptive Educational Systems (AES) by grouping the chosen characteristics into four classes: Personal Information, behavior, context and progress/knowledge. An ontology-based learner modelling approach is proposed in [2] to adapt learning contents to learner. Four main classes were proposed, namely: Personal data, context, cognitive data and activity data.

In order to adapt the learning profile to the learning environment, the authors in [23] focus on the following behavioral analysis and evaluation, the detection of learning styles, the development of the learner's profile that takes into account the knowledge, preferences and attitude of the learner for learner profile modeling. Authors in [24] proposed an approach based on the semantic student profile to predict learning preference of the students based on their learning interest and style. With the advent of e-learning, even school orientation starts to be done remotely. The author in [25], proposed a framework of an ontology system called for personalized course recommendation. The approach aims to integrate the information from multiple sources based on the hierarchical ontology similarity with a view to enhancing the efficiency and the user satisfaction and to provide students with appropriate recommendations. The proposed user profile consists of two main parts. The first part is the personal attributes and education attributes of the user and the second is the user's rating of the previously recommended course.

The links between higher education and the world of work are the most controversial. Most of the controversy revolves around the mismatch between higher education opportunities and the needs of the word of work. A better understanding of the relationship between education and the world of work helps to pinpoint the reality of the problems that higher education encounters. Studying the gap between the results of higher education and the needs of industry was the author's objective in [26], by establishing an ontological relationship between the skill requirements of market occupations and the profile of learners of higher education to ensure continuous alignment between student profiles and industry. The classes used by the author to establish the ontology of the student profile are divided into five sub-models: Common model, Education model that represents the education profile, Student model, Application model and Occupation model. Table I presents a second comparison of the research studies dealing with the construction of ontologies in the academic domain for different objectives. Each research study is presented in a row of the table, including the reference, publication year, ontology language used, the tool and the main ontology model classes

used (to define and categorize the different concepts used, to a standardization of concepts is proceeded).

B. Integration of Machine Learning Techniques with Ontologies in Different Domains

ML learning integration with ontologies has proven to be successful in many decision-support systems. For this, artificial neural network methods, logical rules based techniques like decision tree, mathematical functions based ones like SVM, probabilistic methods like naive Bayesian classifier and some others, are used over ontologies concepts or data. The author in [27], used an environmental ontologies of lakes with the K-means clustering algorithm to group lakes according to the average nitrogen concentration into two groups (PoorIn and RichIn). In [28] and [29], the authors proposed an artificially intelligent predictive model for a manufacturing network by developing an ontology model based on decision tree algorithm. In [30], the main goal of the research is enhancing ontology matching by using techniques coming from different fields such as ML, Information Retrieval and Graph Matching to discover correspondences between semantically related entities of ontologies by transforming the ontology matching task into a classification task in ML (match or not match category) using Decision Tree J48 model. Authors in [5], proposed an ontology-based decision tree where the principle was using characteristics of the elements and the relations between them to find the feature super-class with the highest information gain instead of using a single vector of characteristics in the model. These classes are used as decision on tree node to obtain more information on the preferences of the user. The relation between ontologies and ML techniques can be described as a reciprocal benefit relationship. In addition to the advantage of the application of ML techniques within the ontology for decision support, the ontology also carries a benefit for Machine Learning techniques especially in the data processing due to its organized structure which is especially the case for Text classification issue. As for the authors in [4] who used a Human Disease Ontology, and tried to carry out a classification problem with and without the use of ontology. The authors found that the ontology based classification stands at a higher level than the classification without ontology by using various ML classifiers.

The study of the research studies cited in the educational field shows that researchers are interested in constructing the ontology for a specific objective, either to adapt the content of e-learning to the profile of each student, or to predict the drop-out or the performance of the student, or for recommendation or guidance. According to this objective, the system is modelled to meet the given purpose. In what follows, a multi-objective decision support system is proposed, developed from an ontology which covers the different concepts of the student with the integration of ML techniques.

TABLE I. COMPARISON OF CONSTRUCTED ONTOLOGIES FOR EDUCATIONAL PURPOSES

Ref	Ontology language	Tool	Ontology Main Classes
[13]	-	-	Personal data, Prerequiste, Preference, learning style, online behavior, social data, psychological data.
[14]	OWL	Protégé	Personal data, Learning style, knowledge, Course information
[15]	-	-	Personal data, knowledge, Online behavior, skills, Interaction, Activity.
[16]	OWL	Protégé	Personal Data, Learning style, Education.
[17]	OWL	Protégé	Learning style
[18]	OWL	Protégé	Personal data, Learning style, knowledge, Course information
[19]	OWL	Protégé	Personal data, Knowledge, Learning style, Cognitive data, Preferences, Motivation, Education, Goals.
[20]	-	-	Personal data, Social data, Knowledge, Cognitive data, Personality, Psychological data.
[21]	OWL	Protégé	Aptitude, Bloom taxonomy, cognitive capability, disability, personality, stereotype, degree, language, history, learner, learning approach, learning style, plan.
[22]	OWL-DL	-	Personal data, Behavior, Context, Progress/Knowledge.
[2]	OWL	Protégé	Personal data, context, Cognitive data, Activity data.
[23]	OWL	Protégé	Personal data, Knowledge, Behavior, Interaction, Skills, Activity, Preferences.
[24]	OWL	Protégé	Personal data, Social data, Education, learning style
[25]	OWL	Protégé	Personal data, Education, Skills
[26]	OWL	Protégé	Common model, Education model, Student, Application model, Occupation model.

IV. GSMONTO: GENERIC STUDENT MODEL ONTOLOGY

A. GSMonto Creation

Based on state of the art and different data sources, the authors build a domain ontology that covers the different aspects of the student needed in building intelligent systems, in the context of keeping up lifelong learning processes, go beyond guaranteeing interoperability between different educational systems or applications via the web, capable of satisfying users educational needs, and that can be used for different purposes (prediction of failure, dropout, orientation, etc.). OWL DL language is used for its ability to be distributed across many systems and its scaling for Web needs with reasonner support. The generic student model ontology (GSMonto) is based on expert knowledge and documentation on the educational field and mainly covers different classes that can be scalable for future perspectives. The ontology creation goes through two fundamental stages, which are the acquisition and the modelling of knowledge. During the construction of ontology, a semantic reasoner is used to deduce logical consequences from a set of facts or asserted axioms (Hermit and Pallet are two Protégé reasoners examples).

The proposed ontology defines a set of 12 upper level classes, namely:

- Student: the main class of the proposed ontology.
- Personal Identity: defines the student in a unique way (first name, family name, personal address, etc.).
- Social Identity: Derived from the belonging of a student to a social group (Nationality, marital status, etc.).
- Digital Identity: Student Identity in Social Media and Web World (digital signature, twitter account, etc.).
- Family Background: Describes the family size, parents' education, parents' job, family structure (if the student has no parent, single parent or two parents), etc.
- Personality: Encompasses information about the psychology (if the student has a normal psychology or suffers from addiction, depression or anxiety), skills and personality type of the student (intuitive, extraverted or introverted).
- Professional Experiences: Internships and jobs that has been done by the student.
- Physical Limitation: student's physical limitation as visual, verbal, hearing, amputation, paralysis, etc.
- Knowledge Profile: Records if the student has general, theoretical or practical knowledge in a specific topic.
- Learning Profile: Encompasses information about the student's learning in presential and online learning platforms, like learning style, interaction preferences (practice, example or principle oriented), learning media (audio, text or video oriented), etc.
- Academic Background: Detail about the formal education that the student has received including high school background, university background (graduate education, certified formations or PhD formation).
- Cognitive Profile: Describes the range of mental processes relating to the manipulation of the information like memory level, intelligence level, etc.

Fig. 1 describes the schematic representation of the proposed ontology. Each class can include data type properties to connect a single subject with some form of attribute data, and object properties to provide the relationships between two individuals from given class.

B. GSMonto Scalability and Instanciation

To ensure the update and the scalability of the ontology, an enrichment of the latter has to be done periodically. This enrichment can be defined following either the knowledge of experts in the educational field, or by the collection of a dataset which carries new concepts from different sources, whether from learning management systems (Moodle, Blackboard, etc.), Enterprise Resource Planning (like Apogee, XML, etc.) or traditional databases. Then, a mapping process is necessary, to convert dataset components or new discovered concepts to the corresponding ontology components. When it comes to the data source, a Metadata analysis is done, to obtain a description of the data. This metadata analysis will be used to create mapping rules, which, for each database component, generate the correspondence in ontology components with the objective of creating an ontological model from a database. If the concepts drawn from the mapping rules and those of the constructed ontology are aligned, an automatic instantiation of the constructed ontology is done in the third. Nevertheless, if the mapping rules shed light on new concepts that don't exist in the ontology already established, and which are supposed to be important, an update must be done so that the ontology is up to date, as shown in Fig. 2 that describes the process of updating and instantiating the ontology.

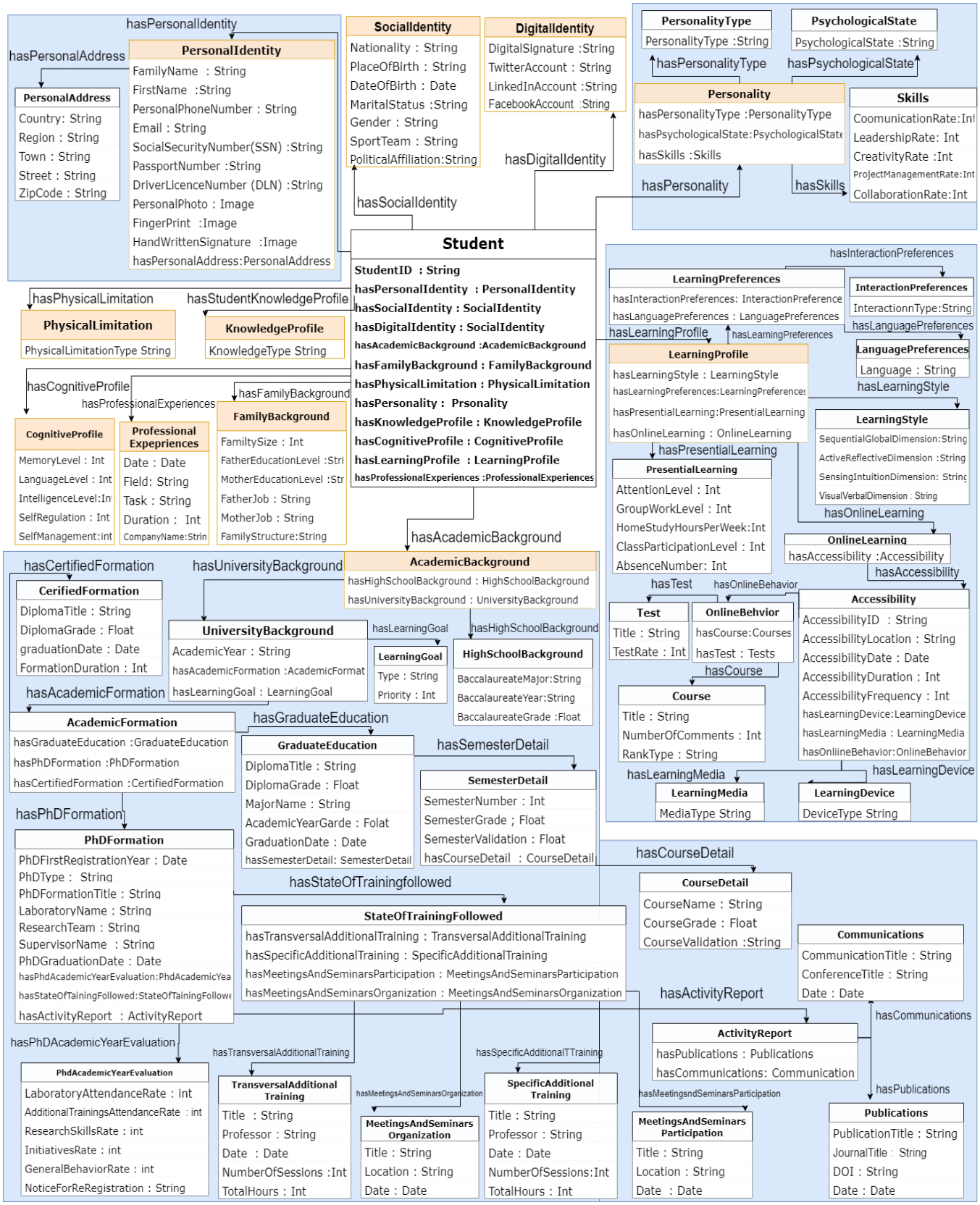


Fig. 1. GSMonto Overview.

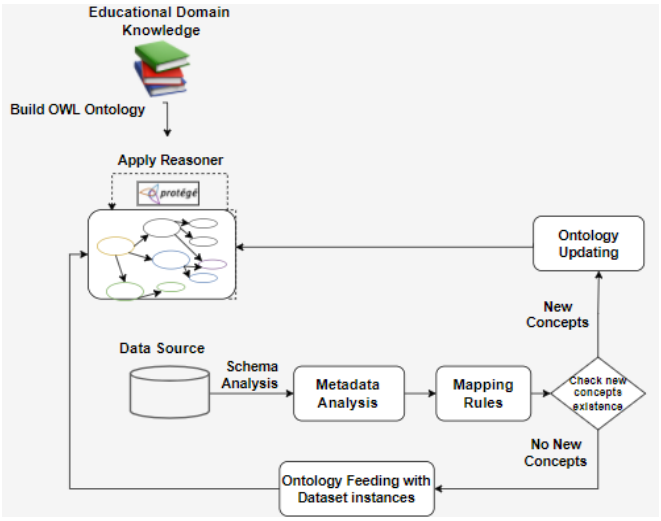


Fig. 2. Ontology Update and Instantiation Process.

V. Proposed Ontology Based Machine Learning System for Student Profile Modeling

This paper proposes an ontology based Machine Learning system using an ontology that covers the different aspects of the student. The proposed system process of the can be divided into four levels. The first level concerns the exploitation of the constructed ontology. According to the objective (prediction of failure, dropout, orientation or other objective) and to the available data, a selection of relevant concepts of the ontology is made to generate a sub-ontology. The second level is the conversion of the created sub-ontology to a dataset using mapping tools. These are based on the use of defined rules to transform the concepts of ontology (classes, individuals, data type properties, etc.) to their equivalents dataset that can be used for the application of ML techniques.

In the third level, the ML process is applied on the generated dataset, starting with data pre-processing, in the case of the presence of anomalies or incorrect values that compromise the quality of the dataset, knowing that the initial ontological presentation of the data already offers the advantage of avoiding the majority of inconsistencies, conflicts and contradictions. Then, a feature selection is carried out to reduce the number of input variables and the computational cost of modelling and, in some cases, to improve the performance of the model. In addition, data partitioning is followed to prepare for the application of the appropriate ML technique. The final level consists in converting the ML algorithm results into Semantics Web Rule Language (SWRL) to be easily integrated into the ontology, thus enriching its expressive power and increasing knowledge about individuals, so the ontology will be more consistent and can include in addition to knowledge, integrated predictive models. Fig. 3 shows the proposed ontology based ML system process for student profile modelling.

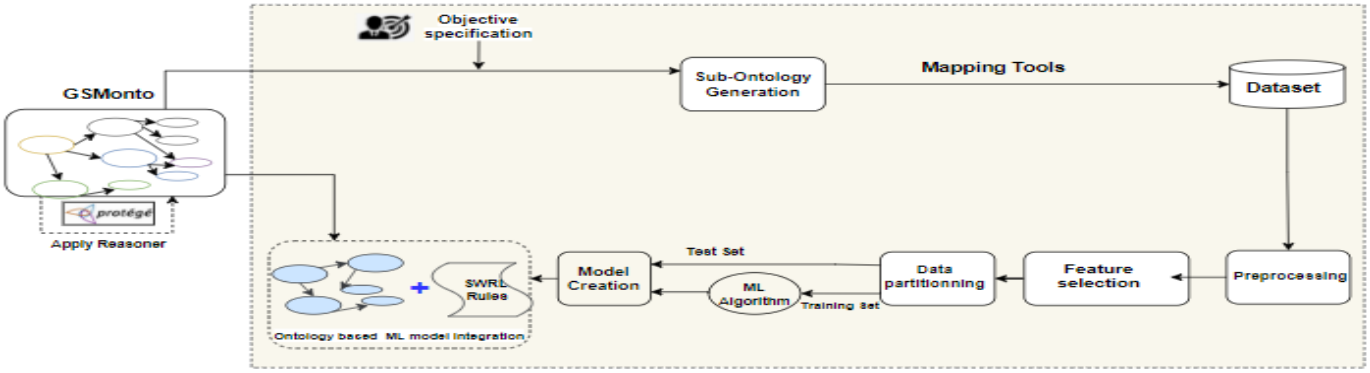


Fig. 3. The Proposed Ontology-Based Decision Support System for Multi-objective Prediction Tasks.

VI. Experiment and Results

In this section, a case study that illustrates the application of the proposed system described in the previous section is presented. The aim of this case study is to apply the proposed system in real situations, whether they are new and/or complex or to extend knowledge. The implementation process of the GSMonto ontology is accomplished with the aid of Protégé tool, which is a free and open source ontology editor for building intelligent systems.

A. Data Source and Metadata Analysis

The dataset used is extracted from the Apogee (Application For Organization and Management of Teaching and Students) database and transferred to a spreadsheet, including 20 academic performance features of 3911 student [31].

One feature relates to the validation result of the academic year (100) and two features relate to the validation of each semester of the two semesters (110 and 120), and the rest of the attributes give the validation results of the courses.

B. Ontology Updating and Feeding

In this case study, no new general context is detected, but detailed sub concepts must be presented in the established ontology, these are the courses studied in each semester. The semester class is divided into two classes (semester 1 and semester 2) and each course column title from the Spreadsheet is transformed to a subclass of each Semester class. To carry out this mapping, Cellfie is used, which is a Protégé Desktop plugin for importing spreadsheet data into OWL ontologies with the intermediary of the mapping rules. The updating process of the ontology includes information about the new courses. For example, a transformation rule expression can be defined to take the name of the spreadsheet cell G1 (column name SemesterOne) and declare an OWL named class that is a subclass of an existing YearResultDetail class which defines

that the SemesterOne class is a subclass of the class YearResultDetail , as follows:

Class: @G1 SubClassOf: YearResultDetail

Fig. 4 illustrates the proposed ontology with the update undergone in relation to new concepts in Protégé.

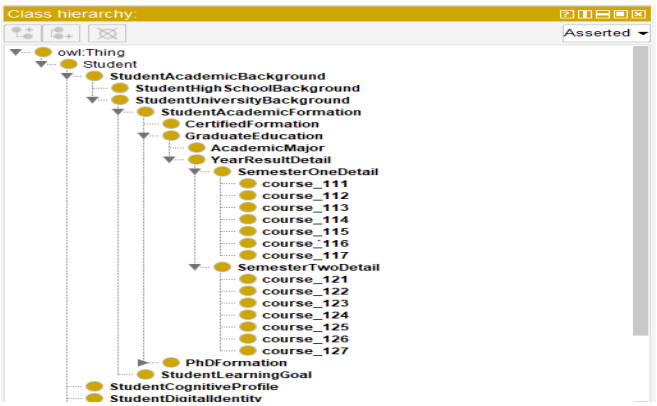


Fig. 4. GSMonto overview after the Updating Process as Diplayed in Protégé.

After updating the new concepts under the GSMonto, Pallet reasoned is used to ensure the compatibility between the concepts. Finally, the ontology feeding is carried out.

C. Machine Learning Integration

The Third step of the proposed process concerns the integration of ML within the ontology created.

An example of prediction concerns the same dataset with which the authors instantiated the proposed ontology. The validation of the academic year is the class to be predicted. The independent variables used for the prediction concerns the validation result of the first semester and its related courses. The algorithm used in this case study is C5.0. R software is used for the implementation of the model. The model performances achieved are 83.6% in accuracy and 81.9% in recall. To implement the proposed model within the ontology, the rules deduced from the decision tree are converted into SWRL rules, using a Python code that maps each rule in the decision tree to its equivalence syntax on Protégé SWRL. The generated decision tree rules are converted into SWRL rules. An example the conversion method is presented in the Fig. 5.

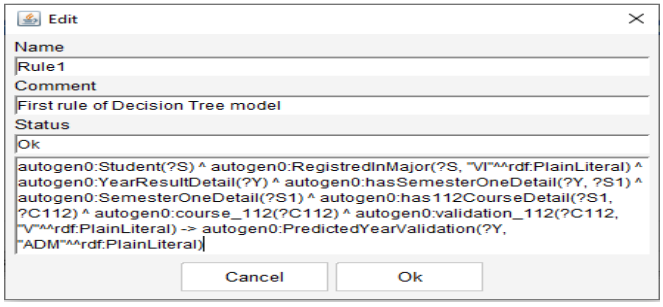


Fig. 5. SWRL Rules Implementation.

For each new student registered with the information of his academic performance in the first semester, the model gives the academic year result prediction. Fig. 6 shows the prediction result for a student: PredictedYearValidation: "ADM", which means that the student will succeed in the academic year according to the model prediction.

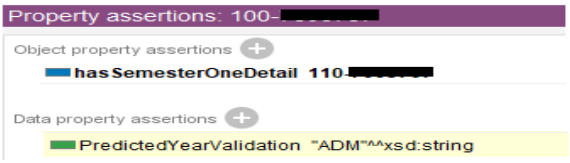


Fig. 6. Prediction Results view in Protégé.

VII. Discussion

In this study, the authors were able to implement a generic ontology that dissects the majority of students' characteristics. The authors were also able to bring together the benefits of standardization of the concepts provided by the ontology with the benefits of machine learning techniques to meet several predictive tasks. As an experiment, the decision tree algorithm was used for predicting student performance, and the results of the algorithm were transformed into SWRL rules to build an ontology-based decision support system.

VIII. Conclusion

The authors have presented a generic ontology proposed for modelling the student profile, which differ from the existing ontologies by its generic aspect that can be adapted to several objectives in the educational field. The authors also proposed a system that combines the proposed ontology with machine learning, using an algorithm based on decision trees and SWRL rules to achieve several objectives such as prediction of failure/abundance, orientation or decision-making. As a future perspective, authors plan to optimize the update process of the ontology by automatically detecting and integrating new concepts. Another goal is to implement other machine learning techniques to meet other objectives and to benefit from the AHP (Analytical Hierarchy Process) technique implemented in a previous paper to propose a module in the system dealing with the missing data as in the case of school dropout.

References

[1] T. Hamim, F. Benabbou, and N. Sael, "Student profile modeling: an overview model", in Proceedings of the 4th International Conference on Smart City Applications, New York, NY, USA, oct. 2019, p. 1-9.

[2] L. Akharraz, A. El Mezouary, and Z. Mahani, "LMOnto: An Ontology-Based Learner Model for Technology Enhanced Learning Systems", International Conference on Advanced Information Technology, Services and Systems. Springer, Cham, 2018. p. 137-142.

[3] H. Yago, J. Clemente, D. Rodriguez, and P. Cordoba, "ON-SMMILE: Ontology Nandwork-based Student Model for MultIple Learning Environments", Data Knowl. Eng., vol. 115, p. 48-67, mai 2018.

[4] S. Malik and S. Jain, "Semantic Ontology-Based Approach to Enhance Text Classification", CEUR Workshop (Vol. 2786, pp. 85-98).

[5] A. Bouza, G. Reif, A. Bernstein, and H. Gall, "SemTree: ontology-based decision tree algorithm for recommender systems", 2008.

[6] N. Guarino and P. Giarandta, "Ontologies and Knowledge Base", Towards very large knowledge bases, 1995, p. 1-2.

[7] O. Lassila, R. R. Swick, W. Wide, and W. Consortium,"Resource Description Framework (RDF) Model and Syntax Specification". 1998.

[8] V. Christophides, "Resource Description Framework (RDF) Schema (RDFS)", in Encyclopedia of Database Systems, L. LIU and M. T. ÖZSU, Éd. Boston, MA: Springer US, 2009, p. 2425-2428.

[9] D. Fensel, I. Horrocks, F. van Harmelen, D. McGuinness, and P. F. Patel-Schneider, "OIL: Ontology Infrastructure to Enable the Semantic Web", IEEE intelligent systems, 16(2), 38-45.

[10] D. L. Mcguinness, R. Fikes, J. Hendler, and L. A. Stein, "DAML+OIL: an ontology language for the Semantic Web", IEEE Intell. Syst., vol. 17, no 5, p. 72-80, sept. 2002.

[11] McGuinness, D. L., and V.Harmelen, F "OWL Web Ontology Language Overview", W3C recommendation, 2004, vol. 10, no 10, p. 2004.

[12] T. Hamim, F. Benabbou, and N. Sael, " Survey of Machine Learning Techniques for Student Profile Modeling", International Journal of Emerging Technologies in Learning, vol. 16, no 04, p. 136, févr. 2021.

[13] S. Bourekkache, O. Kazar, M. Abik, S. Tigane, and L. Kahloul. "Ontology based approach for representing the learner profile and learning styles". In : 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE, 2019. p. 1-6.

[14] D. Paneva, "Use of Ontology-Based Student Model in Semantic-Oriented Access to the Knowledge in digital libraries", In proc. of HUBUSKA Fourth Open Workshop (pp. 31-41).

[15] S. Ulfa, D.B . Lasfeto and C. Kurniawan,. "Modelling The Learner Model Based Ontology In Adaptive Learning Environment". Journal of Disruptive Learning Innovation (JODLI) (2019).

[16] B. Ciloglugil and M. M. Inceoglu, "A Learner Ontology Based on Learning Style Models for Adaptive E-Learning", in Computational Science and Its Applications – ICCSA 2018, mai 2018, p. 199-212.

[17] A. E. Labib, J. H. Canós, and M. C. Penadés, "On the way to learning style models integration: a Learner's Characteristics Ontology", Comput. Hum. Behav., vol. 73, p. 433-445, août 2017.

[18] A.MUNASSAR and A.ALI, "Semantic web technology and ontology for E-learning environment". Egyptian Computer Science Journal, 2019, vol. 43, no 2, p. 88-100.

[19] O.Zine, A.Derouich and A.Talbi, "IMS Compliant Ontological Learner Model for Adaptive E-Learning Environments". International Journal of Emerging Technologies in Learning, 2019, vol. 14, no 16.

[20] A.Abyaa, M.IDRISSI, and S.Bennani, "An adult learner's knowledge model based on ontologies and rule reasoning". In : Proceedings of the Mediterranean Symposium on Smart City Application. 2017. p. 1-6.

[21] D. Nurjanah, "LifeOn, a ubiquitous lifelong learner model ontology supporting adaptive learning", in 2018 IEEE Global Engineering Education Conference (EDUCON), avr. 2018, p. 866-871.

[22] H. N. M. Ferreira, T. Brant-Ribeiro, R. D. Araujo, F. A. Dorca, and R. G. Cattelan, "An Automatic and Dynamic Student Modeling Approach for Adaptive and Intelligent Educational Systems Using Ontologies and Bayesian Nandworks", in IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2016.

[23] A.Korchi, N. El Amrani El Idrissi and L. Oughdir. "Modeling and Implementing Ontology for Managing Learners' Profiles". International Journal Of Advanced Computer Science And Applications, 2017, vol. 8, no 8, p. 144-152.

[24] T. Sheeba and R. Krishnan, "Semantic Predictive Model of Student Dynamic Profile Using Fuzzy Concept", Procedia computer science, 2018, vol. 132, p. 1592-1601.

[25] M.Ibrahim, Y.Yang, D. L.Ndzi., G.Yang and M. Al-Maliki. "Ontology-based personalized course recommendation framework". IEEE Access, 2018, vol. 7, p. 5180-5199.

[26] H. Gasmi and A. Bouras, "Ontology-Based Education/Industry Collaboration System", IEEE Access, vol. 6, p. 1362-1371, 2018.

[27] M. Stocker, M. Rönkkö, F. Villa, and M. Kolehmainen, "The Relevance of Measurement Data in Environmental Ontology Learning", in International Symposium on Environmental Software Systems. Springer, Berlin, Heidelberg, 2011. p. 445-453.

[28] X.-B. Tang, G.-C. Liu, J. Yang, and W. Wei, "Knowledge-based Financial Statement Fraud Dandection System: Based on an Ontology and a Decision Tree", Ko Knowledge Organization, 2018, vol. 45, no 3, p. 205-219.

[29] Z. M. A. Khan, S. Saeidlou, and M. Saadat, "Ontology-based decision tree model for prediction in a manufacturing nandwork", Production & Manufacturing Research, 2019, vol. 7, no 1, p. 335-34.

[30] D. H. Ngo, "Enhancing Ontology Matching by Using Machine Learning, Graph Matching and Information Randrieval Techniques", (Doctoral dissertation, Université Montpellier II-Sciences et Techniques du Languedoc).

[31] N. Sael, T. Hamim, and F. Benabbou, "Multilevel Hybrid System Based on Machine Learning and AHP for Student Failure Prediction", International Journal of Computer Science and Network Security, Vol. 19 No. 9 pp. 103-112, 2019.

Towards a New Metamodel Approach of Scrum, XP and Ignite Methods

Merzouk Soukaina, Elkhalyly Badr, Marzak Abdelaziz, Sael Nawal
Department of Mathematics and Computer Sciences, Hassan II University- Casablanca
Faculty of Sciences Ben M'sik, Casablanca, Morocco

***Abstract*—The agile approach is a philosophy that aims to avoid the traditional management approach problems. It concentrates on the collaborative approach, using iterative and incremental development. The client receives a first production version (increment) of his software, faster thanks to agile methodologies. Project needs are influenced by the rapid expansion of technologies, particularly after the emergence of the Internet of Things (IoT). They are becoming larger and more complex. IoT provides a standardization and unification of electronic identities, digital entities, and physical objects. Consequently, interconnected devices can retrieve, store, send, and process data easier from both physical and virtual worlds. Scalable methods such as SAFe, LeSS, SPS, and others are existing methodologies ameliorated and dedicated to large projects. These methods are tough to adopt and do not consider the physical side of the project, according to IoT enterprise teams. Based on their managerial and IoT expertise, they suggest their own methods (Ignite | IoT Methodology and IoT Methodology). Model Driven Architecture (MDA) was coined by the Object Management Group (OMG) in 2000 to develop perpetual models that are independent of the technical intricacies of the execution platforms. The purpose of this paper is to propose a metamodel for each methodology among: Scrum, XP, and Ignite.**

Keywords*—*Agile software development; scrum; extreme programming; XP; internet of things; IoT; Ignite | IoT Methodology; IoT Methodology; metamodel; MDA; OMG

I. Introduction

For decades, projects have been managed with the classic (or traditional or predictive) approach, which is characterized by gathering the requirements, defining the product, developing, and testing it before the delivery. This is the "waterfall" model [1] or its adaptation, the "V" model [2].

One of the main weaknesses of 'waterfall' approach is that the design errors are often not discovered until the time of deployment. At this time, the project is almost complete, and errors are often costly to recover.

Agile methods avoid this weakness by executing iterative and incremental development that is carried out in a collaborative spirit, with the right amount of formalism. They generate a high-quality product while considering the modification needs of the customers.

Thanks to agile methods, the client participates in the realization of the project (prioritize, select items to be implemented on current iteration, do the functional tests, etc.) and obtains very quickly a first production release of his software, by using one of these methods: the XP method, the SCRUM method, the DSDM method, the FDD method, etc. [3].

Projects are becoming larger and complicated as the technology industry expands, especially after the emergence of the Internet of Things. The latter is defined as a network of interconnected electronic devices, which enables electronic identities, digital entities, and physical objects to be standardized and united. As a result, being able to recover, store, transmit, and process the associated data without interruption across the physical and virtual worlds [4].

The technology evolution leads project management experts to try different management methods or to improve existing ones. SAFe [5] [6], LeSS [7], SPS [8] and others are among the methods dedicated to large projects. IoT experts find that these methods, despite being dedicated to large and complex projects, they are complex in use and do not address the physical part of the project. At this level, IoT companies' teams propose their own methodologies based on their managerial and IoT experience. These methodologies are Ignite | IoT methodology (Ignite) [9] and IoT Methodology [10].

In 2000, the Object Management Group (OMG) coined the term Model Driven Architecture (MDA) to create perennial models that are independent of the technical minutiae of the execution platforms. This approach necessitates the use of a variety of models, including CIM, PIM, PSM, and others. As a result, the various formalisms that enable the building of models that are both sustainable and productive had to be explicitly specified. The MetaObject Facility (MOF [16]) standard, which was designed by OMG specifically for this purpose, supports the establishment of modeling formalisms in the form of metamodels. These are made up of a collection of metaclasses linked together through meta-association [11].

This article aims to present first of all the Scrum, XP and Ignite methodologies, describing their processes, artifacts, and roles. Then, it proposes a metamodel for each of these methods using MOF standards.

The rest of the paper is organized as follows: Section 2 presents the research work related to the field of project management and model engineering. Section 3 describes the methodology followed in this work. Section 4 is specific to the definition of the metamodel principle, the presentation of the proposed metamodel, and finally the description of the method

components that are the basis of the proposal. Section 5 presents the paper's discussion and conclusion.

II. Related Work

During this work, there are research works related to the metamodels of project management methods.

Many companies that use agile processes might benefit from using a process measurement framework, for example, to evaluate their process maturity. Ernesto et al. [12] offer a metamodel for the construction of specialized data models for agile development processes in their study. Then, they demonstrate how their metamodel can be utilized to derive a Scrum process model.

The traditional approaches of software development (e.g., RUP, waterfall) and agile approaches (e.g., Scrum, XP) are the two most popular software development strategies nowadays. Hybrid methods can also be used because both approaches offer advantages. Darko and Zeljko's work [13] demonstrates how to use metamodels to create new hybrid software engineering methodologies. They build a common metamodel by combining the metamodels of the traditional waterfall method with extreme programming. The new hybrid method development and development workflow are then built on top of this shared metamodel.

Given the rapid advancement of technology, the necessity for project management in terms of methodology and new concepts continues to develop. Hamzan and Belangour [14] constructed a framework for generating a metamodel that they used to project management to provide a generic metamodel of project management. This approach is founded on two project management methodologies which are "PRINCE 2" and "Scrum". The goal of this research is to validate and apply this methodology to all aspects of IT governance, then merge the metamodels to build a global metamodel that covers all IT governance domains.

The Agile Project Management Framework (APMF) is a collection of fine-grained project management techniques used in agile methodologies that is quickly gaining traction as a viable alternative to traditional project management frameworks. However, both frameworks have flaws that prevent developers from improving one in order to accept the other. Merging the two systems into a Unified Project Management Framework (UPMF) is a reasonable option. In order to achieve this goal, Mahsa and Raman [15] propose a project-management technique metamodel as a common abstract substrate for fusing the traditional framework with its agile counterpart. By abstracting the fine-grained parts of APMF, the proposed Agile Project Management Methodology Metamodel (APM3) was created. An analytical analysis of the project management procedures of seven important agile approaches was undertaken using APM3's generic agile metamodel.

A standardization of the methods concerning Scrum, XP and Ignite is proposed. This standardization is a metamodeling of the phases, the artifacts, and the whole ecosystem.

III. Methodology

The carrying out of the work is done by applying the Scrum methodology as shown in the Fig. 1. The first step consists in defining a list of tasks in the form of user stories that should be carried out throughout the work and to refine them. This list is not definitive, as it will be updated during the work. The next step is prioritizing the list according to the importance of the user stories and its sequence. The second step is to select the user stories to be done for each sprint. Finally, an increment is produced, and the next sprint is started, and so on.

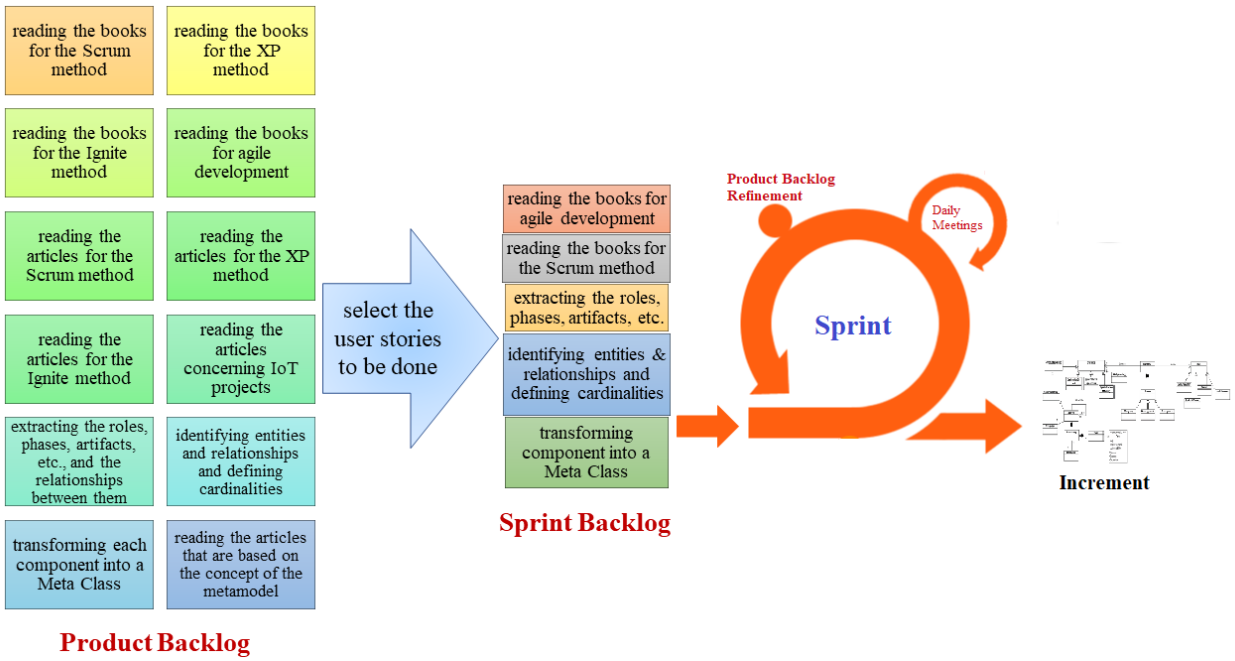


Fig. 1. The Work using Scrum Process.

The list of user stories contains: reading the books for the Scrum method; reading the books for the XP method; reading the books for the Ignite method; reading the books for agile development; reading the articles for the Scrum method; reading the articles for the XP method; reading the articles for the Ignite method; reading the articles concerning IoT projects; extracting the roles, phases, artifacts, etc., and the relationships between them; identifying entities and relationships and defining cardinalities; transforming each component into a Meta Class; reading the articles that are based on the concept of the metamodel.

- The first Sprint concerns the realization of the Scrum metamodel.
- The 2nd Sprint concerns the realization of the XP metamodel.
- The 3rd Sprint concerns the realization of the Ignite metamodel.
- The 4th Sprint concerning the proofreading, writing and correction of the proposed metamodels.

Moreover, daily meetings are held for the discussion regarding the work and the problems encountered in the realization of the current activity.

IV. Metamodel Approach

A. Definition

The Object Management Group (OMG) defined Model Driven Architecture (MDA) in 2000.

This approach advocates the massive use of models and offers the first answers to how, when, what and why to model. It includes the definition of several standards, notably UML, MOF, and XMI. The main objective of MDA is the development of perennial models, independent of the technical details of the execution platforms, in order to allow the automatic generation of the entire application code and to obtain a significant gain in productivity.

MDA necessitated the employment of a variety of models. As a result, it was necessary to explicitly specify the many formalisms that permit the construction of models that are both sustainable and productive. The MOF [16] standard, created by OMG for this purpose, provides support for establishing modeling formalisms in the form of metamodels. According to MOF, any model must respect the structure defined by its metamodel. A metamodel is thus composed of a set of metaclasses. The latter has a name and contains attributes and operations, also called meta-attributes and meta-operations. A meta-association is a binary association between two metaclasses. A meta-association has a name, and each of its ends can have a role name and a multiplicity [11].

B. Extreme Programming Methodology

1) Definition: Extreme programming, or XP, is a method proposed by Kent Beck that applies the old development principles to the extreme. It divides the project into subprojects applying the traditional development steps in each subproject in an iterative way and continuous integration (incremental) which reduces the change cost [17] [18] [19] [20]. Fig. 2 shows the evolution of the Waterfall model towards extreme programming.

2) XP Metamodel: Fig. 3 shows the metamodel of the XP method. This metamodel is based on the transformation of the method's components into metaclasses and the relationships between them into meta-associations.

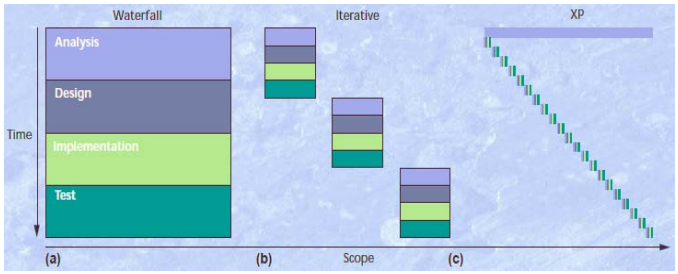


Fig. 2. Evolution towards XP Method: (a) Waterfall Model, (b) Spiral Model, (c) XP [18].

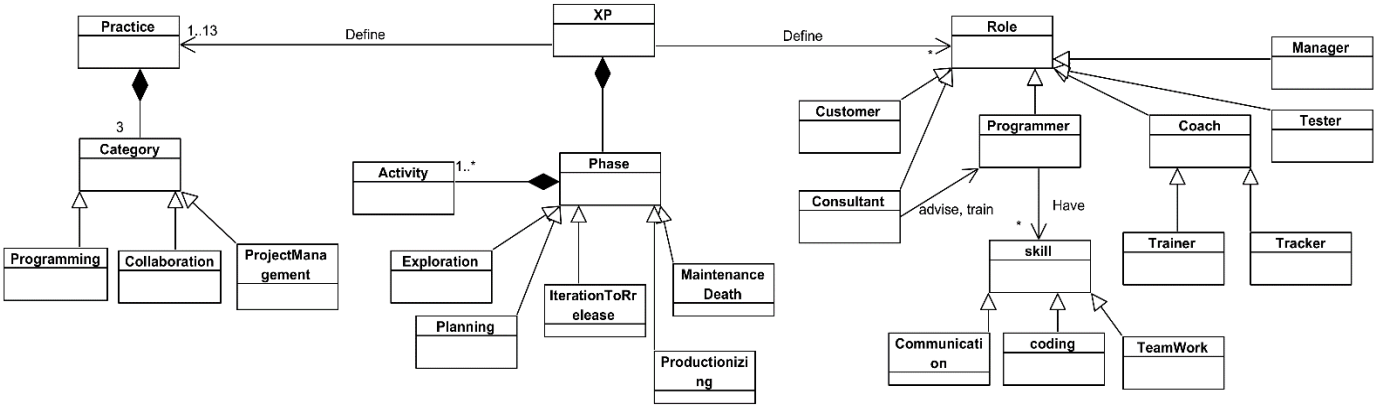


Fig. 3. Proposed Metamodel for XP Method.

3) Component:

a) Phases: The XP method consists of five phases viz., exploration, planning, iteration to release, productionizing, maintenance, and death. Each phase consists of a set of activities. These phases are described according to [20], [21], and [22].

- Exploration: The customer uses index cards to write user stories presenting his needs, as shown in the Fig. 4. These user stories are estimated by developers one by one in terms of the time needed to implement them and the implementation risk. This phase should take a few weeks to a few months and defines the technologies, tools and architectures that will be used in the project. Fig. 4 shows an example of a user story card used in the C3 project.

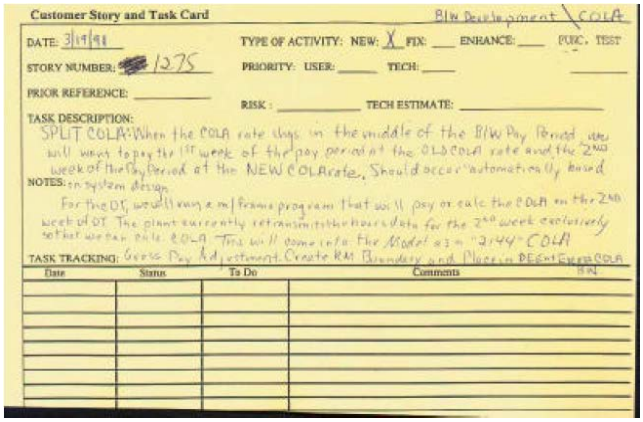


Fig. 4. C3 Project user Story Card [21].

- Planning: Commitment schedule meeting is done, after all user stories are written by the customer and estimated by the developers, to define the priority of each story and which ones are needed for the current release. In addition, the customer writes the functional tests based on the user story cards. The team transforms these cards into tasks with an estimate of each (the Fig. 5 shows a task card used in the C3 project). The meeting ends when the list of stories and the schedule are validated.

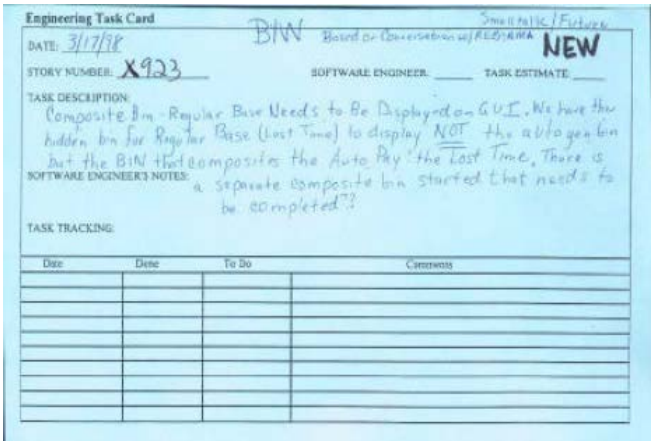


Fig. 5. C3 Project Engineering Task Card [21].

- Iteration To Release: This phase consists of breaking down the schedule of commitments set in the previous phase in a series of iterations. Each iteration follows the phases of the classical approach (designing, coding, testing, and integration). Furthermore, functional tests are applied at the end of each iteration to verify the functioning of the story.

- Productionizing: The system is ready for production at the end of the last iteration. In this case, it is necessary to ensure the performance of the system before delivering it to the customer. To do this, extra testing is done. The ideas and proposals reported are documented for later implementation during, for example, the maintenance phase.

- Maintenance and Death: The maintenance phase is triggered after the first release to the customer. The team must keep the system running in production while the new iteration is in production. It may also require the integration of new people into the team and the modification of the team structure. The death phase is the phase that describes the end of the project when the customer is satisfied and has no stories to implement. At this point, the system documentation is written. This phase also consists of closing down the system if it does not deliver the desired results or if it becomes too costly for further development.

b) Roles: Furthermore, XP defines six roles viz., Customer, Consultant, Programmer, Coach, Tester, and Manager that are described according to [18], [21], and [22]. The consultant is responsible for advising and training the programmer that having communication, coding, and teamwork skills.

- Customer: The client is responsible for defining the requirements because he writes the user stories. In addition, he defines the priority for each card and writes the functional tests which are used at the end of the iteration to check that the stories work. There is a special role in the XP method called on site customer. This is often a domain expert representing the customer.

- Consultant: In most XP projects, there are no professionals due to the rules and practices. A consultant will be employed in these circumstances to supply this information. The consultant's job is to provide expertise. One or two team members will meet with the consultant and ask several of the technical questions before attempting to fix the problem.

- Programmer: The programmer is responsible for the analysis of the design code, etc. He writes the program code as simple as possible. He is required to be competent in communication, coding, and teamwork.

- Managing part of XP project: The project management part of XP is divided into two roles such as coach and tracker Coach: The system manager participates in the management meetings. His role is to guide the team away from the process. It is necessary to be calm; to understand alternative practices that need it and could

help the current set of problems; what the ideas behind XP are; how it relates to the current situation and how other teams use XP. Tracker: Acts as the conscience of the team. He/she must track to determine if the Iteration Schedule and Commitment Schedule can be met. This work gives him data that is used to give feedback to developers on the quality of their estimates. Also, it helps him get feedback on the team's next estimates. In addition, the tracker is required to be proficient in collecting the necessary information without disrupting the whole process.

- Tester: The programmers have a large portion of the duty of testing, so the role of the tester in an XP team is particularly customer-centric. However, someone needs to run all the tests regularly, disseminate the test findings, and guarantee that the testing instruments are in good working order.
- Big Boss: Courage, confidence, and the occasional insistence that they do what they say they will be what the team most needs from the Big Boss. They aren't complaining; they genuinely aren't. They want Big Boss to know as soon as possible if things aren't going as planned, so he can react as quickly as feasible. If it works, he will be golden because he will have a team that's productive, satisfied with its clients, and does everything they can to avoid surprising him.

c) Practices: The method's practices are grouped into three categories, such as programming, collaboration and project management, which are described according to [21], [22], [23], and [24].

Programming category

- Simple design: The simplest solution that will work is always implemented by developers. They do not, for example, design generic mechanisms if the urgent necessity does not necessitate it.
- Refactoring: Developers are not hesitant to go back over the written code to make it "cleaner," to remove any unused components, and to prepare it for the addition of the next feature. More generally, this practice suggests a continuous design approach that highlights the application's structure as it develops.
- Test-first programming, unit tests, developer tests: Even as they are writing the code, developers generate automated tests for it. This enables them to gain a deeper understanding of the problem before creating the code. In addition, to gradually build up a battery of tests that enables them to make changes to the application fast and with confidence.
- Acceptance Tests, Customer Tests: Through participating in the writing of acceptance tests, the client expresses his wants and the programmers' objectives very explicitly. Acceptance tests, like unit tests, must be automated in order to ensure that the product does not regress on a daily basis.

Collaboration category

- Pair Programming: The developers always work in pairs on the same machine when coding for the application - this is an "extreme" type of code review that both developers actively collaborate to resolve issues they discover. The pairs change regularly, so everyone must work with all other team members early or later.
- Collective code ownership: All developers in the team may be required to work on all parts of the application. Furthermore, they have a duty to improve the code they work on, even if they are not the original authors.
- Coding standards: Developers follow coding rules defined by the team itself. This ensures that their code is consistent with the rest of the application, and therefore facilitates the intervention of other developers.
- Metaphor: Developers do not hesitate to use metaphors to describe the internal structure of the software or its functional issues. This facilitates communication and ensures a certain homogeneity of style throughout the design, the ideal being to describe the system in its entirety by a single metaphor.
- Continuous integration: Developers synchronize their work as frequently as feasible, at least once a day. This decreases the frequency and severity of integration issues, while also ensuring that a current version of the software is always available.

Project Management Category

- Frequent releases: The team delivers software releases at a regular rate, as high as possible, depending on the client's preferences. This enables both the team and the client to guarantee that the product meets the client's expectations and that the project remains on schedule.
- Planning game: In dedicated sessions done on a regular schedule throughout the project, the client and the development team collaborate on project planning.
- On-site customer, whole team: The customer is literally incorporated into the development team, allowing him to set priorities and specify his wants clearly, notably by directly addressing programmers' inquiries and taking advantage of the instant feedback provided by a frequent-delivery application.
- Sustainable pace: The team adopts schedules that allow it to keep the energy required to create high-quality work and efficiently follow other project procedures.

C. Scrum Methodology

1) Definition: The word "scrum" and the method's idea are derived from a rugby strategy that entails "bringing an out-of-play ball back into the game" through collaboration [25] [26]. The method was created in the early 1990s to manage the systems development process. It is a framework that focuses on how team members should work together and always ready to reorient so as to create a flexible system in an ever-changing environment. Scrum helps an organization's existing

engineering processes by requiring frequent management activities to systematically identify gaps or impediments in the development process, as well as the techniques that are employed [27] [19].

2) Scrum metamodel: The Fig. 6 shows the metamodel of the Scrum method. This metamodel is based on the transformation of the method's components into metaclasses and the relationships between them into meta-associations.

3) Component:

a) Phases: Scrum process consists mainly of three phases viz; Pre-Game, Development and Post-Game, which are described according to [20] and [22].

- Pre-Game: Pre-game consists of two sub-phases, the planning phase, which serves to define the system being developed. Defining the customer requirements and estimating the effort needed to implement each requirement. The list of requirements is always updated with new requirements. This phase ends with the definition of the project team, tools, and resources. The architecture phase consists of the high-level design of the system based on the requirements determined in the previous phase and the preparation of preliminary plans for the content of the releases.
- Development: The development phase or game phase is a black box where the system is developed in sprints that comprise the traditional software development phases. In addition, Scrum identifies environmental and technical variables. Then, aims to control them through various Scrum practices during the sprint, which is planned to take from one week to one month.
- Post-Game: The closing phase of the release after the customer has reached his goal, and he has no change or other requirements. Release preparation includes the integration, testing, and documentation of the final system.

b) Roles: Scrum defines the roles, described according to [28], which are Scrum Master, Product Owner, Team, and Stakeholders.

- Scrum Master: He is the team leader and at the same time a team member, he helps the team to understand the Scrum methodology, to create a high-value increment. He ensures that there are no obstacles that stop the progress of the product and that the team respects the work schedule. It helps the product owner to define and manage the product backlog. Then, facilitates the collaboration of stakeholders according to the demands or needs and to remove the barriers between them and the team [27].
- Product Owner: This is the most important role in this method. For the reason that it is the representative of the customer and responsible for defining the product vision and following its transition to the product backlog list. The latter is used by the product owner to check that the requirements are developed. The most significant skill that the product owner is supposed to have is written and oral communication [29].
- Team: Responsible for the project from planning to delivery of an increment. It is self-organized, works together, and takes account of the probability of change in requirements.
- Stakeholder (s): It is a person or a group of people or organizations that have a relationship with the project who are the clients [30].

c) Practices: It also defines seven practices divided into two categories, events, and artifacts to be applied in different phases to avoid chaos caused by the unpredictability and complexity. The practices, which are described according to [22], [27], and [31], are: Daily Meeting, Sprint Planning Meeting, Sprint Review Meeting, Effort Estimation, Sprint, Sprint Backlog, Product Backlog. Sprint has a Velocity based on the Sprint Backlog, which is produced from the Product Backlog. In addition, it consists of a set of User Stories that contain Elements and contain Tasks of different types.

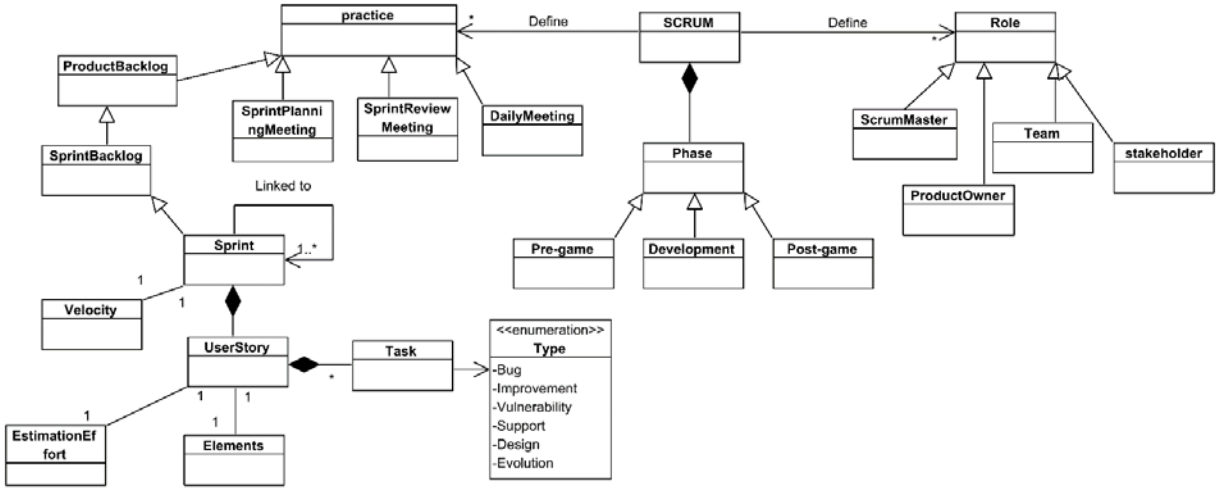


Fig. 6. Proposed Metamodel for Scrum Method.

Events category

- Daily Meetings: Scrum meetings are led by the Scrum Master. In addition to the Scrum team, management can also attend the meeting. The Scrum Team's Developers attend the Daily Scrum, which lasts 15 minutes. Developers are those who are actively working on Sprint Backlog items, such as the Product Owner or Scrum Master. Daily Scrums improve communication, identify barriers, stimulate quick decision-making, and thereby avoid the need for additional meetings.
- Sprint Planning Meeting: The Scrum Master organizes a Sprint Planning Meeting that is a two-section meeting. In the primary section of the meeting, customers, users, management, the Product Owner, and the Scrum Team determine the objectives and functionalities of the subsequent Sprint. The Scrum Master and the Scrum Team hold the second section of the meeting, which focuses on how the product increment is implemented in the course of the Sprint.
- Sprint Review Meeting: The Sprint Review's goal is to examine the Sprint's results and make recommendations for future changes. In an informal meeting, the Scrum Team and Scrum Master present the outcomes of their work to management, customers, users, and the Product Owner. The participants evaluate the product increment and make decisions about the next steps. The review meeting may result in the addition of new Backlog items and possibly a change in the system's direction.

Artifact's category

- Sprint: The procedure of adjusting to changing environmental variables is known as sprint (requirements, time, resources, knowledge, technology, etc.). In a Sprint, where ideas are becoming valuable, the Scrum Team arranges itself to generate a new executable product increment over the course of thirty calendar days. Sprint Planning Meetings, Sprint Backlog, and Daily Scrum meetings are the team's working tools. Each Sprint may be in concept of as a small project. Burndowns, burn-ups, and cumulative flows are all techniques for forecasting progress. While they have proven to be valuable, they do not take the place of empirical evidence. What's going to occur in complicated environments is unknown. Only what has already occurred can be used to make decisions in the future. If the Sprint Goal is no longer relevant, the Sprint may be cancelled. Only the Product Owner has the right to terminate the Sprint. During the Sprint, no modifications are made that might endanger the Sprint Goal; quality is maintained; the Product Backlog is adjusted as required; and, as additional information becomes available, with the Product Owner, the Scope can be clarified and renegotiated.
- Sprint Backlog: A Sprint Backlog is created at the start of each Sprint. The Sprint Backlog is made up of the Sprint Goal (why), the Product Backlog items chosen for the Sprint (what), and an actionable plan for producing the Increment (how). The Sprint Backlog is a strategy created by and for developers. The items are chosen in the Sprint Planning meeting by the Scrum Team, in collaboration with the Scrum Master and the Product Owner. Which are based on prioritized items and Sprint Goals. The Sprint Backlog, in contrast to the Product Backlog, is stable until the Sprint is finished. A new iteration of the system is deployed once all the items in the Sprint Backlog have been accomplished.
- Product Backlog: Based on existing knowledge, the Product Backlog describes everything that is required in the final product. As a result, the Product Backlog describes the project's tasks. Features, functionality, bug fixes, issues, requested improvements, and technology upgrades are all examples of backlog items. The list also includes issues that must be resolved before other Backlog items may be completed. Product Backlog items can be created by a variety of actors, including the customer, project team, marketing and sales, management, and customer support. The Product Owner oversees keeping the Product Backlog updated.
- Effort Estimation: The Product Owner and the Scrum Team oversee effort estimation, which is an iterative procedure. The Product Owner collaborates with others as the backlog is generated to predict how long it will take to develop. He or she consults with developers, technical writers, quality assurance staff, and others who are familiar with the product and technology in order to arrive at the estimate. Because the product owner and the team are experienced in estimating, the estimate will be accurate. The Product Owner develops estimates for every item, beginning with the highest priority backlog.

D. Ignite Methodology

1) Definition: Originally from industry. The founders of this methodology are based on the analysis of manufacturing and industry, connected vehicles, Smart Energy and Smart Cities. Through collecting best practices of IoT strategy management and project execution. It is open source and covers all aspects of IoT development. It addresses various IoT stakeholders, namely product managers, project managers, and solution architects [9] [32] [33] [34] [35].

2) Ignite metamodel: Fig. 7 shows the metamodel of the Ignite method. This metamodel is based on the transformation of the method's components into metaclasses and the relationships between them into meta-associations.

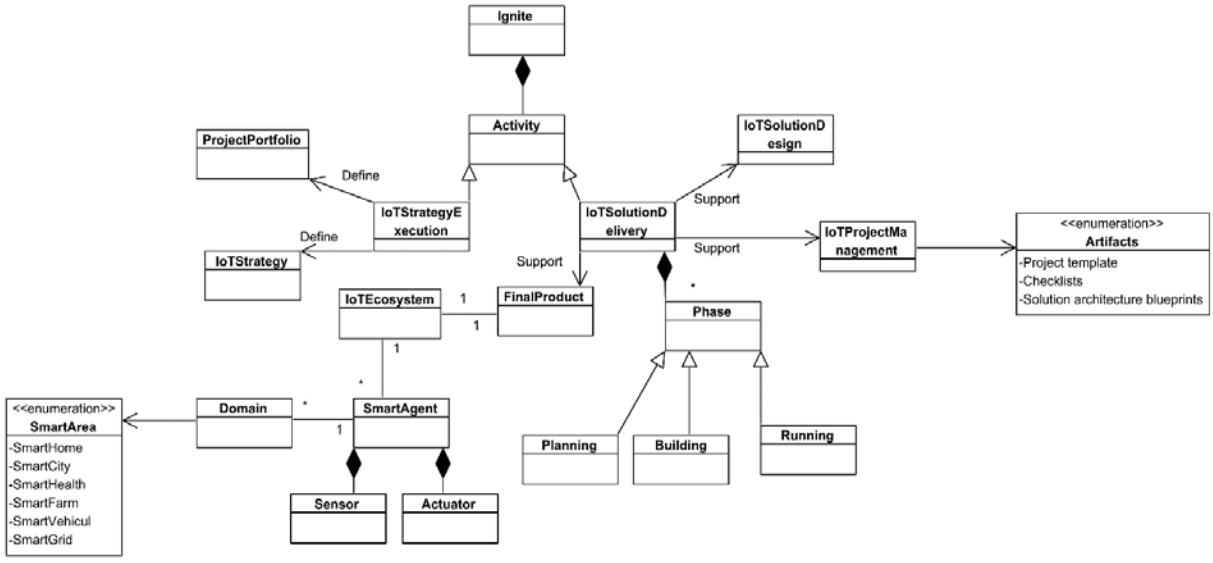


Fig. 7. Proposed Metamodel for Scrum Method.

3) Component: All components of this methodology are described according to [9].

a) Phases: Two activities make up the Ignite process (as shown in Fig. 8). The first is to define the strategy and prepare the organization to adopt IoT, then create and manage the portfolio of IoT projects to support the IoT Strategy. This activity is called IoT Strategy Execution. The second activity, called IoT Solution Delivery, is used to execute the famous three phases such as Plan, Build and Run to deliver a solution.

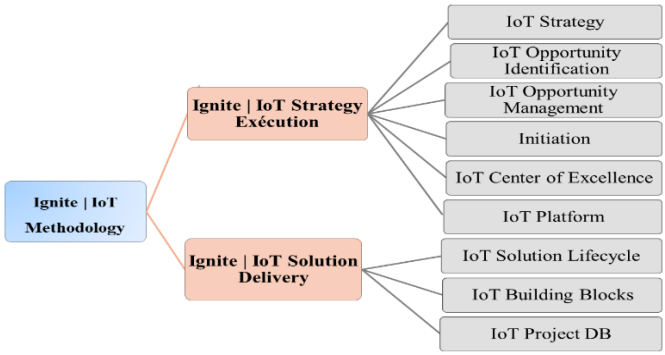


Fig. 8. Ignite | IoT Methodology Activities.

Ignite | IoT Strategy Execution

IoT Strategy, IoT Opportunity Identification, IoT Opportunity Management, Initiation, IoT Center of Excellence, and IoT Platform are the six domains of the Ignite | IoT Strategy Execution framework.

- IoT Strategy: The extent and speed with which a corporation should migrate toward IoT should be reflected in its IoT strategy. The Internet of Things strategy must have a vision, goals, and guiding principles. It should also provide a high-level overview of how IoT-related business areas should create strategic alliances and partner ecosystems. Finally, it must oversee the portfolio of IoT prospects and projects, as well as budgeting and IoT roadmap management.
- IoT Opportunity Identification: The generation of IoT solution innovation ideas, can take two forms: an open process that taps into the creativity of employees, customers, and developers, or a more structured approach in which ideas are derived from a specific context, such as the company's value chain. Ideas that show the most promise should be fleshed out further, perhaps using idea refinement templates.
- IoT Opportunity Management: The most promising ideas are then improved as part of the IoT Opportunity Management process after passing the first quality gate. In order to examine the utility and the business case, a more complete business model must be created. The following Impact & Risk Assessment step guarantees that all conceivable results of the business model are taken into account.
- Initiation: An IoT opportunity can be moved to the Initiation stage once it has been authorized. Management must decide how to best set up the effort at this stage, e.g., as a dedicated internal project, a spin-off, or even an M&A project. These activities connect to the Ignite | IoT Solution Delivery for internal initiatives.
- IoT Center of Excellence: An IoT Center of Excellence (CoE) can assist new projects in gaining traction faster. For instance, by offering IoT consulting and alter management support, IoT maturity evaluations can assist a company in determining where it stands in terms of IoT adoption.
- IoT Platform: Large enterprises may find it beneficial to provide a shared IoT Platform that many projects can use to create their solutions. An IoT application platform, connectivity solutions, and technical and functional standards are typically included.

Process

- Generate Idea: In a large company, there are usually two approaches to produce ideas: open idea generation (green field technique) or a more structured idea generation approach. The latter approach is generally carried out in a top-down manner. It typically entails a thorough market investigation by an internal strategy team or an external consulting agency. Open idea creating is more likely to produce disruptive ideas. As a result, companies should have numerous channels in place to collect these ideas, including employees, consumers, and even developers.
- Refine Idea: Many good ideas aren't particularly attractive when they're initially formed. Before they will really persuade potential stakeholders, they need care in order to grow and mature. Thankfully, there is no shortage of ideation methodologies that promote idea refining, such as the St. Gallen Business Model Navigator™ and the Innovation Project Canvas. The detailed idea sketch, which is the product of the idea-refinement phase, can be used for presentations at the next quality gate level. It can be used to develop the business model after it has been approved.
- Business Model Development: it consists of four phases (shown in the Fig. 9) viz., Strategic embedding (it lays the foundations of the business model and ensures consistency with the IoT strategy or the company's IoT vision. The implementation of "future proofing" should indicate how the business model intends to address future challenges.); Value proposition (To increase the attractiveness of the offer for customers, the proven approach of segmentation of target groups, formulation of the value offer and definition of customer channels can be used.); Customer journey (The explanation of the end-to-end solution from the customer's perspective serves in highlighting the characteristics of the proposal that the consumer finds important. Another benefit of establishing the customer journey is that it guarantees that all relevant consumer channels have been identified.); Value added (The value added can be demonstrated once the solution has been defined. The capabilities of the parties are the network's constituent elements: they are a combination of technology, resources, and know-how that they can bring in to assist the solution.); Business case (There are numerous techniques and templates for calculating business cases, but the recommendation is to use the same one for all IoT activities, as this makes comparing business models easier.); Strategic impact and subsequent business models (The house's chimneys represent two non-monetary effects of a business model that must be considered alongside the business case. The second chimney, “subsequent business models” is extremely specific to the IoT: it is very usual for teams to come up with exciting new ideas on how to utilize the data. Additionally, build new services while designing the business model and gathering all the associated data with connected devices.).

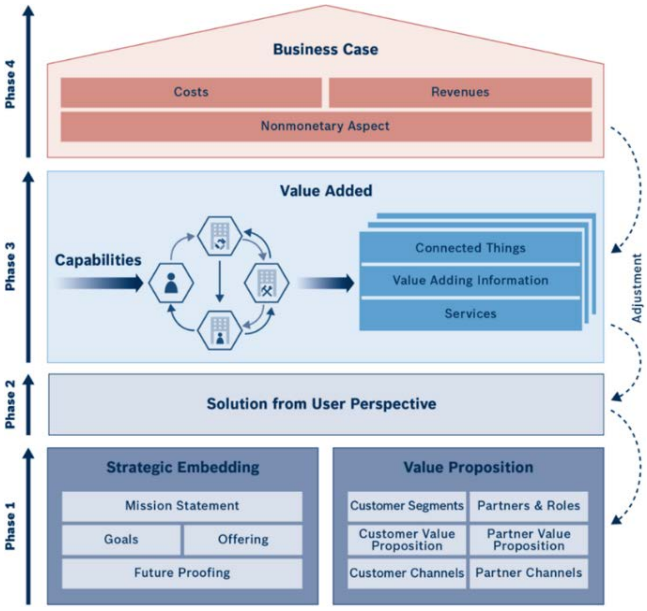


Fig. 9. Builder of IoT Business Model [9].

- Impact And Risk Analysis: Business models, and business cases in particular, deal with future value flows, so they are subject to uncertainty. To promote transparency for decision-makers and to generate the tasks necessary to resolve these uncertainties, it is critical to underline the degree of uncertainty within the business model. Various future scenarios are suggested that support the provided parameters. Trends or cause-and-effect relationships can be used to accomplish this. In the context of the strategy, it is crucial to check those aspects of the business model that create effect and value.

2 Ignite | IoT Solution Delivery

- By providing project templates, checklists, and solution architectural blueprints, the mission is to make IoT best practice applicable in the form of a technology-independent, reusable, open-source methodology that supports IoT solution design as well as the implementation and management of IoT projects. The following is a breakdown of Ignite | IoT Solution Delivery.
- The IoT Solution Lifecycle focuses on the planning, development, and execution of IoT solutions encloses the following elements. Initial Project Design: The elements established as part of the generic IoT Building Blocks, such as project self-assessment employing IoT Project Dimensions, solution architecture employing IoT Architecture Blueprints, and technology selection employing IoT Technology Profiles, are all used in this design blueprint. Project workstreams and project organization: The top-level organization and workstreams generally found in an IoT solution project are defined in this blueprint. There is a checklist for every workstream, as well as a list of common dependencies between them.

- IoT Building Blocks includes reusable artifacts such as IoT Project Dimensions, IoT Architecture Blueprints, and IoT Technology Profiles from successful projects.
- IoT Project DB is a repository of reference projects that have been analyzed in order to identify best practices for the IoT Solution Lifecycle and Building Blocks.

Process

- IoT Project Initiation: A requirements study, which is more in-depth than the analysis performed during the business model building phase, is a significant factor in the Ignite | IoT Project Initiation phase. A tiny team of subject matter specialists is generally in charge of project initiation. A business analyst with strong domain expertise and a clear vision for the solution's functional features should also be part of the team.
- Initial Solution Design: Initial Solution Design consists of a collection of key artifacts that include analysis, projections, and planning, as well as functional and technical design artifacts. Even though they might be developed concurrently, it is generally more practical to group them, as shown the Fig. 10. Analysis, Projections, Planning: It was created to aid with analysis, projections, and planning. They contain: Problem Statement, Stakeholder Analysis; Site Survey; Solution Sketch; Project Dimensions; Quantity Structure; Milestone Plan.

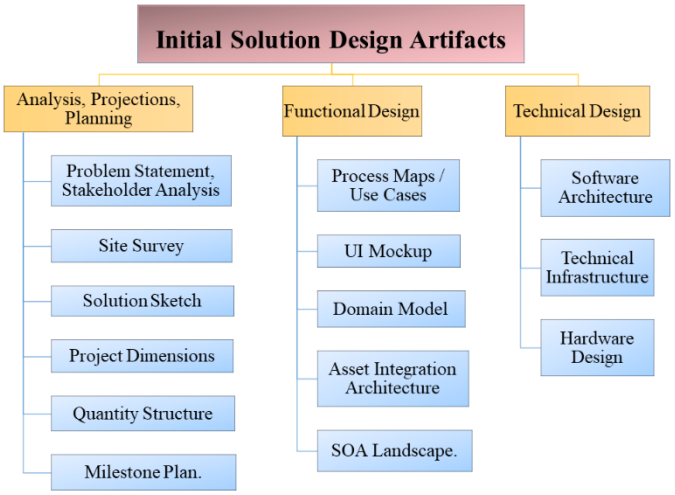


Fig. 10. Initial Solution Design Artifacts.

Functional Design contains: Process Maps / Use Cases; UI Mockup; Domain Model; Asset Integration Architecture; SOA Landscape. Technical Design contains: Software Architecture; Technical Infrastructure; Hardware Design.

- Plan: After the funding decision on Ignite | IoT Strategy Execution, this phase begins. A small, but devoted team usually creates an initial project plan, which includes a solution definition, based on the ideas and criteria from the business planning phase. This could be an RFP (Request for Proposal) document or the initial list of high-level epics that will need to be broken down into more detailed user stories later in the Build phase. The initial team will often oversee sourcing (internally or externally) the larger team that will eventually create the solution during the planning phase.
- Build: A larger team or teams often execute the build phase. Keep in mind that, particularly in the IoT, the work is frequently with various, multidisciplinary teams. It's worth noting that, due to the often highly dynamic nature of IoT projects, planning continues during the build phase. Each sprint will be meticulously planned, especially if an Agile approach is used. The higher-level papers developed during the planning phase will frequently need to be updated to reflect new or changing needs, as well as lessons gained from prior sprints.
- Run: The project team is typically disbanded, and the solution is handed over to a line organization around the time of the IoT solution's Start of Production (SOP). This line organization will set up an integrated DevOps organization in modern enterprises, which will deal with both the solution's continuous development and operations. DevOps for IoT can be more challenging than typical DevOps due to the potentially extremely distributed nature of IoT systems.

V. Discussion and Conclusion

The previous section presents the Scrum and XP agile methods and the dedicated IoT project method Ignite. Moreover, it presents the metamodel of each one with their components that have been translated into metaclasses and meta-associations.

The Scrum and XP methods are based on almost the same principles, with a very clear definition of roles, unlike the Ignite method.

Furthermore, another difference between Ignite and the Scrum and XP methods is that Ignite divides the project realization process into two sub-processes called Strategy Execution and Solution Delivery activity, whereas Scrum and XP have an ecosystem whose components are chained.

To sum up, the paper presents the standardization of the Scrum, XP and Ignite methods as metamodels based on their components and the fundamental MDA principles. These metamodels are the beginning of the forthcoming contribution concerning a Framework used for Industry 4.0.

References

[1] L. Sherrell, “Waterfall Model,” in Encyclopedia of Sciences and Religions, A. L. C. Runehov and L. Oviedo, Eds. Dordrecht: Springer Netherlands, 2013, pp. 2343–2344. doi: 10.1007/978-1-4020-8265-8_200285.

[2] I. Graessler, J. Hentze, and T. Bruckmann, “V-MODELS FOR INTERDISCIPLINARY SYSTEMS ENGINEERING,” 2018, pp. 747–756. doi: 10.21278/idc.2018.0333.

[3] S. Merzouk, S. Elhadi, A. Cherkaoui, A. Marzak, and N. Sael, “Agile Software Development: Comparative Study,” SSRN Electron. J., 2018, doi: 10.2139/ssrn.3186323.

[4] A. Abddessamad, S. Cherkaoui, M. Sm, A. Abdelaziz, M. Marzak, and H. Mh, “Review on Embedded Systems and the Internet of Things: Comparative Study,” Assoc. Comput. Mach., p. 7, 2021, doi: 10.1145/3454127.3457636.

[5] R. Knaster, SAFe 4.0 distilled: applying the Scaled Agile Framework for Lean software and systems engineering. Boston, MA: Addison-Wesley, 2017.

[6] "What is SAFe | Scaled Agile," SAFe® Enterprise Solutions. https://www.scaledagile.com/enterprise-solutions/what-is-safe/

[7] C. Larman, B. Vodde, and B. Jensen, Large-scale scrum. Dpunkt, 2017.

[8] G. Verheyen, "Scaled Professional Scrum – NexusTM," p. 5.

[9] D. Slama, F. Puhlmann, J. Morrish, and R. M. Bhatnagar, Eds., Enterprise IoT: Enterprise IoT: strategies and best practices for connected products and services. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2016.

[10] "IoT Methodology – The Internet of Things project lifecycle guide for creative, technical and business people." http://www.iotmethodology.com/ (accessed May 30, 2020).

[11] X. Blanc and O. Salvatori, « MDA in action model-driven software engineering », MDA en action ingénierie logicielle guidée par les modèles. Paris: Eyrolles, 2005. [Online]. Available: https://www.eyrolles.com/Informatique/Livre/mda-en-action-9782212115390/

[12] E. Damiani, A. Colombo, F. Frati, and C. Bellettini, "A Metamodel for Modeling and Measuring Scrum Development Process," in Agile Processes in Software Engineering and Extreme Programming, vol. 4536, G. Concas, E. Damiani, M. Scotto, and G. Succi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 74–83. doi: 10.1007/978-3-540-73101-6_11.

[13] D. Androcec and Z. Dobrovic, "Creating Hybrid Software Engineering Methods by Means of Metamodels," p. 6.

[14] H. Ibrahim and B. Abdessamad, "Project Management Metamodel Construction Regarding IT Departments," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, 2019, doi: 10.14569/IJACSA.2019.0101029.

[15] M. H. Sadi and R. Ramsin, "APM3: A Methodology Metamodel for Agile Project Management.," 2009, pp. 367–378.

[16] "MetaObject Facility | Object Management Group." https://www.omg.org/mof/ (accessed Nov. 27, 2021).

[17] A. Anderson, R. Beattie, and K. Beck, "Chrysler goes to extremes," Distrib. Comput., 1998.

[18] K. Beck, "Embracing change with extreme programming," Computer, vol. 32, no. 10, pp. 70–77, Oct. 1999, doi: 10.1109/2.796139.

[19] S. Merzouk, "A Comparative Study of Agile Methods: Towards a New Model-based Method," vol. 9, no. 4, p. 8, 2017.

[20] S. Merzouk, A. Cherkaoui, A. Marzak, N. Sael, and F.-Z. Guerss, "The proposition of Process flow model for Scrum and eXtreme Programming," Assoc. Comput. Mach., p. 6, 2021, doi: 10.1145/3454127.3457627.

[21] T. Dudziak, "Extreme programming an overview," Methoden Werkzeuge Softwareproduktion WS, vol. 1999, pp. 1–28, 2000.

[22] P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta, "Agile software development methods: Review and analysis," 2002.

[23] K. Beck, Extreme programming explained: embrace change. addison-wesley professional, 2000.

[24] Jean-Louis Bénard, Laurent Bossavit, Régis Médina, and Dominic Williams, "EXtreme Programming: project management", EXtreme Programming: gestion de projet. Paris: Eyrolles, 2004. [Online]. Available: https://www.eyrolles.com/Informatique/Livre/gestion-de-projet-extreme-programming-2eme-tirage-2005-9782212115611/

[25] K. Schwaber, "SCRUM Development Process," in Business Object Design and Implementation, J. Sutherland, C. Casanave, J. Miller, P. Patel, and G. Hollowell, Eds. London: Springer London, 1997, pp. 117–134. doi: 10.1007/978-1-4471-0947-1_11.

[26] J. Sutherland and K. Schwaber, "Nut, Bolts, and Origins of an Agile Framework," p. 224.

[27] K. Schwaber and J. Sutherland, "The Scrum Guide The Definitive Guide to Scrum: The Rules of the Game," Nov. 2020, [Online]. Available: https://scrumguides.org/scrum-guide.html

[28] S. A. Shinde, "Analysis of Agile Project Management with Scrum Method and Extreme Programming," IJSETR, vol. 4, p. 7, May 2015.

[29] S. Oomen, B. De Waal, A. Albertin, and P. Ravesteyn, "How can Scrum be succesful? Competences of the scrum product owner," 2017.

[30] A. Pham and P. V. Pham, Scrum in action: agile software project management and development. Boston: Course Technology PTR, 2012.

[31] K. Schwaber and M. Beedle, Agile Software Development with Scrum, Illustrée., vol. 1. Prentice Hall, 2002. [Online]. Available: https://books.google.co.ma/books?id=BpFYAAAAYAAJ

[32] I. Jacobson, I. Spence, and P.-W. Ng, "Is There a Single Method for the Internet of Things?," Queue, vol. 15, no. 3, pp. 25–51, 2017, doi: https://doi.org/10.1145/3106637.

[33] G. Görkem, T. Bedir, and T. Eray, "IoT System Development Methods," in Internet of things challenges, advances, and applications, Chapman & Hall/CRC Press, 2018, pp. 141–159.

[34] S. Merzouk, A. Cherkaoui, A. Marzak, and S. Nawal, "IoT methodologies: comparative study," Procedia Comput. Sci., vol. 175, pp. 585–590, 2020, doi: 10.1016/j.procs.2020.07.084.

[35] S. Rahman, "Comparative analysis about the challenges and implications of IoT development methodologies," p. 55, 2018.

Towards a Computational Model to Thematic Typology of Literary Texts: A Concept Mining Approach

Abdulfattah Omar
Department of English, College of Science & Humanities
Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
Department of English, Faculty of Arts, Port Said University, Egypt

***Abstract*—In recent years, computational linguistic methods have been widely used in different literary studies where they have been proved useful in breaking into the mainstream of literary critical scholarship as well as in addressing different inherent challenges that were long associated with literary studies. Such computational approaches have revolutionized literary studies through their potentials in dealing with large datasets. They have bridged the gap between literary studies and computational and digital applications through the integration of these applications including most notably data mining in reconsidering the way literary texts are analyzed and processed. As thus, this study seeks to use the potentials of computational linguistic methods in proposing a computational model that can be usefully used in the thematic typologies of literary texts. The study adopts concept mining methods using semantic annotators for generating a thematic typology of the literary texts and exploring their thematic interrelationships through the arrangement of texts by topic. The study takes the prose fiction texts of Thomas Hardy as an example. Findings indicated that concept mining was usefully used in extracting the distinctive concepts and revealing the thematic patterns within the selected texts. These thematic patterns would be best described in these categories: class conflict, Wessex, religion, female suffering, and social realities. It can be finally concluded that computational approaches as well as scientific and empirical methodologies are useful adjuncts to literary criticism. Nevertheless, conventional literary criticism and human reasoning are also crucial and irreplaceable by computer-assisted systems.**

Keywords*—Computational linguistics; concept mining; data mining; empirical methodologies; semantic annotators; text clustering; typology

I. Introduction

Despite the wide applications of computational and statistical approaches in different disciplines in humanities, many critics still argue against the usefulness of these approaches in literary studies. So far, most literary critics reject the use of computer technology and statistical and computational methodologies in the analysis and interpretation of literary texts [1]. In light of this argument, this study seeks to evaluate the reliability of computational and statistical approaches to literary studies, and more specifically to the thematic typologies of literary texts. In other words, it seeks to see whether computational and statistical approaches, which have long been rejected by many literary critics, can be usefully used in generating typologies that best reveal the thematic features within these texts.

The study takes Thomas Hardy's prose fiction production as an example. The great reputation Hardy received as a novelist and the thematic richness of his texts have always made his novels and short narratives a target for critics and commentators [2]. From the time Hardy published Wessex Edition, where he provided a broad classification of his works, a critical response has focused on the questions of thematic classification of his work. Critics have adopted different approaches for investigating Hardy's thematic treatment of his texts [3-5]. Nevertheless, questions are often raised regarding the reliability of these classifications. In this regard, Hardy's prose fiction represents a good opportunity to test the reliability of computational and statistical methods in the thematic typology applications to literary texts.

To carry out the objectives of the study, concept mining is used. The rationale is that concept mining methods have been usefully used in text clustering and text classification applications. They have been effectively used for generating reliable clustering and classifications that explored the underlying meanings and themes within texts. Unlike conventional text clustering approaches including vector space clustering (VSC), concept mining focuses on identifying the patterns that are associated with concepts in similar texts [6, 7]. The underlying premise is that concept mining can be usefully used for identifying and recognizing the thematic patterns which can thus be used in developing a thematic typology that expresses the thematic interrelationships within literary texts.

The remainder of this article is organized as follows. Section Two surveys the literature on the thematic typology of literary texts. Section Three proposes the research questions. Section Four describes the methodological framework of the study. Section Five describes the data collection and processing procedures. Section Six reports the results of the study. Section Seven is a discussion and interpretation of the results. Section Eight is conclusion.

II. Literature Review

Different critical approaches have been used in the thematic typology studies of literary texts. These have usually been traditionally based on qualitative methods that tended to focus on identifying and revealing the underlying meanings that are

conveyed within the texts [8]. In this regard, thematic classifications have been carried out according to different concepts including purpose, content, and period [9]. Despite the popularity of the concepts of content and purpose, typologies based on these two concepts are always associated with subjectivity being largely based on the intuitive readings and making general observations based on such readings. One major problem with the adoption of these concepts in the identification of themes is that they are based on personal thoughts and emotional reactions, and thus they are subjective.

The period concept, on the other hand, is one of the most dominating methods in the thematic classification of literary texts. The underlying premise of this concept is that every period has its characteristic thematic features that can be usefully used in grouping texts through time. Taking English literature as an example, it is usually classified under the headings the Anglo-Saxon Literature, the Medieval Literature, the Elizabethan Literature, the Neoclassical literature, the Romantic literature, the Victorian literature, the Modernist literature, and the Postmodern literature.

Following this tradition, different thematic typologies have been developed in the critical study of Thomas Hardy's prose fiction. According to Plietzsch [10], the first person to suggest a general classification of Hardy's novels is thought to be Edmund Gosse in 1890. Gosse's classification included only ten novels, which were all that Hardy had written so far [11]. In this regard, Gosse's study misses many of the themes Hardy was to develop in subsequent work. With its limitation, Gosse's classification was nevertheless a step forward in generating a broad classification of Hardy's works.

Following Gosse's attempt, Thomas Hardy provided a broad classification of his prose and verse works. In the Wessex Edition, Hardy [12] classified his novels and short stories into three categories: (1) Novels of Character and Environment, (2) Romances and Fantasies, and (3) Novels of Ingenuity. One obvious observation about Hardy's classification is that there is no clear-cut relationship between the thematic accounts Hardy himself suggested for this classification and the way the texts were finally classified. Plietzsch [13] argues that it seems that Hardy ranked his texts in this particular order as a result of the responses from the public and literary critics which he had received.

The implication for the present study is that the reliability of such a classification is thus questionable. First, Hardy did not provide definite criteria for his classification. Second, some of the texts that used to be regarded as minor by the public and critics in Hardy time are now regarded as major works [14, 15]. Furthermore, the classification does not include all of Hardy' works. A Changed Man and other Stories, for instance, was published one year after the Wessex Edition. In view of this, it excludes some important works that represent the thematic development of Hardy's career as a novelist.

In response to Hardy's work, Abercrombie [16] classified the novels and short stories based on their artistic significance into four categories: (1) Minor Novels, (2) Annexes, (3) Dramatic Form, and (4) Epic Form. In spite of its success in drawing connections between texts, Abercrombie's classification, however, raises many questions concerning replicability and objectivity, since his criteria are subjective and undefined.

Harvey [14] suggested an alternative typology where he divided Hardy's prose works into three main categories: major novels, lesser novels, and short stories. The main criterion of this classification is subject matter. Harvey considered only the social and realistic novels to be major novels. Other novels were classified under the category of minor novels. Once again, the classification lacked any objective criteria [15, 17].

Another approach to the typology of Hardy's prose fiction texts can be traced in grouping Hardy's novels and short stories based on literary criticism perspectives. The underlying principle is that critics have come to classify Hardy's works under different headings including tragedy, women, religion and philosophy, Wessex and regionalism, nature and landscape, social change, and pastoral. One major problem with these critical discussions is that their perception of Hardy's work is very narrow in the sense that they are almost restricted to what Hardy calls 'Novels of Character and Environment'. Furthermore, they ignore important thematic concepts within the texts as they usually focus on just one aspect of his writings. Equally important, such reviews are always based on some biographical elements or historical accounts which again raise questions regarding the objectivity and reliability of such typologies and classifications.

In the face of the problems associated with the conventional classifications of Hardy's prose works, recent studies built on the advances in the application of computational data processing for analyzing and classifying novels and short stories in a way that is both objective and replicable [18, 19]. Motivation has been to understand Thomas Hardy better as a literary artist in an objective, replicable, therefore scientific way. In a recent study, Omar [20] employed centroid-based lexical clustering methods for identifying the thematic structures in Hardy's prose fiction. The novels and short stories were clustered into four classes, where each class or group of texts share the same thematic features based on the lexical profiles of each class.

Despite the reasonable success of such classification in drawing a thematic mapping of Hardy's novels and short stories based on objective grounds, questions regarding the use of vector space clustering (VSC) in exploring the salient thematic features of the texts are often raised. VSC is based on what is known as bag of words techniques where context is not considered at all. In this regard, VSC is not capable of accounting for all the linguistic and contextual features of texts. In the face of these limitations, this study proposes the use of concept mining methods for generating a thematic typology that best captures the underlying meanings, concepts, and thematic patterns of literary texts taking the novels and short stories of Thomas Hardy as an example.

III. Research Questions

Despite the extensive literature on the thematic typology of literary texts including Thomas Hardy's prose fiction writings, almost all of the relevant work is theoretically driven. That is, classification criteria are selected by the critic based on some critical theory or framework supported by personal knowledge

and evaluation of the texts. Moreover, many existing accounts follow the stereotypical classifications of what might be called Hardy Critical Industry. In other words, many of Hardy's commentators are willing to agree with conventional, well-known evaluations of Hardy even though such evaluations conflict with their critical presuppositions. Two examples of this are given. First, many commentators have favored the idea of classifying Hardy's works into major and minor novels in relation to subject-matter, and many studies use such dichotomy without giving reasons for its adoption. Second, many thematic reviews of Hardy use the term 'Wessex novels' in reference to nine or ten of Hardy's novels without explaining why these nine or ten texts should constitute variants of the same theme apart from the fact that they are about Wessex [21].

It can also be claimed that many of the classifications of Hardy's work followed Hardy's own classification of his works. The problem is that Hardy did not set clearly defined criteria for his classification. Furthermore, some classifications based on philological methods are greatly biased. In the face of this problem, this study seeks to answer the following research questions:

- Can computational linguistic methods be usefully used in addressing the limitations of the conventional approaches of literary criticism regarding thematic typologies of literary texts?
- How can computational models in general and concept mining methods in particular be used in developing a thematic typology of literary texts with reference to the prose fiction writings of Thomas Hardy?
- What is the future of computational approaches and scientific and empirical methodologies in literary studies?

IV. Method

In different natural language processing (NLP) applications including text clustering and text classification, concept mining is a process that has been used to provide an automated categorization of documents based on their content [22, 23]. It is a workflow that is used to discover implicit and explicit relationships, useful associations and groupings in a set of documents or data collection with the purpose of detecting similar documents in a large corpora and classifying them by topic [24, 25]. It can provide thus powerful insights into the meaning, provenance, and similarity of documents [26-28]. The assumption is that each word in a given document relates to several possible concepts which make it possible to cluster documents based on their content. The underlying principle of concept mining is the conversion of words into concepts. This is done in two subsequent steps. First, documents are reduced into a sequence of words that describes the content. Second, these words are mapped into concepts [29].

In this way, given that we have a number of documents on generative grammar; concept mining is possible by identifying relationships and generating facts based on the data within collection and the dimensions of the subject. These can be something like Chomsky and generative grammar, theoretical linguistics and generative grammar, Phrase Structure Rules (PSR) and Generative grammar, deep and surface structures in generative grammar, etc. Documents can also be classified by topic as WH-movement, linguistic competence, etc.

In this way, concept mining is based on clustering or grouping semantically-similar texts together. Text clustering is the process of automatically grouping natural language texts according to an analysis of their information/semantic content. In other words, clustering is a task of dividing given data into defined set of clusters and it is the task of classification to structure these clusters and sort them into categories according to a group structure known in advance [30, 31]. In concept mining processes, text clustering starts by discovering and finding groups that have similar content, and then organizing our perceptions of these groups into categories. In other words, clustering places documents into natural classes and generating taxonomies that best describe the patterns within the datasets [32].

V. Data Collection Procedures

For generalizability purposes, the study is based on all the novels and collections of short stories written by Thomas Hardy. Three sources were used for data collection. These are shown as follows.

- Chadwyck-Healey Literature Collections is a commercial product with authoritative full-text databases that offers coverage of English literary works from 1477 to the present.
- The Gutenberg Project is the oldest producer of free e-books on the Internet with a large volume of collections produced by thousands of volunteers. The project was founded in 1971 by Michael Hart.
- The Thomas Hardy Short Story Page (http://darlynthomas.com/hardyshortstories.htm-) includes all collection of short stories and the individual, excluded and collaborative stories written by Thomas Hardy.

Hardy has 14 published novels. These are listed below.

1) Desperate Remedies
2) Under the Greenwood Tree
3) A Pair of Blue Eyes
4) Far from the Madding Crowd
5) The Hand of Ethelberta
6) The Return of the Native
7) The Trumpet-Major
8) A Laodicean
9) Two on a Tower
10) The Mayor of Casterbridge
11) The Woodlanders
12) Tess of the D'Urbervilles
13) Jude the Obscure
14) The Well-Beloved

In his life time, Hardy also published four collections of short stories. These are A Group of Noble Dames [33], Life's Little Ironies [34], Wessex Tales [35] and A Changed Man and

other stories [36]. These account for "thirty-seven stories in all [37]. For the first three collections, texts were abstracted from the Wessex Edition [12]. A Changed Man and other stories was published one year after the publication of the Wessex Edition. So it was not included in that edition. The data was abstracted from Macmillan & Co 1913 edition which includes as well the novella The Romantic Adventures of the Milkmaid.

A. Wessex Tales

The short stories used in this study are the contents of the Wessex Tales collection of the 1912 *Wessex Edition*. These are shown as follows.

The Three Strangers
A Tradition of Eighteen Hundred and Four
The Distracted Preacher
The Withered Arm
Fellow-Townsmen
Interlopers at the Knap

B. Life's Little Ironies

The stories in this collection had been written at different periods but were assembled in 1893 for publication under the title given. In 1912, Hardy reassembled the stories for the Wessex Edition as indicated below. The texts used for the data of the study are the ones in the Wessex Edition. The stories in this collection are listed below.

An Imaginative Woman
For Conscience's Sake
The Fiddler of the Reels
To Please His Wife
On the Western Circuit
A Few Crusted Characters
A Tragedy of Two Ambitions
The Son's Veto

C. A Group of Noble Dames

In his preface to *A Group of Noble Dames*, Hardy [38] indicates that the tales were first published in periodicals six or seven years before being collected and published in 1891. Hardy published the tales once again in the same form in which they appeared in the 1891 edition for Wessex Edition in 1912. The stories in this collection are shown below.

The First Countess of Wessex
Barbara of the House of Grebe
The Marchioness of Stonehenge
The Lady Icenway
The Duchess of Hamptonshire
Anna, Lady Baxby
Lady Mottisfont
Squire Petrick's Lady
The Honourable Laura
The Lady Penelope

D. A Changed Man and other Stories

Although the tales in this collection represent an important stage in Hardy's development, they were not collected in one volume until 1913. Interestingly, Hardy used the expression minor novels instead of short stories. This may suggest that the short story was not a fully-fledged literary genre yet. Stories in this collection are shown below.

A Changed Man
Alicia's Diary
A Tryst at an Ancient Earthwork
A Committee-Man of 'The Terror
The Waiting Supper
The Grave by the Handpost
What the Shepherd Saw
Master John Horseleigh, Knight
A Mere Interlude
The Duke's Reappearance
Enter a Dragoon
The Romantic Adventures of a Milkmaid

E. Excluded and Collaborative Stories

These are the stories which were not included in the collected volumes published during Hardy's life. They were collected and edited by Pamela Dalziel [39] in Thomas Hardy: The Excluded and Collaborative Stories. Dalziel argues that although the stories occupy a significant position in the professional career of Hardy as a novelist, they have not received due critical treatment from critics and biographers. She also stresses that these stories are not thematically coherent: "Each story is an individual work, meriting treatment as such, and its unique conditions of composition and publication (or non-publication) have been carefully considered, in so far as they are now recoverable, when making editorial decisions" [39]. In making these stories available in one volume, Dalziel addresses some limitations in Hardy scholarship.

F. A List of Hardy's Excluded and Collaborative Stories

This collection consists of 10 stories including both excluded and collaborative stories of Hardy. The Spectre of the Real is the only story acknowledged by Hardy to be a collaborative tale [39]. It was written in collaboration with Florence Henniker. Blue Jimmy: the Horse Stealer and The Unconquerable were written in collaboration with his wife Florence Dugdale-Hardy. However, Hardy never admitted Florence's role in the two stories [39]. The texts in this collection are listed below.

How I Built Myself a House
The Thieves Who Couldn't Help Sneezing
The Doctor's Legend
Our Exploits at West Poley
Destiny and a Blue Cloak
Old Mrs. Chundle
The Spectre of the Real
The Unconquerable
An Indiscretion in the Life of an Heiress
Blue Jimmy: The Horse Stealer

G. Unpublished Work

The Poor Man and the Lady is Hardy's first novel. He sent the manuscript of the novel to Macmillan who refused to publish it on the grounds that the book creates a world which is entirely dark. Accordingly, Hardy took it to Chapman & Hall where George Meredith recommended him not to publish it. Meredith thought that the text was socialist, injudiciously

provocative, and full of indiscriminate satire. As a result, Hardy gave up the idea of publishing it and it is commonly said that he burned it [40]. Nevertheless, many critics often have argued that the novel may have been partly drawn on for the short story An Indiscretion in the Life of an Heiress [41, 42]. However, Hardy always stressed that the short story is different from the novel [43]. The text of this work is based on Weber's edition of Hardy' lost novel. Weber [44] claimed that there was an earlier record for the novel Hardy destroyed. He realized that the novel deserves a critical attention; therefore, he revived it. He gathered information from six sources together which equipped him, as he claims, with a detailed knowledge of The Poor Man and the Lady and used it for the synopsis of the novel he produced.

A corpus was thus built from the electronic texts of the novels and short stories. Codes were used in reference to the texts as shown in Table I.

TABLE I. THE CORPUS

Code	Title
Hardy01	Desperate Remedies
Hardy02	Under the Greenwood Tree
Hardy03	A Pair of Blue Eyes
Hardy04	Far from the Madding Crowd
Hardy05	The Hand of Ethelberta
Hardy06	The Return of the Native
Hardy07	The Trumpet-Major
Hardy08	A Laodicean
Hardy09	Two on a Tower
Hardy10	The Mayor of Casterbridge
Hardy11	The Woodlanders
Hardy12	Tess of the D'Urbervilles
Hardy13	Jude the Obscure
Hardy14	The Well-Beloved
Hardy15	The Poor Man and the Lady
Hardy16	Wessex Tales
Hardy17	Life's Little Ironies
Hardy18	A Group of Noble Dames
Hardy19	A Changed Man and other Stories
Hardy20	Excluded and Collaborative Stories

It was decided to consider each of the collection of short stories as a single document. Short stories were not considered as separate documents. The rationale is that concept mining is more effective with long documents. In this regard, it was thought that concept mining would work better with the collections of short stories than individual stories.

VI. DATA ANALYSIS

For the extraction of the concepts, concept-based text representation was used. The documents were converted into clauses or what is referred to as 'bag of concepts'. Documents were represented as strings of concepts, where each document was represented by a given number of vectors. For this purpose, the study adopts the extraction model developed by Kim, et al. [45] shown in Fig. 1.

Given the high dimensionality of the corpus, it becomes impossible for any concept extraction or mining system to deal with these huge datasets effectively. In concept extraction applications, just like other clustering applications, high dimensionality is a serious problem that has adverse impacts on the reliability of the clustering performance [46]. With high dimensionality data in the clustering applications to literary texts, semantic similarity or relatedness is not accurately computed [47].

In the face of this problem, dimensionality reduction is carried out. The purpose is to keep only the distinctive features or variables. Also, concept-frequency inverse document frequency (CF-IDF) is used. CF-IDF is a weighting scheme for discovering the key concepts within datasets that is based on term-frequency inverse document frequency (TF-IDF) that tends to rank the concepts based on their frequency in relation to document frequency [48-50].

As a final step, semantic similarity between the texts is computed. Thus is a process whereby metrics are used for weighting or ranking similar concepts based on a concept taxonomy [51]. Semantically similar concepts can be thus grouped or classified together as shown in Fig. 2.

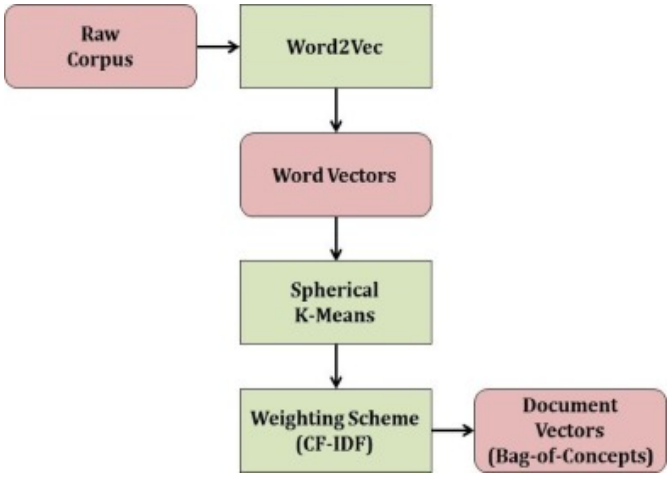


Fig. 1. Bag of Concepts Model Developed by Kim, et al. [45].

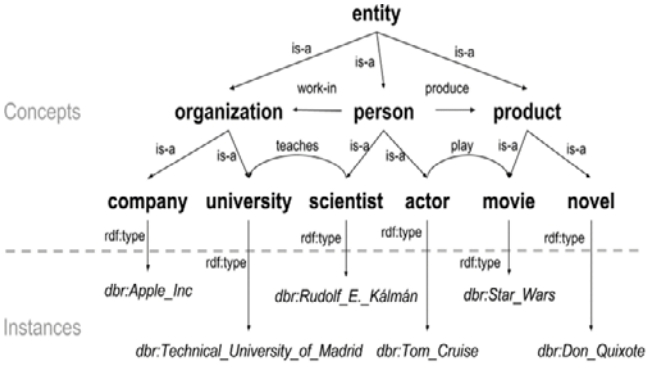


Fig. 2. An Example of Computing Semantic Similarity [51].

Most of metrics are designed for computing the semantic similarity with focus on the structure of the semantic network between concepts (e.g., path length and depth), or only on the Information Content (IC) of concepts [51]. These metrics, however, are not appropriate for the thematic typologies of literary texts which require identifying the conceptual similarities within texts. In so doing, the method developed by Zhu and Iglesias [51] for computing the conceptual similarity within texts is used for clustering and categorizing the selected texts based on their type concepts.

The selected texts are categorized based on their conceptual similarity into five categories. These are class conflict, Wessex, religion, female suffering, and social realities. First, the concept of class conflict is represented through romantic relations, mismatched marriages, elopement, and inequalities. Texts in this group included The Poor Man and the Lady, Life's Little Ironies, Far From the Madding Crowd, and A Group of Noble Dames. Second, the concept of Wessex is associated with a number of other concepts including rural life, local traditions, and industrialization. The concept of Wessex is extensively represented in Desperate Remedies, Under the Greenwood Tree, A Pair of Blue Eyes, The Hand of Ethelberta, The Trumpet Major, A Laodicean, and The Well-Beloved. Third, the concept of religion is represented through Christian beliefs and rituals, faith, morality, human existence, spiritual doubt, Evangelicalism and biblical references. Texts grouped under the concept of religion included Jude the Obscure, Tess of the D'Urbervilles, and Wessex Tales. Fourth, the concept of female suffering is represented through divorce, sexuality, struggle of women, oppression, and loss of chastity. Texts in this group included Two on a Tower, An Indiscretion in the Life of an Heiress, and A Changed Man and Other Tales. Finally, the concept of social realities is represented through poverty, social security, inequalities, and forces of oppression. Texts categorized under the heading of social realities included The Woodlanders. The Return of the Native and The Mayor of Casterbridge.

VII. Analysis and Discussions

Based on computing the conceptual similarity within the datasets, the texts were grouped into five main categories. These included class conflict, Wessex, religion, female suffering, and social realities. These are shown as follows.

A. Class Conflict

Class conflict is one of the central concepts in Hardy's novels and short stories including The Poor Man and the Lady and A Group of Noble Dames. In The Poor Man and the Lady, for instance, Hardy describes the ugly face of class conflict. This is represented in the love story between Miss Allamont and Will Strong. Miss Allamont is the squire's daughter and his heiress and Will Strong is a son of a peasant working on the estate of the Squire. In spite of the class gap between both, Miss Allamont takes a romantic interest in Will, and this is strongly rejected by her parents. Being rejected, Will moves to London where he achieves a striking success and becomes a public figure. However, he is still rejected by the family. This leads the two lovers to marry in secret and live away from her family. Soon her life is endangered and she dies. It was thus obvious that Hardy did not like class differences of his age and tended to represent the hypocrisy of the Victorian age in his books. Hardy describes the sufferings of the lower classes and the severe laws that threaten their lives in these books [52].

B. Wessex

Wessex is a dominating theme in many of Hardy's novels and short stories. In these texts, Hardy stressed the death of England's rural life along with its old customs and local traditions. Many critics consider Hardy as the greatest novelist in the form of regional fiction, and they think that the best example of regional fiction is Hardy's Wessex novels [53-55]. In the Wessex texts, Bullen [56] argues, Hardy succeeded at linking human behavior with the physical world. He adds that the works of Thomas Hardy to the historic place he was concerned with in his writings indicating that he was nostalgic for the past of England and that he distrusted modern civilization.

C. Religion

Religion is one of the central themes in Hardy's prose fiction. In his novels and short stories, the Bible and biblical names are obviously frequent. Influenced by controversies of the age, Hardy used what came to be known as the evolutionary narrative envisaged an alternative to a narrative which assumes that God created the world in its present state. Hardy expressed morality in a unique way. Morality is not based on traditional Christian beliefs. Rather, it is a social construct enforced by human intelligence rather than divine authority [57].

Many critics claim that any thematic discussion of Hardy's works has to consider religion as a crucial element in understanding and interpretation [58]. Hardy's texts cannot be fully understood without critical considerations of religious background and influence. According to Fergusson [59], the avoidance of such religious dimensions in thematic discussions results in interpretative gaps and loss of many thematic concepts in the literary texts.

D. Female Suffering/Tragedy

The texts included in this category reflect Hardy's preoccupation with the Victorian women and their sufferings. Hardy's women are destined to suffer. They are victims to the merciless conditions of the age. Women's suffering was deeply rooted in the hypocrisies of the Victorian society which was male dominated and obsessed with the idea of woman virginity. Many critics have advocated the idea that the texts of Hardy address the low position of women in the Victorian society and the strict laws that tended to deprive them of their independence. Morgan [60] argues that Hardy's texts reflect his sympathy towards women and his deep concern with their sufferings. Hardy introduces the sufferings of his women with a peculiar pathos and shows them as victims of male-dominated society. Likewise, in her book The Feminist Sensibility in the Novels of Thomas Hardy, Kaur [61] stresses that Hardy had sympathy for the women's cause and their sufferings. She stresses that Hardy is 'feministic', an artist with feminist sensibilities.

The results of the study agree with different feminist readings of Thomas Hardy that consider the writings of Hardy as a cry against the injustices done to the Victorian woman,

and an assertion her rights. The feminist investigations of Hardy's work often involve a discussion of sexuality and the sensual dimensions of the texts [60, 62, 63]. The main assumption of feminist readings of Hardy is that his texts reflect in one way or another the individual and social pressures the Victorian woman had to experience [61, 64-67]. The central concept in such reviews is that Hardy's works both depict and resist the male-centered culture and the oppression of the Victorian woman. At this point, much of the feminist reading of Hardy's prose fiction praises his progressive exploration, understanding, and support of women issues at a time of social crisis and change [66, 68].

E. Social Realities

In this class of novels and short stories, Thomas Hardy was concerned with depicting the contemporary social issues of his age. Levine [69] argues that that Hardy was a Victorian social critic since his writings depict the sufferings of England's working class and society's responsibility for their tragic fates. That is, Hardy's mind was preoccupied with improving conditions of society. He marks Hardy as a realistic writer who thought his role to express the joys and woes of the victims of merciless conditions of life. Likewise, Reid [70] argues that Hardy's works represent a cry against the excesses of modern civilization and injustices of modern societies.

It can be concluded that concept mining methods can be used for identifying the conceptual similarities within texts and generating reliable typologies of the novels and short stories of Thomas Hardy. Texts were successfully categorized based on their thematic concepts of class conflict, Wessex, religion, female suffering, and social realities. Although the thematic typology of Thomas Hardy based on the concept mining methods agrees in principle with the previous classifications based on fundamental philological approaches, the results reported here are testable, replicable, and thus reliable.

VIII. Conclusion

This study reviewed the literature regarding the thematic typologies of literary texts focusing on the thematic classification of Thomas Hardy's novels and short stories. It was obvious that the typologies of literary texts including those of Thomas Hardy have been traditionally based on philological methods with no consideration of empirical methodologies. With the development of computational approaches, quantitative and statistical methods have come into use. The majority of these typologies, however, are largely based on standard clustering methods or more specifically VSC theory. Despite the effectiveness of this methodology in generating classifications that are based on objective and replicable methods, underlying meanings and concepts were not fully explored. This is attributed to the absence of context in VSC applications which are largely based on 'bag of words' methods. In the face of these limitations, this study proposed a thematic typology based on concept mining methods. The findings indicated that concept mining was usefully used in generating a thematic typology of the novels and short stories of Thomas Hardy that revealed the thematic patterns of the texts. These thematic patterns would be best described in these categories: class conflict, Wessex, religion, female suffering, and social realities. Although the study was limited to the novels and short stories of Thomas Hardy, the results can be extended to other literary texts. It can be finally suggested that the computational and quantitative methods will be central components in the future of thematic typology research.

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 fulfilment of the current research project.

References

[1] R. G. Potter, "Literary Criticism and Literary Computing: The Difficulties of a Synthesis," Computers and the Humanities, vol. 22, no. 2, pp. 91-97, 1988.

[2] S. Gilmartin, Thomas Hardy's Shorter Fiction: A Critical Study. Edinburgh: Edinburgh University Press, 2007.

[3] R. D. Morrison, Thomas Hardy: A Companion to the Novels. McFarland, Incorporated, Publishers, 2021.

[4] S. Gatrell, Thomas Hardy's vision of Wessex. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2003, pp. xvii, 264 p.

[5] P. Gossin, Thomas Hardy's Novel Universe: Astronomy, Cosmology, and Gender in the Post-Darwinian World. London; New York: Routledge, 2017.

[6] W. Lu, Y. Zhou, J. Yu, and C. Jia, "Concept Extraction and Prerequisite Relation Learning from Educational Data," in The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-19), 2019, pp. 9678- 9685: Association for the Advancement of Artificial Intelligence.

[7] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," Journal of Biomedical Informatics, vol. 100, p. 100057, 2019/01/01/ 2019.

[8] M. M. Louwerse and W. van Peer, Thematics: Interdisciplinary Studies. John Benjamins Publishing Company, 2002.

[9] A. García-Berrio, A Theory of the Literary Text. De Gruyter, 2016.

[10] B. Plietzsch, "Hardy's Classifcation of his Works " 2003.

[11] E. Gosse, "Thomas Hardy," in Thomas Hardy; The Critical Heritage, R. G. Cox, Ed. (Critical heritage series, New York: Barnes & Noble. First published in The Speaker (13 September 1890). , 1970, pp. 167-172.

[12] T. Hardy, The works of Thomas Hardy in prose and verse. With prefaces and notes. (Wessex edition.). London: Macmillan & Co, 1912, p. 23 vol.: plates; maps. 23 cm.

[13] B. Plietzsch, The Novels of Thomas Hardy as a Product of Nineteenth Century Social, Economic, and Cultural Change. Berlin: Tenea Verlag Ltd, 2004.

[14] G. Harvey, The complete critical guide to Thomas Hardy (The complete critical guide to English literature). London: Routledge, 2003, pp. x, 228 p.

[15] P. Widdowson, Hardy in history : a study in literary sociology. London ; New York: Routledge, 1989, p. 260.

[16] L. Abercrombie, Thomas Hardy. A critical study. Martin Secker: London, 1912, p. 8º.

[17] P. Widdowson, On Thomas Hardy : late essays and earlier. Basingstoke: Macmillan, 1998, pp. x, 218.

[18] A. Omar, "Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction," Journal of the Chicago Colloquium on Digital Humanities and Computer Science, vol. 1, no. 2, pp. 1-12, 2010.

[19] A. Omar, Addressing Subjectivity in Thematic Classification of Literary Texts: A Fresh Look at the Prose Fiction of Thomas Hardy. Berlin: Lap Lambert Academic Publishing, 2015.

[20] A. Omar, "Identifying Themes in Fiction: A Centroid-Based Lexical Clustering Approach," Journal of Language and Linguistic Studies, vol. 17, no. Special Issue 1, pp. 580-594, 2021.

[21] M. Ford, Thomas Hardy: Half a Londoner. Harvard University Press, 2016.

[22] S. Shehata, Concept Mining: A Conceptual Understanding Based Approach. Waterloo, Ontario: University of Waterloo, 2009.

[23] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," in Intelligent natural language processing: Trends and Applications: Springer, 2018, pp. 373-397.

[24] G. M. Borkar, L. H. Patil, D. Dalgade, and A. Hutke, "A novel clustering approach and adaptive SVM classifier for intrusion detection in WSN: A data mining concept," Sustainable Computing: Informatics and Systems, vol. 23, pp. 120-135, 2019.

[25] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high ‐ dimensional databases: A review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, p. e1300, 2019.

[26] M. Looks, A. Levine, G. A. Covington, R. P. A. L. R. P. Loui, and J. W. Lockwood, "Streaming Hierarchical Clustering for Concept Mining," in Aerospace Conference, 2007 IEEE, 2007, pp. 1-12.

[27] L. Fang, M. Mehlitz, F. Li, and H. Sheng, "Web Pages Clustering and Concepts Mining: An approach towards Intelligent Information Retrieval," Cybernetics and Intelligent Systems, 2006 IEEE Conference, pp. 1-6, 2006.

[28] J. Han and M. Kamber, Data mining : concepts and techniques. San Francisco, Calif. ; London: Morgan Kaufmann, 2001, pp. xxiv, 550 p.

[29] K. Li, H. Zha, Y. Su, and X. Yan, "Concept Mining via Embedding," in 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 267-276: IEEE.

[30] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge: Cambridge University Press, 2019.

[31] C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[32] M. W. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval. Springer New York, 2013.

[33] T. Hardy, A group of Noble Dames. New York: Harper and brothers, 1891, pp. 5 p.l., 292 p.

[34] T. Hardy, Life's Little Ironies. Leipzig: B. Tauchnitz, 1894, p. 295 p.

[35] T. Hardy, Wessex Tales. New York: Harper & brothers, 1896, pp. vii, 290, [1] p.

[36] T. Hardy, A Changed Man, The Waiting Supper, and Other tales, Concluding with The Romantic Adventures of a Milkmaid (Thomas Hardy's works. The Wessex novels.). London: Macmillan & co., 1913, pp. vii, 412, [1] p.

[37] P. Widdowson, "Into the Hands of Pure-minded English Girls": Hardy's Short Stories and the Late Victorian Literary Marketplace," in A companion to Thomas Hardy, K. Wilson, Ed. no. Blackwell companions to literature and culture) Malden, MA: Wiley-Blackwell Pub., 2009, pp. 364-378.

[38] T. Hardy, A Group of Noble Dames, Wessex Edition ed. (The works of Thomas Hardy in Prose and Verse. With prefaces and notes. (Wessex Edition.). [The Wessex novels. II. Romances and fantasies.]). London: Macmillan and Co, 1912, p. 235.

[39] P. Dalziel, "Thomas Hardy: The Excluded and Collaborative Stories." Oxford: Clarendon Press, 1992, p.^pp. Pages.

[40] E. Gosse. (1928, January 22) Thomas Hardy's Lost Novel. London Times.

[41] T. Coleman, "An Indiscretion in the Life of an Heiress." London: Hutchinson 1976, p.^pp. Pages.

[42] R. G. Cox, Thomas Hardy; the critical heritage (Critical heritage series). New York: Barnes & Noble, 1970, pp. xlvii, 473 p.

[43] R. L. Purdy and M. Millgate, "The Collected Letters of Thomas Hardy (hereafter Collected Letters)." Oxford: Clarendon Press, 1978, p.^pp. Pages.

[44] C. J. Weber, "An Indiscretion in the Life of an Heiress. Hardy's "lost novel" now first printed in America and edited with introduction and notes by Carl J. Weber." Baltimore, MD.: The Johns Hopkins Press. , 1935, p.^pp. Pages.

[45] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," Neurocomputing, vol. 266, pp. 336-352, 2017/11/29/ 2017.

[46] A. Omar, "Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 99-107, 2020.

[47] A. Omar, "Classifying literary genres: a methodological synergy of computational modelling and lexical semantics," Texto Livre: Linguagem e Tecnologia, vol. 13, no. 2, pp. 83–101, 2020.

[48] S. Agarwal, A. Singhal, and P. Bedi, "Classification of RSS feed news items using ontology," in 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 491-496: IEEE.

[49] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong, "News recommendations using CF-IDF," in Proceedings of the International Conference on Web Intelligence, Mining and Semantics 2011 (WIMS 2011), 2011.

[50] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak, "News personalization using the CF-IDF semantic recommender," in Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011, pp. 1-12.

[51] G. Zhu and C. A. Iglesias, "Computing Semantic Similarity of Concepts in Knowledge Graphs," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 1, pp. 72-85, 2017.

[52] I. Clark, Thomas Hardy's Pastoral: An Unkindly May. Palgrave Macmillan UK, 2016.

[53] C. Pickford, The Rural-urban Bind in Thomas Hardy's Regional Novels. Manchester Metropolitan University, 2012.

[54] R. Pite, Thomas Hardy: Selected Writings. Oxford: Oxford University Press, 2021.

[55] K. Snell, The Bibliography of Regional Fiction in Britain and Ireland, 1800–2000. London; New York: Routledge, 2017.

[56] J. B. Bullen, Thomas Hardy: The World of his Novels. Frances Lincoln, 2013.

[57] N. Vance, Bible and Novel: Narrative Authority and the Death of God. Oxford: Oxford University Press, 2013.

[58] R. Franklin, Thomas Hardy and Religion: Theological Themes in Tess of the D'Urbervilles and Jude the Obscure. Eastbourne: East Sussex: Sussex Academic Press, 2021.

[59] D. Fergusson, Faith and Its Critics: A Conversation. Oxford: Oxford University Press, 2011.

[60] R. Morgan, Women and Sexuality in the Novels of Thomas Hardy. London; New York: Routledge, 2006.

[61] M. Kaur, The Feminist Sensibility in the Novels of Thomas Hardy. New Delhi: Sarup & Sons, 2005.

[62] T. R. Wright, Hardy and the Erotic (Macmillan Hardy Studies). Palgrave Macmillan 1989.

[63] M. R. Higonnet, The Sense of sex: feminist perspectives on Hardy. Urbana: University of Illinois Press, 1993, p. 270 p.

[64] P. Ingham, Thomas Hardy: A Feminist Reading Hemel Hempstead: Harvester Wheatsheaf, 1989.

[65] J. Thomas, Thomas Hardy, Femininity and Dissent: Reassessing the Minor Novels. New York: Macmillan, 1999.

[66] M. Jacobus, "Tess's Purity," Essays in Criticism, vol. 26, pp. 318-38, 1976.

[67] M. Jacobus, " Women Writing and Writing about Women." London: Croom Helm, 1979, p.^pp. Pages.

[68] P. Boumelha, Thomas Hardy and Women: Sexual Ideology and Narrative Form. Brighton: Harvester Wheatsheaf, 1982.

[69] G. Levine, Reading Thomas Hardy. Cambridge: Cambridge University Press, 2017.

[70] F. Reid, Thomas Hardy and History. Springer International Publishing, 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, digital humanities, discourse analysis, and translation studies. ORCID: 0000-0002-3618-1750.

Educational Data Mining in Predicting Student Final Grades on Standardized Indonesia Data Pokok Pendidikan Data Set

Nathan Priyasadie, Sani Muhammad Isa
BINUS Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta
Indonesia, 11480

***Abstract*—Educational Data Mining has been implemented in predicting student final grade in Indonesia. It can be used to improve learning efficiency by paying more attention to students who are predicted to have low scores, but in practice it shows that each algorithm has a different performance depending on the attributes and data set used. This study uses Indonesian standardized students' data named Data Pokok Pendidikan to predict the grades of junior high school students. Several prediction techniques of K-Nearest Neighbor, Naive Bayes, Decision Tree and Support Vector Machine are compared with implementation of parameter optimization and feature selection on each algorithm. Based on accuracy, precision, recall and F1-Score shows that various algorithm performs differently based on the high school data set, but in general Decision Tree with parameter optimization and feature selection outperform other classification algorithm with peak F1-Score at 61.48% and the most significant attribute in are First Semester Natural Science and First Semester Social Science score on predicting student final score.**

Keywords*—Educational data mining; student performance; classification models; feature selection; parameter optimization

I. Introduction

Educational data mining is a rapidly growing multidisciplinary area of study devoted to studying and developing techniques for extracting useful information from enormous amounts of data generated in educational settings [1]. Because information technology has been significant in improving the area of education over the last decade, nearly every institution now maintains a student information system [1]. This information includes student demographic, parent information, scores etc. Applying data mining techniques to educational processes can be beneficial in identifying important trends, performance summaries, and insights, which will assist students in identifying areas for improvement. An institution's academic performance, life cycle management, course selection, retention rate measurement, and grant money management may all be considered [2].

Predicting student grades is one of educational mining's applications. Grades are critical components of education since they act as a barometer of a student's competency and performance within that institution. Predicting a student's final grade might also encourage a school to improve its teaching techniques and create a more pleasant learning environment [1]. By providing additional support to students who were previously projected to have lower grades, it is possible to enhance learning efficiency and the overall student grade [3]. Finally, a high score improves a student's chances of admission to a more prestigious higher education program.

Data from Organization for Economic Co-operation and Development (OECD) shows that Indonesian student ranked 72 out of 77 countries on Programme for International Student Assessment (PISA) report in 2018, and this rank tends to stagnate for the last 10-15 years. It can be concluded that education in Indonesia is still lagging compared to other countries.

The purpose of prediction is to determine the value of an unknown variable that correspond to the student [1]. In Indonesia there have been several researches that investigate student performance prediction using Naive Bayes (NB) [4], Decision Tree (DT) [5], Support Vector Machine (SVM) [6], K-Nearest Neighbor (K-NN) [7] and Regression Analysis [8] but there is no research that predicts student grades by using standardized national socio-demographic aspects of students such as the Data Pokok Pendidikan (DAPODIK).

The purpose of this study is to find the best algorithm to predict student final score using standardized DAPODIK data combined with student historical grade from three public junior high schools in Indonesia. With standardized data, schools throughout Indonesia can determine the best method to predict student grades in their schools. It can be used to improve learning efficiency by providing additional support to students who were previously projected to have lower grades. This study will compare four different algorithms, Naive Bayes, Decision Tree, K-Nearest Neighbors and Support Vector Machine with two data mining optimization methods, parameter optimization (PO) and feature selection (FS).

II. Related Work

Mengash [9] in their study found that using Artificial Neural Networks to predict student performance of 2039 Computer Science students at a Saudi Public University from 2016 to 2019 had an accuracy rate of greater than 79%, outperforming other classification techniques such as Decision Trees, Support Vector Machines, and Naive Bayes. It compares various pre-admission criteria (high school grade average, Scholastic Achievement Admission Test score, and General

Aptitude Test score). The findings indicate that the Scholastic Achievement Admission Test score is the most reliable predictor of future student performance of any pre-admission criteria. As a result, admissions systems should give this score a higher weight.

Rifat et al. [10] perform research to predict students' performance using transcript data from a Bangladeshi institution. The authors utilized six cutting-edge classification algorithms (Gradient Boosted Tree, Random Forest, Tree Ensemble, Decision Tree, Support Vector Machines and K-Nearest Neighbor) to forecast students' final grades. The findings indicated that the Random Forest algorithm performed the best, with an accuracy of 94.1%, followed by the Tree Ensemble method.

Yao et al. [11] perform research to determine the final score of secondary school students utilizing their personal data. The data set contains a variety of factors, including parent information, student health status, financial status and. attendance etc. With feature selection, the J48 algorithm achieved the highest accuracy of 84.39%, whereas without feature selection, the OneR algorithm achieved the highest accuracy of 84.19%.

Saa et al. [12] gathered data on student demographics, course teacher information, student general information, and prior performance from a private institution in the United Arab Emirates using various algorithms (Decision Tree, Random Forest, Gradient Boosted Trees, Deep Learning, Naive Bayes, Logistic Regression and Generalized Linear Model). With 75.52% accuracy, the Random Forest method topped the other classifiers, followed by the Logistic Regression technique.

Fairos et al. [13] conduct research to predict student performance using Universiti Technology Cawangan Kelantan and Universiti Teknologi MARA Cawangan Negeri Sembilan student data with total 631 transcript from 2013 to 2016, with various attributes such as gender, all the course enrolled by student including the course grade. They develop a model to predict student performance using K-Nearest Neighbor, Naive Bayes, Decision Tree and Logistic Regression Model. It shows that Naive Bayes outperform other classification algorithm with 89.26% accuracy.

Based on previous study shows that data set plays a big role in determining which algorithm is the best for predicting student final score. On this research a standardized data set is used to determine which algorithm is best to be applied throughout high school in Indonesia.

III. Methodology

This research uses The Cross Industry Standard Process for Data Mining (CRISP-DM) [14]. CRISP-DM is the most used methodology for developing Data Mining projects; it consists of six steps as visualized in Fig. 1. The first step is business understanding where the purpose is to provide context for the objectives and data. The second step is data understanding where its purpose is to determine what can be expected and accomplished from the data. The third step is data preparation where it involves cleaning, integrating, and formatting the data [15]. The fourth step is modelling where the Naive Bayes, Decision Tree, Support Vector Machine and K-Nearest Neighbor algorithm are used then optimized using feature selection and parameter optimization method to produce the best prediction model. The last step is Evaluation of each model based on accuracy, precision, recall and F1-Score.

A. Data Understanding

The first step is to collect data from various sources, locate and gather data for training and validate the algorithm, which may be spread over many spreadsheets, databases, or webpages. This research uses data from 3 high schools from the Jakarta class of 2020 and 2019. With a total of 926 student data each with 33 variables, in xlsx format with Table I attributes.

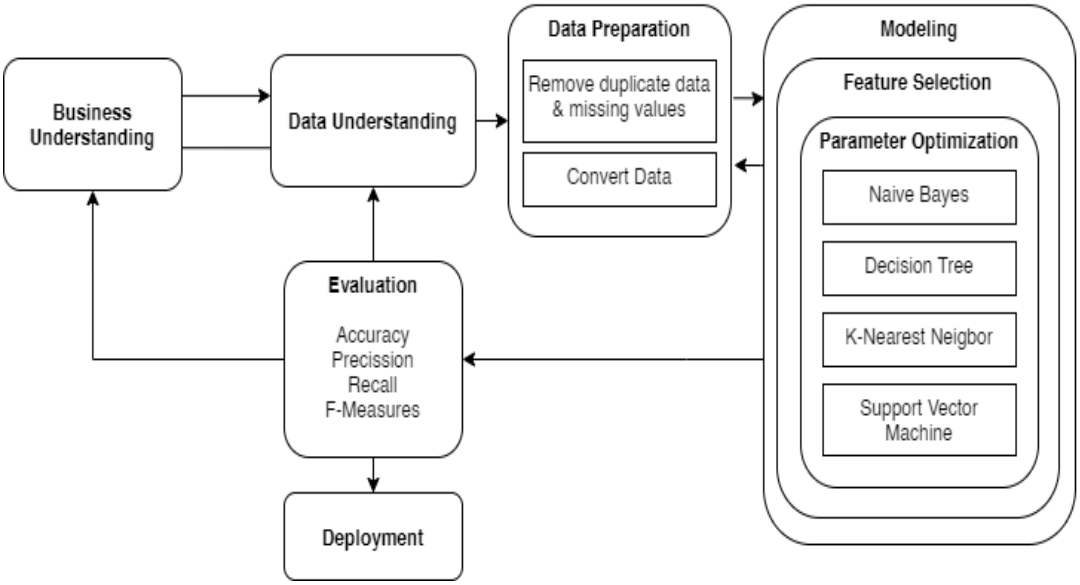


Fig. 1. CRISP-DM Methodology.

TABLE I. DATA ATTRIBUTES

Variables	***Description***	***Possible Value***
Final Grades	Average of Student's Final Grade	A, B, C, D, E, F, G
Entrance Grades	Student's Final Grade in Primary School	A, B, C, D, E, F, G
Gender	Student's Gender	Male, Female
Type of living	Student's living types	Living with Parents, Boarding House, Living with Guardian Others
Transportation Method	Student's Transportation Method to School	Car, Motorcycle, Bicycle, Public Transportation, Taxibike, On Foot, Others
Father's Education	Father's latest Education	None, Primary School, Junior High school, Senior High school, Diploma, Bachelor's Degree, Master's Degree
Father's Occupation	Father's latest Occupation	General employees, Entrepreneur, Merchant, Deceased, Laborer, Government Employees/Soldiers/Police, Others
Father's Income	Father's Monthly Income	No Income, <Rp 500.000, Rp 500.000 - Rp 999.999, Rp 1.000.000 - Rp 1.999.999, Rp 2.000.000 - Rp 4.999.999, Rp 5.000.000 - Rp 20.000.000, >Rp 20.000.000
Mother's Education	Mother's latest Education	None, Primary School, Junior Highschool, Senior Highschool, Diploma, Bachelor's degree, Master's degree
Mother's Occupation	Mother's latest Occupation	General employees, Entrepreneur, Merchant, Deceased, Laborer, Government Employees, Soldiers/Police, Others
Mother's Income	Mother's Monthly Income	No Income, <Rp 500.000, Rp 500.000 - Rp 999.999, Rp 1.000.000 - Rp 1.999.999, Rp 2.000.000 - Rp 4.999.999, Rp 5.000.000 - Rp 20.000.000, >Rp 20.000.000
First and Second Semester's Grades	Grades in Religion, Civic, Bahasa Indonesia, English, Mathematics, Natural Science, Social Science, Art and Culture, Sports	A, B, C, D, E, F, G

After being collected, the data is combined into a single data set.

B. Data Preparation

This step is to delete duplicate data or have empty attributes leaving 759 student data remaining. Then, numerical attributes are converted into categorical attributes based on Table II rules.

TABLE II. MAPPING RULES

Numerical Values	**Categorical Values**
Above 95	A
90-94	B
85-89	C
80-84	D
75-79	E
70-74	F
Below 70	G

Then the data is divided into two categories: training and testing, The training data set comprises 80% of the total data will be used to train the algorithm for classifying student data, whereas the testing data set comprises 20% will be used to evaluate the trained model's performance.

C. Modelling

Classification is a concept that refers to the act of classifying things according to information about one or more of its attributes, as well as categorizing them according to a collection of already classified items [16]. This research uses RapidMiner software which has a large collection of classification and optimization methods [17].

Naive Bayes is one of the simplest and most frequently used classification methods [18]. This method is based on Bayesian theory of probability, which assumes that a class is independent of each other [19]. With a simple concept, Naive Bayes uses a conditional probability model with P(six) as the probability of the class and assumes that the value of a predictor (x) in a particular class (c) does not depend on the value of other predictors. Naive Bayes can be described in the following equation 2.2.

$$ax(t) = Ax(t) + Bu(t) + B1w(t) \quad (1)$$

Naive Bayes has the advantages of being fast and efficient in using memory, able to handle quantitative data and discrete data, resilient to noise and only requires a small amount of data for classification and can handle missing values by ignoring values during probability calculations [20].

Decision Tree is a method for classifying by looking for differences between classes and dividing them using attributes by making a diagram in the form of a tree. This method uses a divide-and-conquer approach, one of the advantages is the ease of reading the model that has been made, with this convenience, information related to the identification of important attributes and relationships between classes can be used for analysis and research in the future [21]. By splitting to determine branches, there are several ways that can be used, such as Gini Impurity which looks for branches that have the most homogeneous results, which means that the results of the division have similar characteristics.

K-Nearest Neighbor performs classification by comparing input with training data like it, each data consists of n-attributes represented by a point on the n-dimensional graph, if given a data whose class is not known then K-Nearest Neighbor will look for several k training data closest to the location with the data [19] . After knowing the number of closest samples, the algorithm can estimate the class of the data based on the number of closest samples, the distance can be measured using several formulas such as Euclidean Distance. The advantage of K-Nearest Neighbor is that it can group a lot of data efficiently and in a fast time [20], but it has the disadvantage that it can become significantly slower with an increasing amount of data.

Support Vector Machine is a method that can be used to classify linear and non-linear data [22]. The way it works is by doing non-linear mapping to change the training data to a higher dimension, in this dimension he looks for the most optimal linear hyperlane separator. With enough nonlinear mappings that have high dimensions, data from the two classes can always be separated by hyperlane. Support Vector Machines finds hyperlane using support vectors and margins [19]. The advantages of Support Vector Machines are that it works well if there is a clear distance between class differences, effective for cases where the number of dimensions is more than the number of sample data, but the disadvantages are that it is not suitable for large data sets and does not perform well for data sets that have a lot of noise.

Feature selection reduces the number of dimensions of the data set thereby reducing processor and memory usage [23]. With this feature selection removes irrelevant attributes from the data set and improves the accuracy of the algorithm. For this study forward selection is used where it starts with an empty attribute set and adds attributes in it until the stopping criterion is met [24]. This method allows avoiding the use of additional memory and processor and improves the accuracy of the algorithm by removing irrelevant attributes from the data set.

Parameter optimization is a technique used to find the best combination of parameters to get the optimum performance of each algorithm. In the approach there are several ways such as through the grid, evolutionary and quadratic. By running iterations according to the provisions, then trying to calculate new parameters that may be between the previous parameters, and after that compare the results of the accuracy of the initial parameters and the parameters of the calculation results. The grid search is originally an exhaustive search based on defined subset of the hyper-parameter space. The hyper-parameters are specified using minimal value (lower bound), maximal value (upper bound) and number of steps [25]. In this case the grid search is used since those best ranges and dependencies are known.

D. Evaluation

In this research method the evaluation will be carried out using multi-class confusion matrix to evaluate each model accuracy, precision, recall and F1-Score.

E. Result and Analysis

Tables below summarizes the algorithm performance for Naive Bayes, Decision Tree, Support Vector Machines and K-Nearest Neighbor without any optimization, and with both feature selection and parameter optimization on different high school testing data set. For more concise tabulation the methods Naïve Bayes, Decision Tree, Support Vector Machines, K-Nearest Neighbor, combined feature selection and parameter optimization are abbreviated to NB, DT, SVM, K-NN and FS+PO, respectively.

On Table III shows result for high school A data set, Decision Tree with optimization shows to be the best overall algorithm and with best F1-Score at 54.01% and the most significant attribute on that algorithm is *First Semester Social Science*, *Second Semester English* and *Gender*, while K-Nearest Neighbor with optimization achieved best accuracy score at 77.36% while Naive Bayes with optimization achieved best recall score at 52.41%.

On Table IV shows result for high school B data set, Decision Tree with optimization shows to be the best overall algorithm and with best accuracy at 85.71% and precision at 64.40% and the most significant attribute is First Semester Religion score, First Semester Natural Science score, First Semester Sports score, First Semester Arts score, First Semester Social Science score and Second Semester Arts score while Naïve Bayes without optimization achieved best recall score at 61.11%.

On Table V shows result for high school C data set, Naive Bayes with optimization shows to be the best overall algorithm on all measurements and the most significant attributes are First Semester Natural Science score, First Semester Social Science score and Gender.

TABLE III. PERFORMANCE ON HIGH SCHOOL A TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	69.81%	60.60%	48.72%	54.01%
	PO + FS	76.92%	55.47%	52.41%	53.89%
DT	-	62.26%	55.90%	39.12%	46.02%
	PO + FS	73.58%	62.46%	50.63%	55.92%
SVM	-	66.04%	32.72%	36.22%	34.38%
	PO + FS	71.70%	60.58%	44.37%	51.46%
K-NN	-	67.92%	58.58%	42.29%	49.11%
	PO + FS	77.36%	38.54%	44.57%	41.33%

TABLE IV. PERFORMANCE ON HIGH SCHOOL B TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	76.79%	53.12%	61.11%	56.70%
	PO + FS	83.93%	60.60%	61.01%	60.80%
DT	-	66.07%	43.75%	42.56%	43.14%
	PO + FS	85.71%	64.40%	58.83%	61.48%
SVM	-	66.04%	32.72%	36.22%	34.38%
	PO + FS	82.14%	60.91%	57.54%	59.17%
K-NN	-	69.64%	47.62%	44.15%	45.81%
	PO + FS	78.57%	53.08%	57.00%	54.97%

TABLE V. PERFORMANCE ON HIGH SCHOOL C TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	48.78%	42.39%	51.62%	46.55%
	PO + FS	79.41%	66.62%	66.62%	66.62%
DT	-	63.44%	50.91%	50.86%	58.88%
	PO + FS	70.73%	57.30%	64.93%	60.87%
SVM	-	60.98%	42.00%	36.70%	39.17%
	PO + FS	68.75%	41.86%	45.56%	43.63%
K-NN	-	58.54%	39.78%	38.17%	38.95%
	PO + FS	75.61%	45.65%	51.49%	48.39%

In general, the experiment shows that feature selection and parameter optimization improve the accuracy of the classifier algorithm up to 62.79%. However, it also shows that various algorithms show different accuracy results with different high school data set. Decision Tree with optimization shows to be the best overall combination to predict student performance on A and B high school data set with peak F1-Score at 61.48%, meanwhile Naive Bayes with optimization shows to be the best combination on high school C data set with 66.62% F1-Score And in almost every data set shows that the most significant attributes are First Semester Natural Science, First Semester Social Science score on predicting student final score.

IV. CONCLUSION

The result of this study found that: (a) Overall best F1-Score is achieved by Decision Tree with feature selection and parameter optimization. (b) In general parameter optimization and feature selection show to improve algorithm performance. (c) The most significant attributes in predicting student score are First Semester Natural Science score and First Semester Social Science score. (d) Even with the same attributes from different schools' data set each algorithm performs differently. With these results it can be concluded that the research has achieved its objectives. But there is a room of improvement on this research since there are lack of data varieties because we're only using data from single province in Indonesia.

REFERENCES

[1] C. Romero and S. Ventura, Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), 2010.

[2] M. Goyal and R. Vohra, "Application of Data Mining in Higher Education," International Journal of Computer Science Issues (IJCSI), 2012.

[3] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, "Analyzing Undergraduate Student's Performance," Computers & Education, 2017.

[4] M. Jannah, "Penerapan Data Mining Prediksi Nilai Ujian Nasional (UN) Siswa SMP Menggunakan Metode Naive Bayes," Jurnal Informatika, Manajemen dan Komputer, 2020.

[5] P. Mayadewi and E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," Seminar Nasional Sistem Informasi Indonesia, 2015.

[6] R. Thaniket, Kusrini and E. T. L., "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine," Jurnal Teknologi dan Rekayasa, 2020.

[7] S. R. Rani, S. R. Andani and D. Suhendro, "Penerapan Algoritma K-Nearest Neighbor untuk Prediksi Kelulusan Siswa pada SMK Anak Bangsa," Prosiding Seminar Nasional Riset Information, 2019.

[8] H. Susanto and Sudiyatno, "Data Mining Untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan Dan Prestasi Masa Lalu," Jurnal Pendidikan Vokasi, 2014.

[9] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," Institute of Electrical and Electronics Engineers, 2020.

[10] M. R. I. Rifat, A. A. Imran and A. S. M. Badrudduza, "Educational Performance Analytics of Undergraduate Business Students," I.J. Modern Education and Computer Science, 2019.

[11] Y. Yao, Z. Chen, S. Byun and Y. Liu, "Using Data Mining Classifiers to Predict Academic Performance of High School Students," Scientific and Practical Cyber Security Journal, 2019.

[12] A. A. Saa, M. Al-Emran and K. Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance," Advances in Intelligent Systems and Computing, 2019.

[13] W. Fairos, W. F. W. Yaacob, S. Azlin, S. Nasir, W. Faizah, N. M. Sobri and C. Mara, "Supervised data mining approach for predicting student performance," Indonesian Journal of Electrical Engineering and Computer Science, 2019.

[14] O. Marbán, G. Mariscal and J. Segovia, Data Mining and Knowledge Discovery in Real Life Applications, 2009.

[15] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, "CRISP-DM 1.0 - Step-by-step data mining guide," 2000.

[16] S. B. Imandoust and M. Bilandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," International Journal of Engineering Research and Application, pp. 605-610, 2013.

[17] S. Slater, S. Joksimovic, V. Kovanovic and R. Baker, "Tools for Educational Data Mining: A Review," Journal of Educational and Behavioral Statistics, pp. 85-106, 2017.

[18] C. C. Aggarwal and C. Zhai, "A Survey of Text Classification," 2012.

[19] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2012.

[20] Defiyanti, J. Sofi and Mohamad, "Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining," Konferensi Nasional Informatika (KNIF), 2015.

[21] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics, 2004.

[22] Boser, G. Bernhard E, V. Isabelle M and V. N, "A Training Algorithm for Optimal Margin Classifiers," Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.

[23] S. A. Ö. T. İ. Esra Mahsereci Karabulut, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technology, 2012.

[24] G. Borboudakis and I. Tsamardinos, "Forward-Backward Selection with Early Dropping," Journal of Machine Learning Research 20, 2019.

[25] I. Syarif, A. Prugel-Bennett and G. Wills, "SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance," Telecommunication Computing Electronics and Control, 2016.

Cyberbullying Detection in Textual Modality

Evangeline D[1], Amy S Vadakkan[2], Sachin R S[3], Aakifha Khateeb[4], Bhaskar C[5]
Assistant Professor, Department of ISE, M S Ramaiah Institute of Technology, Bangalore, India[1]
Student, Department of ISE, M S Ramaiah Institute of Technology, Bangalore, India[2, 3, 4, 5]

***Abstract*—Cyberbullying is the use of technology to harass, threaten or target another individual. Online bullying can be particularly damaging and upsetting since it is usually anonymous and it's often hard to trace the bully. Sometimes cyberbullying can lead to issues like anxiety, depression, shame, suicide, etc. Most of the cyberbullying cases are not revealed to the public and the number of cases reported to the legal system is only few. Certain victims do not reveal their bully experiences out of shame or due to difficult procedures for reporting to the legal system. Our cyberbullying detection system aims to bring cases involving cyberbullying under control by detecting and warning the bully. Such cases are also reported to appropriate authorities, which can then be verified and necessary actions can be taken depending on the situation. The technology stack used for implementation include Flask, Scikit learn, Chat application APIs, Firebase, HTML, Javascript and CSS. The model was tested on classifiers like SVM, KNN, Logistic regression and Random Forest. F1 score was used as a metric to assess the four models. While analyzing the performances of these models, it was observed that Random Forest Classifier outperformed all the models. F1 score of 93.48% was achieved using the Random Forest Classifier.**

***Keywords*—Cyberbullying detection; support vector machine (SVM); kNN (k nearest neighbor); logistic regression; random forest classifier**

I. Introduction

The most common type of online bullying is mean comments which includes use of aggressive and pejorative words, threats, profile hacking etc. [11][12][13][14]. Nearly 8 out of 10 individuals are subject to the different types of cyberbullying in India. Out of these around 63% faced online abuses and insults, and 59% were subject to false rumours and gossip for degrading their image. 64% of victims receive an aggressive instant message when they are bullied. 7 in 10 young people experience cyberbullying before they hit the age of 18. About 37% of children between 12 and 17 years experienced cyberbullying at least once. One in four children fall victim to cyber bullies. In just one year, cyberbullying of teenagers and Indian women has increased by 36%. Only 4.6% of the cases are reported to the authorities and the rest go unnoticed or are hidden by victims to save themselves from further damage. Cases of cyber stalking or bullying of women or children increased by more than 36% from 2018 to 2020, data released recently by the National Crime Records Bureau showed.

Most of the comments that are posted on social media are noticed by people, but a large number of cases involving cyberbullying in messages are not shown in the public by victims to protect themselves from shame. These events can severly impact the one getting bullied and can sometimes lead to suicide. Currently studies and projects that are carried out in this area only include models to classify single sentences as a comment on bullying or not. What differentiates our model from existing ones is that we capture the message, the context and details of the bully and report it to authorities. With the use of our cyberbullying detection system, messages indicating cyberbullying can be detected and reported so that such events do not go unnoticed.

The rest of the paper is organized as follows: Section II summarizes contemporary research works carried out in cyberbullying detection. While Section III elaborates the methodology employed, Section IV discusses the results as an outcome of our work carried out. Section V concludes the findings and details the future work.

II. Literature Review

Many existing approaches were proposed in [1]. Authors have worked on the detection of Cyberbullying over comments posted on Instagram. Preprocessing techniques performed include unimportant character removal and removal of stop words. The machine learning model used was the Linear Support Vector Machine (LSVM). The metrics used for evaluation were accuracy, precision and recall. The model was designed only for detection of highly negative social media posts, more features and detailed labelling surveys can improve accuracy. In [2], authors have worked on identifying tweets related to cyberbullying by using PHP and HTML with MySQL and Twitter API. The preprocessing steps carried out were removal of punctuations and emotional icons. The detection of cyberbullying in tweets was done using a simple keyword search, in which each word present in the tweet was compared with the words in the dataset. The model did not consider the context of tweet, accuracy, informal language and abbreviations.

In [3], authors worked on the detection of cyberbullying on ASKfm which is a platform that allows users to ask and answer questions anonymously. Data cleaning steps done were removal of white spaces and replacement of abbreviations. The preprocessing steps consisted of tokenization, POS tagging and lemmatization. The machine learning model used was SVM (Support Vector Machine) and the evaluation metric considered was F1 score. The model only detected cyberbullying, it did not warn the bully or report the same to authorities. In [4], authors worked on identification and classification of Cyberbully Incidents using Bystander Intervention Model. The proposed model focused mainly on the analysis of direct intervention by bystanders. The dataset used consisted of posts and activities from facebook. Data preprocessing steps included were removal of whitespaces and

stop words. The machine learning model used for implementation was Random Forest Classifier. A limited dataset with limited accuracy was used for binary classification. In [5], the authors worked on the design of Semantic Framework for detecting Cyberbullying on social media. The main focus of the paper was detection of cyberbullying using Semantic learning. Removing stop words and Tokenization were the preprocessing steps performed. Sentiment Analysis was done on comments from posts. The model concentrated only on binary classification of comments and did not include reporting of comments related to cyberbullying to authorities. In [6], the authors' approach is validated on a dataset of over 3000 images along with peer-generated comments posted on the Instagram photo-sharing network, running comprehensive experiments using a variety of classifiers and feature sets. In this work, methods for detecting cyberbullying in commentaries following shared images on Instagram are developed. Classification of images and captions themselves are potential targets for cyberbullies. Standard k-fold validation technique is used to train data. It is a lengthy process. The training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. In [7], the paper proposes a supervised machine learning approach for detecting and preventing cyberbullying. Several classifiers are used to train and recognize bullying actions. The evaluation of the proposed approach on cyberbullying dataset shows that Neural Network performs better. Preprocessing is done, and then feature extraction is performed. The extracted features are fed into a classification algorithm to train, and test the classifier and hence use it in the prediction phase. Two classifiers, namely, SVM and Neural Network are used. This model of detecting cyberbullying patterns is limited by the size of training data. Thus, a larger amount of cyberbullying data is needed to improve the performance. In [8], authors worked on aggressive text detection for cyberbullying. The proposed model automatically maps a document with an aggressiveness score, thus treating aggressive text detection as a regression problem and explores different approaches for this purpose. These include lexicon-based, supervised, fuzzy, and statistical approaches. The different methods were tested over a dataset extracted from Twitter and compared them against human evaluation. The results favored approaches that considered several features particularly the presence of swear or profane words. The approaches used could be refined to better handle difficult cases, testing more supervised approaches using a larger dataset and building a framework for cyberbullying automatic identification. In [9], the authors have focused on detection and mitigation of cyberbullying in English and Arabic. The dataset was taken from Facebook posts and Twitter feeds. Their preprocessing involved removing tweets other than English and Arabic. Naive Bayes and Support Vector Machine were used to build the model. Recall and Precision were used as a metric for evaluating the built model. Enhancing the performance measures achieved by the system by using hybrid training models, such as combinations of Distance Functions, NB and SVM would make the model more effective in detecting cyberbullying. In [10], the authors have used methodologies to extract texts sent by the user and network based attributes are used to study the properties of bullies and aggressors. Dataset was taken from youtube comments and Twitter handles. The preprocessing includes replacing abbreviations with full phrases. Multiple models were used such as Logistic Regression, SVM and Gradient Boost. These models were only able to achieve an accuracy between 70 and 75% and it was only able to give a binary classification.

Some gaps were identified in the works discussed in this section. In [1], cyberbullying detection is limited only to English and Arabic. In [2][3][8], reporting to authorities is not done at all. In [6][7], only a limited dataset was employed. The works mentioned in [5][6] performs only with limited accuracy. Hence, in our paper, we have focused on working with large dataset and improving accuracy. We have also worked on reporting to authorities.

III. METHODOLOGY

The system consists of two main parts, a backend that captures messages on a chat application and calculates the probability that the sent message belongs to categories like toxic, severe toxic, obscene, insult or identity hate and a front end which displays messages that have a high probability of the message belonging to one of the six categories mentioned. The system is designed with the following objectives.

- Detection of cyberbullying in text messages.
- Reporting to appropriate authorities for initiating suitable action.

The technology stack used for backend are Flask, Scikit learn and Firebase. Flask is a micro web framework written in Python which contains third-party python libraries used for developing web applications. Flask was considered for this system since it was easy to integrate it with the machine learning model which was also written in python. Scikit Learn is one of the most useful libraries for machine learning in python and contains several tools and implementations of many machine learning algorithms. This has been used to determine the probabilities of messages being a bully message with the help of a random forest classifier.

Firebase which is a Backend-as-a-service provided by Google was chosen as the database to store all the details related to cyberbullying messages sent on the chat application. Email notifications are also triggered when database updates are made. The front end web application that displays the dashboard is built using HTML, CSS and Javascript. This dashboard is used by authorities to view details associated with the cyberbully messages.

A. Dataset

Toxic comment dataset was chosen for this purpose which was sourced from Wikipedia's talk page edits. This dataset was posted on Kaggle by the Conversation AI team, a research initiative founded by Jigsaw and Google. It essentially consists of 1,59,571 rows and 7 columns. Each row has a unique comment along with its comment id and the labels that it belongs to. Various labels that describe the comment in the dataset are - toxic, severe_toxic, obscene, threat, insult and identity_hate. Libraries used include numpy, re, panda, nltk, Matplotlib, wordcloud, sklearn, pickle.

B. Exploratory Analysis

Exploratory analysis is the process of analysing the dataset and displaying it in the form of graphs for better understanding of how the data is structured. A bar graph is plotted to see the unique comments that fall under different categories.

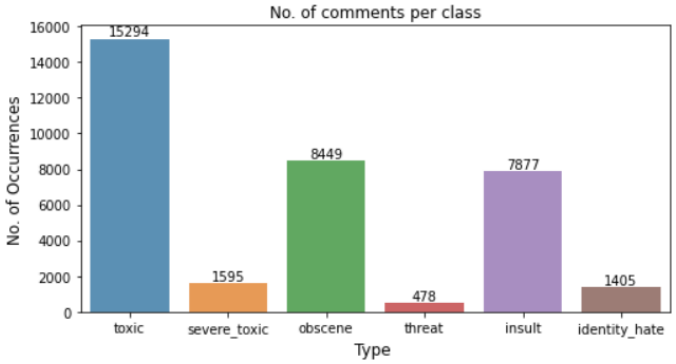


Fig. 1. Bar Chart Showing Number of Comments Per Class.

Fig. 1 shows that there are 15,294 toxic messages, 1595 severe_toxic messages 8449 obscene messages 478 threat messages, 7877 insult messages and 1405 identity_hate messages.

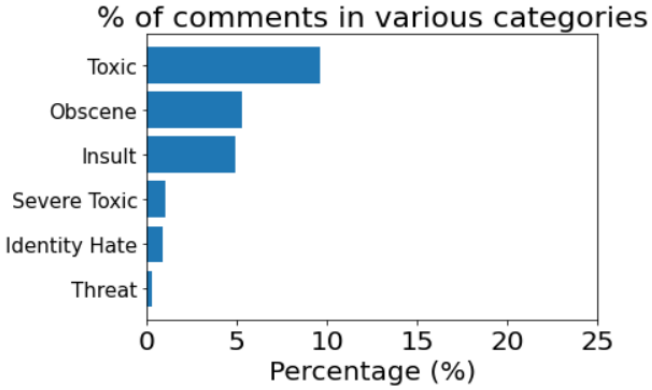


Fig. 2. Percentage of Comments in Various Categories.

Fig. 2 represents the percentage of comments that belong to a particular hate label. As we can see that only about 30% of comments fall under one of the given labels and a certain set of preprocessing has to be done in order to avoid class imbalance problems, see Fig. 3.

Fig. 3. Word Cloud.

C. Data Preprocessing

Data preprocessing is a technique that involves transformation of raw data into a format that can be utilized by machine learning models. The comments present in the dataset contain characters that are not essential in determining if a message is related to cyberbullying, hence such characters must be removed to increase the efficiency of the machine learning models. Various preprocessing steps have been carried out to clean the raw data. They include removal of numbers in combination with letters, replacing punctuations with white spaces, conversion of the comments to lower case, removal of escape sequences and replacing it with white spaces, removal of non ascii characters, removal of stop words, etc.

Since the data has lot of non-toxic comments and less toxic comments, training the model can make it prone to identifying more non-toxic comments rather than toxic comments. Hence, balancing of non-toxic and toxic comments has been done. After the basic preprocessing of the data, the dataset was separated into six sections. Each section contains the comments along with one of the categories in the dataset. Comments that were toxic or belonged to other forms of toxicity made up for less than 10% of the comments in the dataset. This led to a class imbalance problem where the class distribution was unequal or skewed. To overcome this issue, we took a subset of the data where the proportion of comments related to toxicity was at least 20% in relation to non-toxic comments. In order to make the subsets first the number of comments under each category was counted. If the category had 5000 rows, another 5000 rows of comments not belonging to the category were considered and the two were combined to form a new section. In case, a category did not have 5000 rows, then 20% of the comments belonging to the category and 80% of the comments not belonging to the category were chosen for the combination. In this manner six different sections or data frames were made per category.

Fig. 4 gives an overview of the data frames after eliminating the class imbalance problem. After the dataset was balanced, it was vectorized using Tfidf vectorizer which tokenizes the documents, learns the vocabulary given in the dataset and calculates the document frequency weighting which is essential in building a machine learning model. As machine learning models do not understand words or letters, they are converted to a machine understandable format using this method. 1-gram is used to split and tokenize the data, where it breaks the sentences into individual words and calculates the document frequency weight using Tfidf vectorizer.

	DF name	No. of pts (1)	No. of pts (0)	Total data pts
Toxic	data_tox_done	5000	5000	10000
Severe Toxic	data_sev_done	1595	1595	3190
Obscene (8449)	data_obs_done	5000	5000	10000
Threat (478)	data_thr_done	478	1912	2390
Insult (7877)	data_ins_done	5000	5000	10000
Identity Hate (1405)	data_ide_done	1405	5620	7025

Fig. 4. Overview of the Dataframes.

D. Backend

The backend was built using Flask, the database used was Firebase and Telegram API. SMTPLib (Simple Mail Transfer Protocol) module, which defines an smtp client session object is used to send emails.

When the server is run, the vectorizer and Random Forest model for each category is unpickling and loaded into the memory. We have used the Telegram API to capture messages sent by users. These messages are first vectorized and then passed to the machine learning model. These models return their respective names and the associated probabilities. The returned probabilities are compared with a threshold to identify bully messages. If the probability exceeds the threshold of 0.7, the model with the highest probability is chosen as the category that identifies the type of bullying. A warning is sent for the first bully message in the conversation, if more of such messages are sent further actions are taken.

A message queue is maintained to hold a few messages before and after the second bully message in the conversation. This queue is used to store the context of the messages being sent. The details of the messages and users such as username, chat id, time stamp, etc., are also stored in the message queue. Once the message queue is filled, the queue is pushed to the database along with the type of bullying and its probability. This event also triggers an email to the authority which alerts them about the update of the database so that they can review the conversation. This service is hosted on Heroku where the service is kept running so that messages can be monitored in real time.

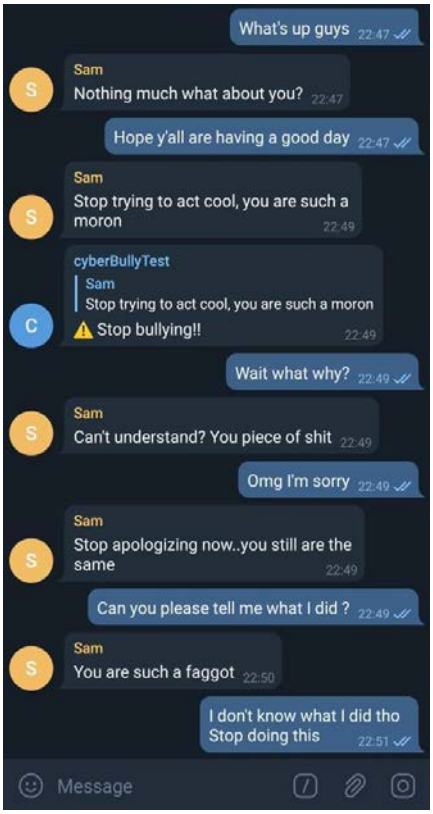


Fig. 5. Snapshot of a Cyberbullying Conversation.

Fig. 5 is a snapshot of a conversation that involves bullying. A warning is sent by the bot as a reply to the first bully message. If the user responsible for sending the message continues to do so by ignoring the warning, the messages are captured as evidence along with their details and these are saved to the database.

E. Frontend

The email that is sent to the authority notifies the authority when the database has been updated and it consists of the dashboard URL. On clicking the URL mentioned in the email for dashboard, the user lands on the login page where the person has to login using the official cyber cell credentials. Firebase authentication was used for this purpose.

Upon successful login, the authority is redirected to the pending page as shown in Fig. 6 which consists of all the conversations along with details like the type of toxicity, level of toxicity, name and user Id of the bully. The authority is given an option to either delete the conversations or approve them if they are legitimate. Approved messages are displayed under the approved tab.

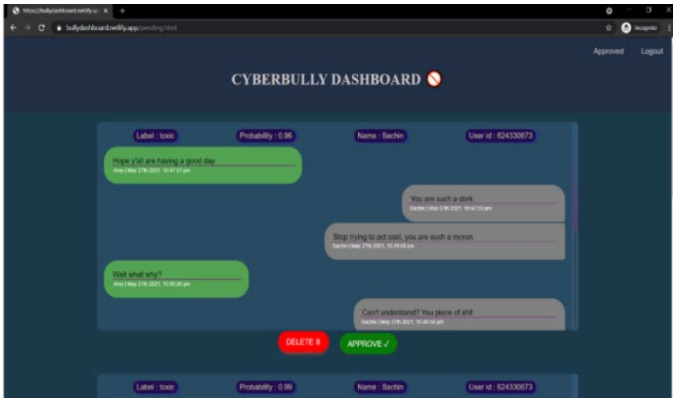


Fig. 6. Pending Reports Displayed in the Dashboard.

IV. Result and Discussion

The evaluation metric chosen was F1 score. F1 score was selected since it takes into consideration both False Positive and False Negative values, our goal was to minimise these values. A table comparing the F1 scores obtained using each of the machine learning models for the six data frames is shown below in Table I.

TABLE I. Comparison of F1 Score for the Models Used

Measure	Logistic Regression	KNN	SVM	Random Forest
F1 Score (Toxic)	0.861234	0.185120	0.876133	0.838055
F1 Score (severe_toxic)	0.927879	0.857416	0.926004	0.934874
F1 Score (obscene)	0.908655	0.519056	0.921378	0.909091
F1 Score (insult)	0.896599	0.257992	0.902619	0.883993
F1 Score (threat)	0.628821	0.720000	0.786765	0.795539
F1 Score (identity_hate)	0.699029	0.230159	0.797516	0.768448

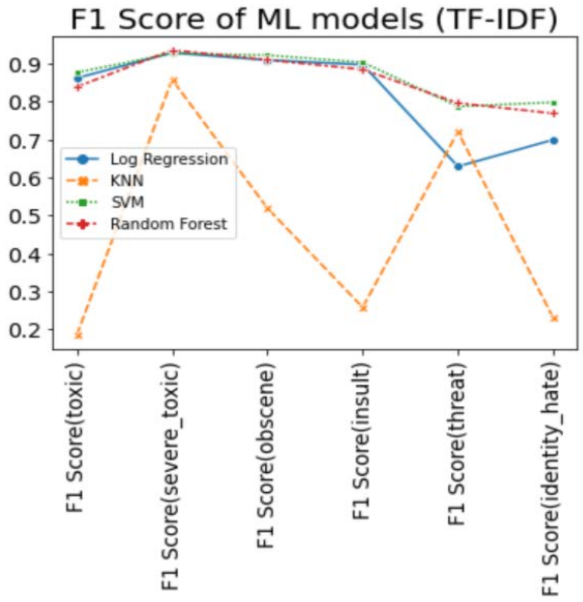


Fig. 7. F1 Score Comparison Plot.

By observing the values in Fig. 7, we see that SVM and Random Forest perform comparatively better than Logistic Regression and KNN classifiers. The plot in Fig. 7 shows that Random Forest represented by the red line and SVM represented by the green line perform better than the rest. So, Random Forest is chosen as the best performing model since results are returned in probabilities which were useful for our cyberbully detection system. Random Forest also makes use of multiple decision trees and hence gives a more accurate result.

V. Conclusion and Future Work

There are many people who are falling prey to cyberbullying, which goes unnoticed. It is high time that a system is made which helps in preventing these crimes. Our cyberbullying detection system aims to bring all these cases of cyberbullying under control by detecting and warning the bully. Then these cases are also reported to appropriate local authorities, which can be verified and required steps and actions can be taken depending upon the situation. Our model is built by using Chat application APIs, Firebase, Flask, Scikit learn, HTML, CSS and Javascript. It also gives high accuracy compared to the existing projects. It can be feasibly used and would show appropriate information whenever one is cyberbullied.

In future, CNN will be applied. Also, the algorithm will be applied on huge datasets and accuracy can be improved further. Certain issues like reducing false alarms, educating the users of the usability feature of cyberbullying detection and reporting to authorities, framing of privacy features of the platform and having moderators to review the conversations will be addressed in future work.

References

[1] Batoul Haidar, Maroun Chamoun & Ahmed Serhrouchni. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. *In : Proceedings of Advances in Science, Technology and Engineering Systems Journal.*

[2] Cynthia Van Hee, Jacobs G, Emmery C, Desmet B, Lefever E & Verhoeven B. (2018). Automatic detection of Cyberbullying in social media text. *In: Proceedings of PLOS One.*

[3] Gurpinder Singh , Vijay Dhir & Vijay Rana. Design of Semantic Framework for Detecting Cyberbullying on Social Media. *In: Proceedings of International Journal of Scientific Research and Reviews.*

[4] Haoti Zhong, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller & Cornelia Caragea. (2016). Content-Driven Detection of Cyberbullying on the Instagram Social Network. *In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence.*

[5] Hosseinmardi, Sabrina Arredondo, Rahat Ibn Rafiq, Richard Han, Qin Lv & Shivakant Mishra. (2015). Detection of Cyberbullying on Instagram Social Network. *In: Proceedings of Association for the Advancement of Artificial Intelligence.*

[6] J.I. Sheeba, S. Pradeep Devaneyan & Revathy Cadiravane. (2019). Identification and Classification of Cyberbully Incidents using Bystander Intervention Model. *In: Proceedings of International Journal of Recent Technology and Engineering (IJRTE).*

[7] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer & Ammar Mohammed. (2019). Social Media Cyberbullying Detection using Machine Learning. *In: Proceedings of (IJACSA) International Journal of Advanced Computer Science and Applications.*

[8] Kshitiz Sahay, Harsimran Singh Khaira, Prince Kukreja & Nishchay Shukla. (2018). Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning. *In: Proceedings of* International Journal of Engineering Technology Science and Research.

[9] Laura P. Del Bosque & Sara Elena Garza. (2014). Aggressive Text Detection for Cyberbullying. In: Proceedings of Springer International Publishing, Switzerland.

[10] Liew Choong Hon & Kasturi Dewi Varathan. (2015). Cyberbullying detection system on Twitter. In: Proceedings of International Journal of Information Systems and Engineering, 1:1.

[11] Monirah Abdullah Al-Ajlan and Mourad Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection" International Journal of Advanced Computer Science and Applications(IJACSA), 9(9), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090927.

[12] Ximena M. Cuzcano and Victor H. Ayma, "A Comparison of Classification Models to Detect Cyberbullying in the Peruvian Spanish Language on Twitter" International Journal of Advanced Computer Science and Applications(IJACSA), 11(10), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111018.

[13] Rolfy Nixon Montufar Mercado, Hernan Faustino Chacca Chuctaya and Eveling Gloria Castro Gutierrez, "Automatic Cyberbullying Detection in Spanish-language Social Networks using Sentiment Analysis Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 9(7), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090733.

[14] Diego A. Andrade-Segarra and Gabriel A. Le´on-Paredes, "Deep Learning-based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks" International Journal of Advanced Computer Science and Applications(IJACSA), 12(5), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120592

Customers' Opinions on Mobile Telecommunication Services in Malaysia using Sentiment Analysis

Muhammad Radzi Abdul Rahim[1], Yuzi Mahmud[3]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Selangor, Malaysia

Shuzlina Abdul-Rahman[2]
Research Initiative Group of Intelligent Systems
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Selangor, Malaysia

***Abstract*—Mobile telecommunication companies in Malaysia have been widely used in the recent decade. There is intense competition among them to keep and gain new customers by offering various services. The reviews of the services by the customers are commonly shared on social media such as Twitter. Those reviews are essential for mobile telecommunication companies to improve their services and at the same time to keep their customers from churning to another company. Hence, this study focuses on the public sentiment on Twitter towards mobile telecommunication services in Malaysia. Data on Twitter was scraped using three keywords: Celcom, Digi, and Maxis. The keywords used to refer to Malaysia's top three mobile telecommunication companies. The timeline for the tweets was between December 2020 until January 2021 and was based on the promotion sales commonly used by the organisation to boost their sales which is called Year End Sales. Corpus-based approach and Machine Learning model using RapidMiner were used in this study, namely, Support Vector Machine (SVM), Naïve Bayes, and Deep Learning. The corpus determines the sentiment from the tweets, either positive, negative, or neutral. The models' performances were compared in terms of accuracy, and the outcome shows that Deep Learning classifiers have the highest performance compared to other classifiers. The results of this sentiment analysis are visualised for easy understanding.**

Keywords*—Sentiment analysis; predictive analytics; RapidMiner; mobile telecommunications

I. Introduction

Sentiment analysis determines a writer's attitude toward a topic or the overall contextual polarity. The author's attitude could be based on his or her judgment or evaluation, affective state (i.e., the author's emotional state at the time of writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader) [1]. In other words, sentiment analysis is the classification of text documents, such as user reviews, newsgroup postings, and blogs, based on the polarity of their opinions.

Twitter, Facebook, and Instagram are examples of online social media that allow users to communicate with people worldwide. People can spread data, opinions, statements, and behaviour via social media [2]. They express their thoughts about items or share personal experiences, and they even can influence politics and businesses [3]. For example, practically every large corporation has a Twitter account to keep track of client comments on their services or products. Twitter is a powerful microblogging platform that allows users to post status updates (called "tweets"). These tweets contain a lot of human expressions, such as likes, dislikes, and contributions to many issues [3].

Malaysia's mobile telecommunications services sector has exploded in the recent decade. The competition among existing and new telecommunication service providers has intensified as they aim to keep and gain new consumers by offering a variety of eye-catching promotions and seasonal events. Customers frequently compare the promotions on Twitter. Opinions expressed on Twitter are frequently more influential than other social media because they are made public for all to see and compare. Hence, the motivation of this paper is to develop a model that could predict which telecommunications companies in Malaysia provide the top mobile services for the users.

The following are the contributions of this paper:

- This research proposes a Sentiment Analysis on customers' opinions for mobile telecommunication services in Malaysia.
- A corpus-based approach and several state-of-the-art machine learning models are compared to get the best model.
- The findings are visualised using Microsoft Power BI to understand the results better.

The remaining of this paper is structured as follows: Section II discusses the related works on mobile communications and sentiment analysis, while Section III describes the study's methodology. Section IV highlights the results, and Section V concludes the paper with future works.

II. Literature Review

A. Mobile Telecommunication Services in Malaysia

Mobile telecommunications are the process of sending, transmitting, and receiving information over a distance to communicate [4]. This type of signal transmission is carried out using a mobile device, such as a cell phone, computer, or other wired or wireless devices. Mobile telecommunications have a generation of network standards which are 1G, 2G, 3G, 4G, and 5G. 1G and 2G only give a means of speaking and texting over mobile phones. 3G allows mobile phone users to

connect to the internet and send or receive any multimedia transmissions. 4G increases the bandwidth significantly while 5G is still under development to improve 5G [5].

A study by Dagli and Jenkins [5] stated that mobile phone services are one of the most prominent growing areas in the telecommunication industry while currently holding more than 1.7 billion customers worldwide and targeting around 80% of the world population as its potential customers. Five (5) big companies dominate the telecommunication industry: Maxis, Celcom, Digi, U Mobile, and Unifi. However, there are three major mobile service providers in Malaysia, namely Maxis, Celcom, and Digi [6]. Smaller carriers with limited coverage in Malaysia and mobile virtual network operators are the rest (MVNOs). Each of these significant corporations has its own set of colours that signify its respective brands. Maxis represents green (post-paid), red (prepaid), blue represents Celcom, and yellow represents Digi.

B. Sentiment Analysis

Sentiment analysis (SA), also known as opinion mining or contextual mining, is a technique used in Natural Language Processing (NLP), computational linguistics, and text analysis to discover, systematically extract, and quantify subjective data. According to Singh et al. [7], the most common approach is machine learning, a method that needs an essential data set for training and learning the aspects and sentiments associated. Sentiment analysis can be used on any material or object in the form of a customer's voice, such as reviews or responses. For example, if a consumer wants to buy something online, they would usually read reviews on the item or product first, which will help them make the best selection possible [8]. Sentiment analysis allows businesses to understand a user's feelings about a product, service, or brand by converting internet comments into emotion and categorising it as positive, negative, or neutral [9].

Machine learning and lexicon-based approaches are the two most utilised techniques in sentiment analysis [10]. The lexicon-based technique generates flawless dictionaries, whereas the machine learning technique concentrates on feature vectors. The general workflow of sentiment analysis consists of a few processes: goal setting, text preprocessing, parsing the content, text refinement, and analysis and scoring [11].

C. Machine Learning Algorithms

1) Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression, and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. SVM is a supervised machine learning model for two-group classification issues that uses classification techniques. SVM models can categorise new text after being given labelled training data sets for each category. It is mainly used in text classification, and it comes with a dataset for high-dimensional training [3].

2) Naïve bayes: Naïve Bayes is the supervised machine learning algorithm that uses Bayes theorem for classification problems [11]. The Bayes' Theorem is used to create a collection of classification algorithms known as Naive Bayes classifiers. In Almonajed and Jukić's paper [3], the researchers mentioned that the Naïve Bayes classifier is a popular supervised classifier, furnishes a way to express positive, negative, and neutral feelings in the web text. Naïve Bayes classifier is valuable and efficient for classification purposes [12]. A family of algorithms share a similar idea: each pair of features being classified is independent of the others. It is a numerical-based approach with easy, fast, and high accuracy [13].

3) Deep learning: Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks [6]. Deep learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results. Along with the success of deep learning in many application domains, deep learning is also used in sentiment analysis in recent years [6]. Deep Learning is based on a multi-layer feed-forward artificial neural network taught via back-propagation and stochastic gradient descent.

III. Research Methodology

Machine learning and lexicon-based approaches are the two most utilised techniques in sentiment analysis [11]. The lexicon-based technique generates flawless dictionaries, whereas the machine learning technique concentrates on feature vectors. The general workflow of sentiment analysis consists of a few processes: goal setting, text preprocessing, parsing the content, text refinement, and analysis and scoring [14].

A. Data Collection

Data was collected by extracting tweets from Twitter using a python tool called Twint. Python is a general-purpose language. It can build anything using Python [15]. Twint is a Python-based Twitter scraping tool that allows scraping tweets from Twitter profiles without using Twitter's API. Twint makes use of Twitter's search operators to scrape tweets from specific people, scrape tweets related to specific themes, hashtags, and trends, and shift out sensitive information from tweets such as e-mail addresses and phone numbers. Twint also uses Twitter to make unique queries that allow it to scrape a Twitter user's followers, tweets they have liked, and whom they follow without requiring any login, API, Selenium, or browser emulation. Tweets scraped were between 1st December 2020 until 31st January 2021. Three keywords are used to scrape the tweets: Celcom, Digi, and Maxis. Any tweets posted on Twitter that contained the keywords and within the durations mentioned above will be automatically scrapped by Twint.

B. Data Preprocessing

1) Data cleaning: There were seven operators used to clean the datasets. The first operator is called the Retrieve operator. After the datasets had been successfully scraped using Python tools, the datasets were stored in a CSV file. Then the CSV was stored in the RapidMiner repository. The

Retrieve operator is used to access the stored datasets in the repository and load them into the RapidMiner's process. Then, the second operator is called Filter Examples. The raw datasets collected consist of multiple languages. Since RapidMiner only detects English, only English tweets were chosen. Python tools used in the data collection phase provided a language detector of the tweets. The tweets were labelled by each detected language in the language attribute. Therefore, the Filter Examples operator selected only data labelled as "en", which means English. After only tweets in English were chosen, the following operator used was called the Select Attributes operator. Only four important attributes from 36 common attributes were selected. The selected attributes were time, date, tweet, and username attributes. Then, the operator will keep only the four attributes selected and remove the remaining 32 attributes.

The next operator used was called Replace operator. The value of tweet attributes might contain special characters or punctuation characters. Those characters cannot be processed to perform a Sentiment Analysis. So, those characters need to be removed. Remove Duplicate operator was the fourth operator used in the data preprocessing process. It might be a spamming situation where the customers spam with the same tweets in the real world. Therefore, the operator will remove duplicate data and only keep one. The last operator used in the data preprocessing process was the Filter Examples operator. The operator was used again but with a different purpose. After the datasets go through several processes, the missing values appear. The values were removed due to some factors, including those from a special character or just a punctuation character. Therefore, the operator will remove data with a missing value. Fig. 1 shows a RapidMiner's operator used in the data cleaning process.

2) Data labelling: Fig. 2 shows a RapidMiner's operators used to extract sentiment. The data labelling process is the process to extract the sentiment and classify the texts as either positive, negative, or neutral. The cleaned datasets have been called by using the Retrieve operator. Then, the Extract Sentiment operator was used to extract the tweet's sentiment. A few models can be used to extract the sentiment, and the VADER model has been used in this project. VADER model will produce a compound score for each sentence. The compound score is a metric that calculates the sum of all the lexicon ratings, which have been normalised between -1(most extreme negative) and +1 (most extreme positive). Table I shows a compound for each sentiment class.

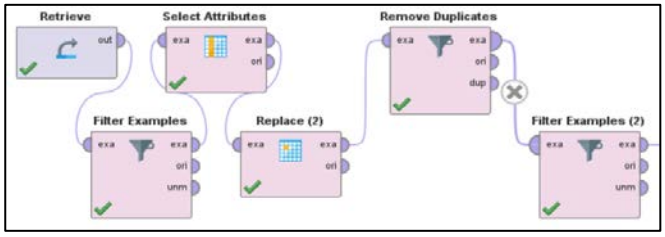


Fig. 1. RapidMiner's Operators used for Data Cleaning.

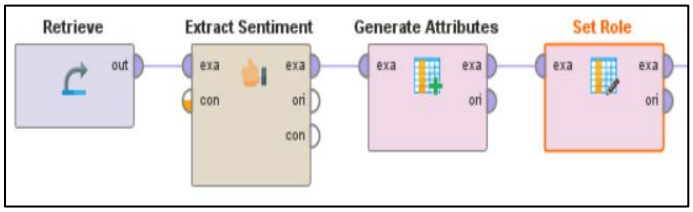


Fig. 2. RapidMiner's Operators use to Extract Sentiments.

TABLE I. COMPOUND SCORE

Sentiment	***Compound Score***
Positive	>= 0.05
Neutral	> -0.05 and < 0.05
Negative	<= -0.05

After three classes of sentiment have been successfully extracted, the following process sets the role for the sentiment attribute. RapidMiner's operator used was called the Set Role operator. The role of the attribute that has been set will describe how other operators handle the attribute. This operator also will transform the sentiment attribute into a special attribute.

3) Data transformation: Fig. 3 shows the operators applied in the Process Document operator. Only attributes that focused on sentiment analysis were involved in this stage, which is tweet attributes. There are five processes involved in the data transformation stage. The first process was Transform Cases. Transform Cases operator has been applied to transform all characters in the tweet attribute into lower case. The second process was stopped word removal. All stop words, such as frequent terms like a and the, are deleted from multiple word queries to improve search performance. The third process was the tokenisation process, where the Tokenize operator was applied. For this project, non-letter characters have been used as a method for the splitting points. The fourth process was the stemming process, where the Stem operator has been applied. This operator stems from English words using the Porter stemming algorithm, which employs an iterative, rule-based substitution of word suffixes intending to shorten the words to a minimum length. The last data transformation process was generating n-grams by using the Generate n-Grams operator. This operator generates the word n-Grams, which refers to all series of consecutive tokens of length n.

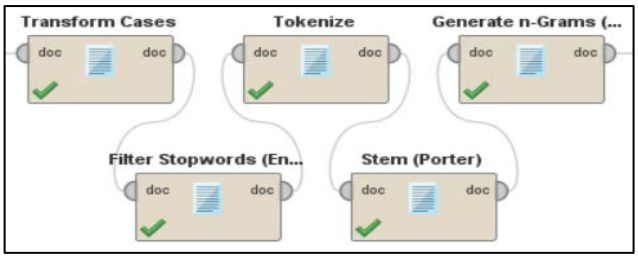


Fig. 3. Operators Applied in Process Document Operator.

C. Model Development

In this project, three machine learning models were compared. The classifiers used were Support Vector Machine (SVM), Naïve Bayes, and Deep Learning. Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input [16-17]. Those Machine Learning models will be applied using RapidMiner. After the datasets had been successfully labelled, the result showed an imbalanced class. Datasets are said to suffer the Class Imbalance Problem when highly imbalanced class distributions. The datasets must be equally distributed for each class to get higher accuracy. There are a few methods that can be used to handle imbalanced data. This study used the under-sampling method to obtain an evenly dispersed amount of data to a model. This method investigates the class with the smallest number. Fig. 4 shows RapidMiner's operators used in the modelling process.

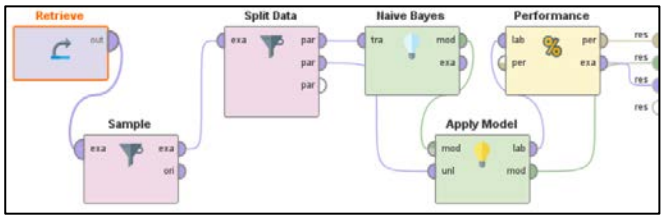


Fig. 4. RapidMiner's Operators used in the Modelling Process.

The first operator used in the modelling process was the Retrieve operator. This operator is used to call the datasets saved in the RapidMiner repository. The second operator used was the Sample operator. This operator was used to solve the imbalanced class problem. The size of the sample has been specified with an absolute sample. Then, the size per class has been set by the smallest number of sizes per class for each dataset. The next operator applied was the Split Data operator. The operator is used to split the datasets into two parts: train data and testing data. The split ratio used in this project was 70:30, 80:20, and 90:10 percent. The three ratios were used to find the best ratio which produced the highest accuracy. After the Split Data operator splits the datasets into training and testing data, the training data will be passed to the classifier operator. In contrast, the testing data will be passed to the Apply Model operator. Since this project will compare three different Machine Learning models, three different classifiers will be used.

After the training data has been successfully processed in the classifier operator, the training data also will be passed to the Apply Model operator like the testing data. Therefore, the Apply Model operator will receive both parts of data: training and testing data. Apply Model operators have two types of input received: Model and Unlabeled data. Model input will be received data that model has been trained, while Unlabeled data input will be received testing data. This operator also will apply a model used on the datasets. Not only receive two types of inputs, but this operator will also produce two types of output: Labelled data and model. The input of the Model type was passed without changing to the output through the output of the model type. While the datasets delivered from the output of the labelled type will be passed to the Performance operator. The last operator applied in the modelling process was a Performance operator. This operator is used to evaluate the performance of classification tasks statistically. This operator will return a list of classification task performance criteria values.

D. Dashboard Development

The goal for developing the dashboard in this project was to visualise customers' reviews and comparisons between each telecommunication company's customers based on Twitter data. This section goes through the visualisation that may be derived from the sentiment data that has been extracted. The subject under discussion is customers' reviews on Malaysia's top three telecommunication services, and the dashboard was developed using Microsoft Power BI.

IV. Result and Discussion

A. Split Ratio Selection

There are three split ratios used in this project which is 70:30, 80:20, and 90:10 percent ratio. Each split ratio produces a different accuracy, and the split ratio which produces the highest accuracy was selected. Table II until Table IV presents the accuracy result by different split ratios for each dataset. Table II shows the accuracy result for the Celcom dataset. As shown in Tables II to IV, generally, the performance of the three classifiers increases as the sample size increases. In Table II, the Deep Learning classifier appears to be the best model, followed by the Naive Bayes and Support Vector Machines classifier. Similar performance can be seen in Digi and Maxis datasets, in which Deep Learning consistently gave the highest accuracy results.

Based on the accuracies in Table II until Table IV, the 90:10 percent split ratio produced the highest accuracy for each classifier and each dataset. Therefore, the 90:10 split ratio will be used for the following process: finding the best N-gram for the better accuracies produced.

TABLE II. Accuracy Result for Celcom Dataset

Classifier	**70:30**	**80:20**	**90:10**
SVM	49.87%	49.94%	50.93%
Naive Bayes	60.84%	61.52%	61.59%
Deep Learning	79.17%	77.81%	81.22%

TABLE III. Accuracy Result for Digi Dataset

Classifier	**70:30**	**80:20**	**90:10**
SVM	37.86%	38.31%	38.72%
Naive Bayes	52.93%	51.31%	55.49%
Deep Learning	65.03%	69.73%	83.73%

TABLE IV. Accuracy Result for Maxis Dataset

Classifier	**70:30**	**80:20**	**90:10**
SVM	44.84%	44.21%	45.86%
Naive Bayes	52.84%	52.60%	53.19%
Deep Learning	76.89%	76.91%	78.25%

B. Model Comparison

1) Support Vector Machine (SVM): The first classifier tested was the Support Vector Machine. Table V shows the accuracy for the SVM model with three conditions: without n-Grams, n-Grams set to 2, and n-Grams set to 4.

TABLE V. ACCURACY RESULT FOR SVM MODEL

Dataset	Accuracy		
	Without n-Grams	**n-Grams = 2**	**n-Grams = 4**
Celcom	50.93%	52.28%	45.18%
Digi	38.72%	49.44%	53.98%
Maxis	45.86%	55.40%	54.99%

For the conclusion, after the three different datasets have been tested using the SVM classifier, by applying the change of parameter with n-Grams equal to 2 has been set, the accuracy for the classifier for Celcom and Maxis datasets were the highest. While for the Digi dataset, by applying the change of parameter with n-Grams equal to 4 has been set, the accuracy for the classifier was the highest.

2) Naïve bayes: The second classifier that has been tested was Naïve Bayes. Table VI shows the accuracy for the SVM model with three conditions: without n-Grams, n-Grams set to 2, and n-Grams set to 4.

TABLE VI. ACCURACY RESULT FOR NAÏVE BAYES MODEL

Dataset	Accuracy		
	Without n-Grams	**n-Grams = 2**	**n-Grams = 4**
Celcom	61.59%	62.44%	63.62%
Digi	55.49%	56.23%	43.49%
Maxis	53.19%	56.58%	48.73%

In conclusion, after the three different datasets have been tested using the Naïve Bayes classifier by applying the change of parameter with n-Grams equal to 2 has been set, the accuracy for the classifier for Digi and Maxis datasets will be the highest. While for the Celcom dataset, n-Grams equals 4 produced the highest accuracy.

3) Deep learning: The final classifier that has been tested was Deep Learning. Table VII shows the accuracy for the SVM model with three conditions: without n-Grams, n-Grams set to 2, and n-Grams set to 4.

TABLE VII. ACCURACY RESULT FOR DEEP LEARNING MODEL

Dataset	Accuracy		
	Without n-Grams	**n-Grams = 2**	**n-Grams = 4**
Celcom	81.22%	78.00%	66.33%
Digi	83.73%	81.22%	41.67%
Maxis	78.25%	74.39%	59.33%

For the conclusion, after the three different datasets have been tested using Deep Learning classifiers, applying the change of parameter with n-Grams equal to 2 and n-Grams equal to 4 has been set, the accuracy for the classifier will be decreased. While without the n-Grams, the classifiers will perform better. It was different from the SVM and Naïve Bayes classifier.

C. Dashboard Visualisation

The dashboard of customers' reviews on Malaysia's top three telecommunication services was developed in 2 pages. Fig. 5 shows the first page of the dashboard. This page was aimed to give a clear view to the readers about the comparison between the top three telecommunication services in Malaysia based on customers' reviews on Twitter. The results are related to customer satisfaction. When the customers feel satisfied, they believe in the brand and become loyal [18]. Based on the visualisation on page 1, it can be concluded that Digi telco has the highest number of positive reviews, while the second-highest was Celcom telco. From that, the readers can decide where the readers might be tended to choose a telco with the highest positive reviews. The readers also were able to read the reviews given by previous users in the table visualisation.

Fig. 5. Page 1 of the Dashboard.

Fig. 6 shows a second page of the dashboard. This page was aimed to provide profound information about the reviews. The readers from the background as a telco's customers might not be interested in this page because most of the telco's customers just want to know the information that can help them decide. At the same time, this page was more focused on telco companies' sites. Based on the line chart on the second page, the telco companies can identify a specific time the reviews were given in the duration selected. From that information, the telco companies might gain new knowledge about the customer's behaviour or what has been suffered by the customers in that time so that the telco companies can decide to maintain customers satisfaction.

Fig. 6. Page 2 of the Dashboard.

V. Conclusion

This study has demonstrated SVM, Naive Bayes, and Deep Learning using a corpus-based approach and machine learning algorithms. Our results showed that the Deep Learning model without n-Grams provides the highest accuracy. The highest value produced for the Celcom dataset is 81.22% which is 17.6% better than the Naïve Bayes model and 28.94% better than the SVM model. Meanwhile, for the Digi dataset, the Deep Learning model produced 83.73% accuracy, 27.5% better than the Naïve Bayes model and 29.75% better than the SVM model. Lastly, for the Maxis dataset, the Deep Learning model produced 78.25% accuracy, which is 21.68% better than the Naïve Bayes model and 22.85% better than the SVM model. The future researcher may consider using a more extensive dataset for better generalisation. This work would reach a wider audience with a mobile application since smartphones are widely used nowadays. Further future work could also include the Malay language during sentiment extraction since Malay is the highest population in Malaysia.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, for the support throughout this research.

References

[1] N. Sakinah Shaeeali, A. Mohamed, and S. Mutalib, “Customer reviews analytics on food delivery services in Social Media: A Review,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 9, no. 4, p. 691, 2020.

[2] M. Bakri C. Haron, S. Z. Z. Abidin, N. Azmina M. Zamani "Visualisation of crime news sentiment in Facebook," International Journal of Engineering & Technology, vol. 7, no. 4.38, p. 955, 2018.

[3] Almonajed, O. and Jukić, S., 2021. "Sentiment Analysis on Twitter Data using Big Data." Journal of Engineering and Natural Sciences, 3(1).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–7319.[4] KDnuggets. 2021. "5 Things You Need to Know about Sentiment Analysis and Classification" - KDnuggets. [online] Available at: <https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html> [Accessed 27 August 2021].K. Elissa, “Title of paper if known,” unpublished.

[4] Dagli, O. and Jenkins, G., 2016. "Consumer preferences for improvements in mobile telecommunication services." Telematics and Informatics, 33(1), pp.205-216. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] Marcopolis.net. 2021. "Top Telecoms in Malaysia | Malaysia's Largest Telecoms." [online] Available at: <https://marcopolis.net/top-telecoms-in-malaysia-malaysia-s-largest-telecoms.htm> [Accessed 27 August 2021].

[6] J. Singh, G. Singh, and R. Singh, "Optimisation of sentiment analysis using machine learning classifiers," Human-centric Computing and Information Sciences, 11-Dec-2017. [Online]. Available: https://hcis-journal.springeropen.com/articles/10.1186/s13673-017-0116-3. [Accessed: 27-Aug-2021].

[7] Etter, M., Colleoni, E., Illia, L., Meggiorin, K. and D’Eugenio, A., 2016. "Measuring Organisational Legitimacy in Social Media: Assessing Citizens' Judgments With Sentiment Analysis." Business & Society, 57(1), pp.60-97.

[8] Medium. 2021. "A Beginner’s Guide to Sentiment Analysis." [online] Available at: <https://medium.com/@mattkiser/a-beginner-s-guide-to-sentiment-analysis-888390a8085a> [Accessed 27 August 2021].

[9] A., V. and Sonawane, S., 2016. "Sentiment Analysis of Twitter Data: A Survey of Techniques." International Journal of Computer Applications, 139(11), pp.5-15.

[10] Godsay, M., 2015. "The Process of Sentiment Analysis: A Study. International Journal of Computer Applications," 126(7), pp.26-30.

[11] N. T. Hazmiza, “Classifying violent elements in role-playing games based on user review using naïve Bayes technique,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1.3, pp. 402–407, 2020.

[12] N. Seman and N. Atiqah Razmi, “Machine learning-based technique for big data sentiments extraction,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 9, no. 3, p. 473, 2020.

[13] Singh, S., 2020. "Twitter Sentiments Analysis Using Machine Learning. International Journal of Scientific Research in Computer Science," Engineering and Information Technology, pp.312-320.

[14] N. S. Mohd Shafiee and S. Mutalib, “Prediction of mental health problems among higher education student using machine learning,” International Journal of Education and Management Engineering, vol. 10, no. 6, pp. 1–9, 2020.

[15] J. Brownlee, “What is deep learning?,” Machine Learning Mastery, 14-Aug-2020. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/. [Accessed: 27-Aug-2021].

[16] Meltwater, E., 2021. "Deep Learning Models for Sentiment Analysis." [online] Underthehood.meltwater.com. Available at: <https://underthehood.meltwater.com/blog/2019/08/22/deep-learning-models-for-sentiment-analysis/> [Accessed 17 October 2021].

[17] Muhammad, I., Farid Shamsudin, M. and Hadi, N., 2016."How Important Is Customer Satisfaction? Quantitative Evidence from Mobile Telecommunication Market." International Journal of Business and Management, 11(6), p.57.

Detecting Server-Side Request Forgery (SSRF) Attack by using Deep Learning Techniques

Khadejah Al-talak[1]
Department of Information Security
University of Tabuk, KSA. Tabuk, KSA
KSA. Tabuk, KSA

Dr. Onytra Abbass[2]
Department of Information Technology
University of Tabuk
Tabuk, KSA

***Abstract*—Server-side request forgery (SSRF) is a security vulnerability that arises from a vulnerability in web applications. For example, when the services are accessed via URL the attacker supply or modify a URL to access services on servers that he is not permitted to use. In this research, various types of SSRF attacks are discussed, and how to secure web applications are explained. Various techniques have been used to detect and mitigate these attacks, most of which are concerned with the use of machine learning techniques. The main focus of this research was the application of deep learning techniques (LSTM networks) to create an intelligent model capable of detecting these attacks. The generated deep learning model achieved an accuracy rate of 0.969, which indicates the strength of the model and its ability to detect SSRF attacks.**

Keywords*—Server-side request forgery (SSRF); machine learning (ML); deep learning (DL); long short-term memory (LSTM)

I. Introduction

With the development and increasing number of applications on the World Wide Web, security violations have increased, as these applications are characterized by being public, making them vulnerable to attacks [1]. Despite the great progress in methods to protect web applications, hackers are searching using advanced technology for loopholes to overcome these protection methods [2]. These applications inherently contain huge and sensitive data that must be protected from intrusion. SSRF attacks exploit any vulnerability within the web application to enter the server and obtain data in illegal ways, so we need mechanisms to defend against these attacks. Web applications use Authentication Systems to confirm the identity of the client communicating with the web application on the server. The username and password are sent from the client to the server in an encrypted form via HTTP, this information can be compromised while it is being sent. One of the main characteristics of the web applications is that it should work without interruption, so when designing these applications, you must take into account that it works even if it is attacked by hackers. Therefore, several security rules and controls must be put in place to protect web applications from attacks that are widespread with the development of technology. The development of technology helps hackers search for potential vulnerabilities in web applications, making the web application vulnerable with the data contained within it being corrupted, lost, or hijacked. There are now basic plans and concepts to reduce the risk of these attacks and protect data [3]. Web applications are exposed to multiple attacks, the most common of these attacks is Cross-Site Scripting (XSS), SQL Injection, DDoS Attack, Malware, Bots, Cross-Site Request Forgery (CSRF) and Server-Side Request Forgery (SSRF). To detect and mitigate attacks against web applications, machine learning techniques are used because of their ability to learn from data. Server-side request forgery (SSRF) is a security vulnerability that arises from a vulnerability in web applications**.**

In this research, different methods are presented in several literatures to detect SSRF attacks that use URLs to perform the attack. in addition to, the LSTM deep learning network was used to build an intelligent model for detecting SSRF attacks. The training of this model was based on a set of data that represents normal data and others infected with attacks. The test results of this model were good, with an accuracy of 96%.

A. Background of SSRF

In a Server-Side Request Forgery (SSRF) attack, the attacker reads and controls internal server resources by using the available functions on the server through web applications [4], as shown in Fig. 1.

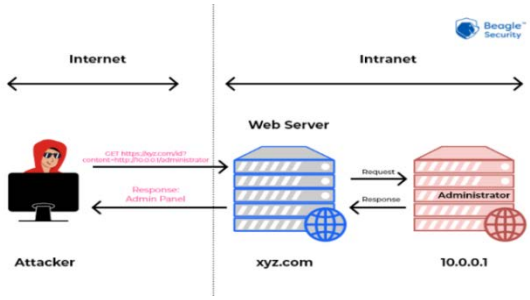


Fig. 1. Server-Side Request Forgery (SSRF) Attack [**4**].

Internal servers behind firewalls can be accessed by the attackers by submitting a URL within a web request to the web application.

A real example of an SSRF attack is the Capital One breach, where the database of Capital One Bank was hacked, and the information of more than 100 million customers was stolen. The attacker uses the Amazon Web Services credentials that were then used to access the Capital One database. Now a list of the main three types of SSRF attacks are explained [5].

- Non-Blind SSRF: As shown in Fig. 2, an attacker can access the data via the HTTP response. Server retrieves the contents of the resource located at the URL submitted, without verification, in an HTTP response to the user is given.

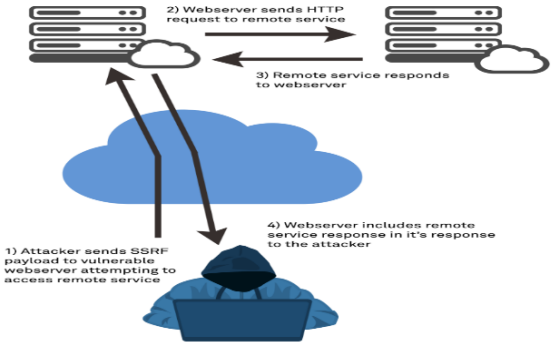


Fig. 2. Non-Blind SSRF [5].

- Blind SSRF: Fig. 3 illustrates this type of SSRF attack. When a web application has SSRF vulnerability but at the same time there is no HTTP response to the attacker. Here the attacker sends his own URL, he can access it, and the server sends an HTTP response to this URL. However, this method detects vulnerability, but it is possible that sensitive data will not be obtained.

Machine learning techniques greatly support artificial intelligence, as computer systems learn to perform human tasks such as classification and prediction. There are many machine learning algorithms and statistical models that are used to analyze data and extract knowledge from it. The process of training machine models produces a system capable of making decisions or recognizing objects. Machine learning algorithms are divided into several types according to the training method and the data available for this process.

Deep learning is a machine learning technique, and it is also considered as developing neural networks by adding new layers to them. Compared to neural networks, deep learning networks improve classifiers, especially as the volume of data increases [6]. As shown in Fig. 5, the performance of deep learning networks generally improves with increasing amounts of training data, unlike other types of machine learning techniques [7].

- Semi-Blind SSRF: Fig. 4 illustrates this type of attack. In the HTTP response, the server does not display all the details but only some of them. Data can be contained in error messages that enable the attacker to learn more information such as request response times, allowing the attacker to validate if a request succeeds.

Deep learning techniques are now used in all areas of artificial intelligence, recognition of voices, faces, text analysis, etc. The process of training deep learning networks often takes longer compared to other machine learning techniques, due to the presence of many parameters in deep learning algorithms.

The difference between neural networks and Deep learning is mainly found in the hidden layer structures. Neural networks contain one layer while deep learning contains several hidden layers. These multiple layers enable deep learning to recognize features in the data without human intervention [8]. Deep learning techniques vary according to several factors, as shown in Fig. 6 [9].

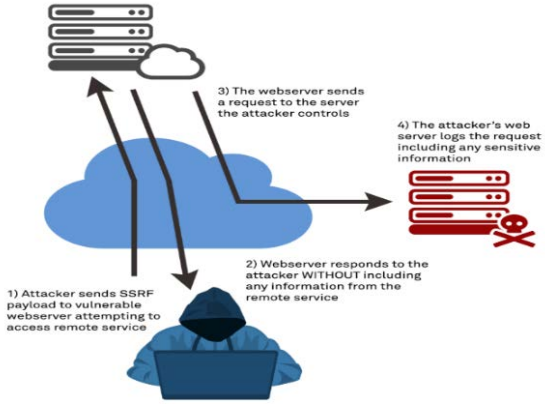


Fig. 3. Blind SSRF [5].

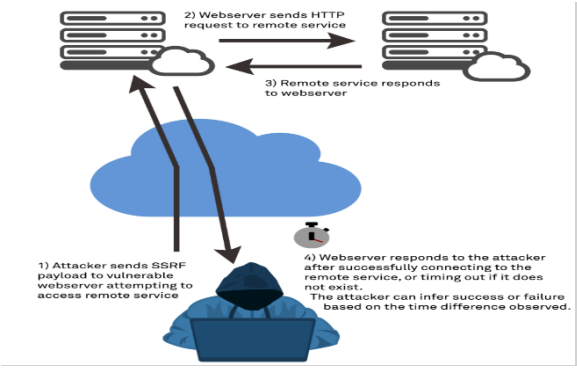


Fig. 4. Semi-Blind SSRF [5].

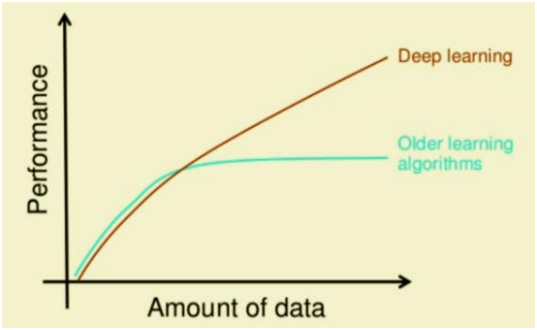


Fig. 5. Machine Learning Techniques Scale with Amount of Data (source [7]).

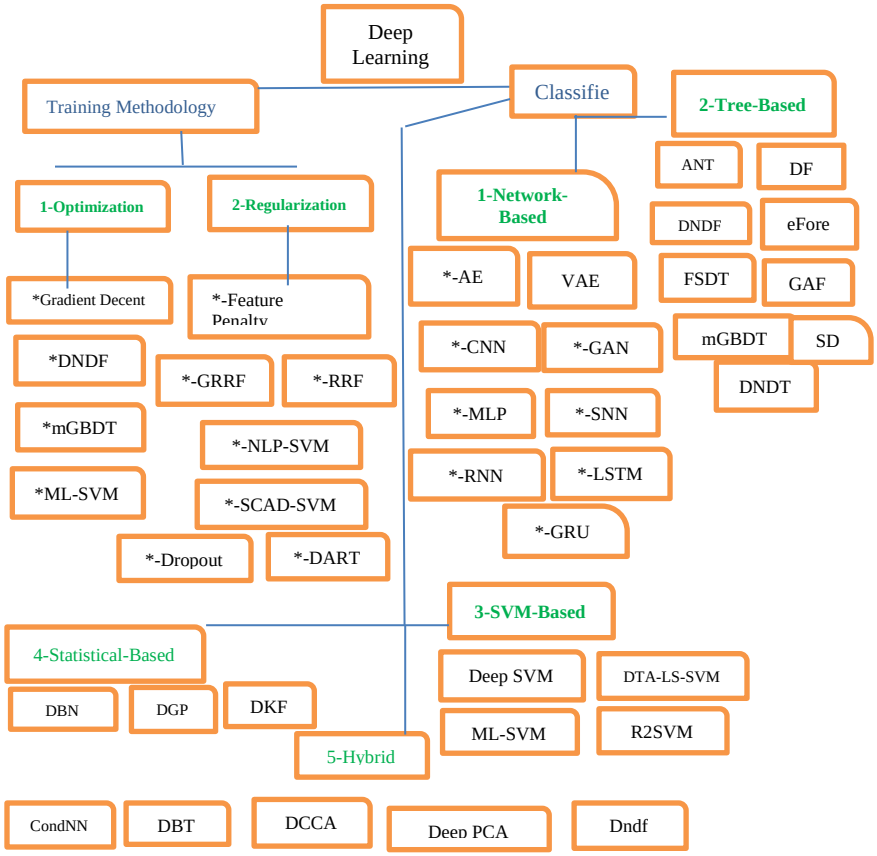


Fig. 6. Deep Learning Methods (source [9]).

II. Related Work

There are several studies concerned with how to detect attacks against web applications, as well as how to face these attacks and overcome them. Various techniques have been used to detect and mitigate these attacks, most of which are concerned with the use of machine learning techniques. we discuss many of the researches concerned with Preventing Server-Side Request Forgery Attacks by discovering them and preventing them from making any threat. Several methods have been used to detect this type of attack. Most of this research was based on artificial intelligence techniques because of their flexibility in application.

In [10], a defensive method was shown to protect web servers from SSRF attacks, and the researcher divided this method into several steps as shown in Fig. 7.

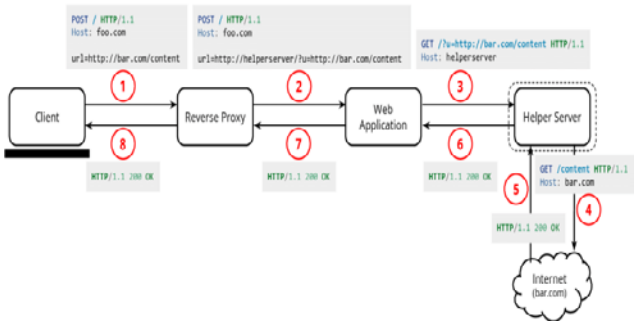


Fig. 7. Overall Architecture to Defenses from SSRF [10].

The reverse proxy located in front of the web application server, in the first step of the methodology, checks all the signals received from service seekers before allowing them to log in and use the application. If the reverse proxy finds a URL embedded in the request from the client to log in to the web application services, it automatically modifies the address and transfers this modified address to the web application server. Helper service takes the original value of the URL and executes it on the server since it cannot use services that are running on the internal network Experimental results show that the proposed solution can prevent all in-band SSRF attacks. Only one of them requires minimal developer collaboration In fact, little increase is observed in the response time in applications for the rest of the applications in the response time.

The researcher at [11] presented a deep learning-based model to discover attacks against web application. This model uses a deep learning technique which is an auto-encoder that is able to learn from the presence of a sequence of words while giving weight to these words according to their presence in the sequence. The model receives an entry request for the web application and then decodes and encodes the requests vector and calculates the reconstruction or loss error. If the loss error value is large then it classifies this request as anomalous requests, and conversely if the value is low then the request is classified as normal requests. The threshold θ is set to determine how small or large the loss error is. The experimental results show that the proposed model can detect web applications attack with low false positive rate and true

positive rate is 1. Because of less volume of labeled categorized anomalous dataset, the proposed classification engine is not 100 percent accurate; however, the classification can be improved with optimized training with a large volume of dataset, which is left as the future scope of the work.

In [12] end-to-end deep learning is applied to detect cyber-attacks astronomically in real-time. The intelligent part of the proposed framework is illustrated in Fig. 8. Authors evaluate the feasibility of an unsupervised/semi-supervised approach for web attack detection based on the Robust Software Modeling Tool (RSMT), which autonomically monitors and characterizes the runtime behavior of web applications. RSMT operates as a late-stage (post-compilation) instrumentation-based toolchain targeting languages that run on the Java Virtual Machine (JVM). It extracts arbitrarily fine-grained traces of program execution from running software and constructs its models of behavior by first injecting lightweight shim instructions directly into an application binary or byte code. These shim instructions enable the RSMT runtime to extract features representative of control and data flow from a program as it executes, but do not otherwise affect application functionality. shows the high-level workflow of RSMT's web attack monitoring and detection system. This system is driven by one or more environmental stimuli, which are actions transcending process boundaries that can be broadly categorized as either manual or automated.

In [13], an accurate and light-weight Android malware detection method is proposed. This method takes an APK-formatted Android app and passes it on to a deep learning network (1-D CNN) to be analyzed to discover the extent of the app's damage. The training and testing steps for this method are shown in Fig. 2. The training process is dependent on 5,000 malwares and 2,000 good wares. We confirmed our method using only the last 512-1K bytes of APK file achieved 95.40% in accuracy discriminating their malignancy under the 10-fold cross-validation strategy as shown in Fig. 9.

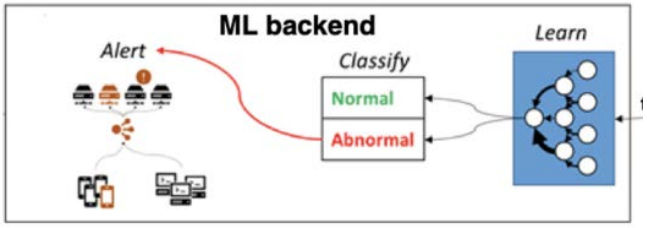


Fig. 8. Train the Classifier to Detect [12].

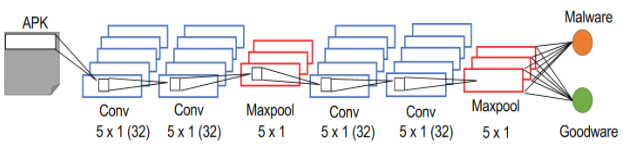


Fig. 9. 1-D CNN Model for Malware Discrimination [13].

In [14] a strategy is presented to discover real-time SSRF activities in the Amazon Web Services environment. This strategy consisted of several steps as follows:

- Detection using VPC Traffic Mirroring
- Detecting SSRF using Zeek
- Detecting SSRF using Suricata
- Detection using iptables

In [15], the researcher introduced a network-based intrusion detection system (NIDS) based on deep learning and machine learning techniques. This system is aimed at detecting intrusion on the network by examining the traffic through it. The general AI-based NIDS methodology for this paper is illustrated in Fig. 10 [16]. Table I covers the literature review evaluation.

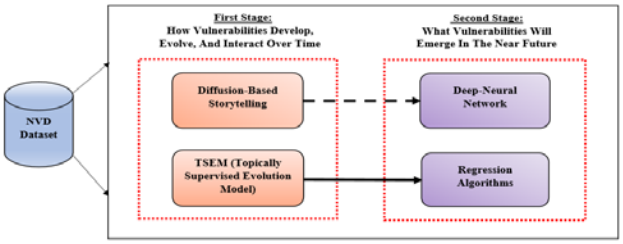


Fig. 10. Proposed Framework in Paper [16].

TABLE I. LITERATURE REVIEW EVALUATION

Paper	Contribution	Tool
[10]	a defensive method using was shown to protect web servers from SSRF attacks	Reverse Proxy
[11]	Detect attacks against web application.	Deep Learning
[12]	Detect cyber-attacks astronomically in real-time.	Deep Learning
[13]	detect malware targeting the Android system	1-D CNN
[14]	A strategy to Detect real-time SSRF activities in the Amazon Web Services environment	VPC Traffic Mirroring, Zeek, Suricata, iptables
[15]	a network-based intrusion detection system	deep learning and machine learning

III. METHODOLOGY

In this Paper, Long Short-Term Memory (LSTM) is used as a deep learning technique to build a model to detect SSRF attacks. Long Short-Term Memory (LSTM) architecture overcomes vulnerabilities in RNN networks [17]. In the first part of our methodology, we collect a dataset. Dataset obtained from the Canadian Institute for Cybersecurity of the University of New Brunswick. This dataset contains many features; we can mention some of them.

This dataset covered all types of SSRF attacks:

- Domain token count, avgpathtokenlen, tld, Arg URL Ratio.
- Number of DotsinURL, Arguments Longest Word Length.
- Spchar URL delimeter Doman, delimeter path, Number Rate.

- Directory Name, Symbol Count Domain, Entropy Domain.

Pre-processing and transformation the data before it is used to train and test machine learning algorithms is essential for creating high-accuracy models. Also, the data that determines whether there is an attack or not, be converted from text to numerical form to facilitate the training process of the deep learning algorithm. Then The text 'benign', 'Defacement' values have been converted to (0-1) to fit the LSTM deep learning network training process Among the important processes of data processing in this Paper is the transformation of data into a form suitable for analysis. Where scaling techniques were used to improve the possibility of recognizing data patterns by deep learning model. All data items is scaled to (-1,1) to enhance the training process.

A. *Learning Methodology*

The LSTM network that used to build the model in this Paper be explained in Fig. 11.

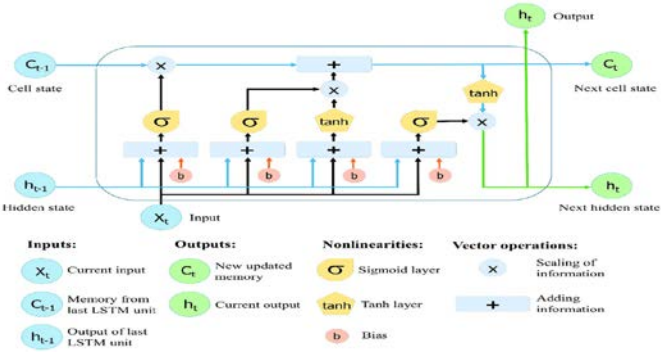


Fig. 11. LSTM Network.

LSTM networks are specifically designed to learn long-range dependencies and avoid problems with RNN. The idea of this type of network is to have a repeating edge inside each cell with weight w = 1. This eliminates the problem of the vanishing gradient problem, since repeated multiplication by 1 neither diverges nor converges to zero. Information flows on two levels from one step to the next while updating the weights is controlled by so-called gates. Gateways use various activation functions to redirect or stop the flow of information. The central change within the LSTM model as opposed to the RNN model or the feedforward model is a more complex type of hidden neuron. LSTM recursive networks contain 'LSTM cells' which have an internal repeat (self-loop), in addition to the external redundancy of the RNN. As can be seen from the previous figure, each element represents a vector with multiple properties that are placed across the grid showing three time steps and a dataset containing three sequential data samples (xt-1, xt, xt+1). The central LSTM cell is shown in detail to reveal the processes within it. The grid appears only until the hidden layers are output (ht-1, ht, ht+1).

LSTM network be used in this Paper to design an intelligent model to predict the SSRF attacks.

In the part of our methodology, we learn a deep learning model to be able to detect SSRF attacks as illustrated in Fig. 12.

In the part of our methodology, we will learn a deep learning model to be able to detect SSRF attacks as illustrated in Fig. 12. To train the deep learning model, the dataset is divided into two parts, the first for the training process and the second for the testing process. The percentage of training data was 80% of the entire dataset, while the percentage of test data was 20%. Then the Feature scaling method was applied to normalize the range of independent variables, where the data range became between -1 and 1. Data now is ready to train and test the model.

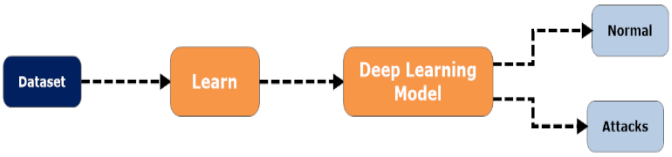


Fig. 12. Training Deep Learning Model.

After that, we can use our proposed model to detect any attacks in the HTTP Request and allow or block this entry according to the absence or presence of an attack as illustrated in Fig. 13.

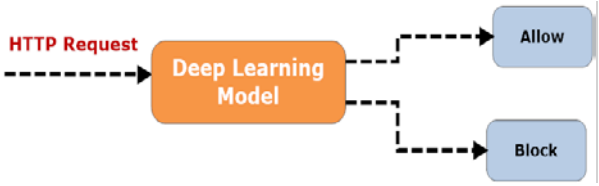


Fig. 13. Use Deep Learning Model to Detect SSRF Attack.

B. *Metrics Measures*

We rely on a set of measures to evaluate the model presented in this Paper, which is based on deep learning techniques. These measures can be summarized as follows:

1) True Positive (TP): The data instances correctly predicted as an Attack by the classifier.

2) False Negative (FN): The data instances wrongly predicted as Normal instances.

3) False Positive (FP): The data instances wrongly classified as an Attack.

4) True Negative (TN): The instances correctly classified as Normal instances.

The confusion matrix, which is generated at the end of the above-mentioned training and testing process, is used to calculate the preceding metrics. The four types (true positives, false negatives, false positives, and true negatives) as well as the positive and negative classifications are used in the template for any binary confusion matrix. Table II shows a 2x2 confusion matrix that can be used to express the four outcomes.

TABLE II. CONFUSION MATRIX PARAMETERS

Actual class	**Predicted class**		
		Class= yes	***Class=no***
	Class =yes	True positive	False negative
	Class=no	Fales positive	True negative

IV. Result and Discussion

Request Forgery (SSRF) attacks on web applications. The basic architecture of web applications and how this architecture is threatened are explained. Also, the types of SSRF attacks are described in this Paper. The proposed system is based on a technique of machine learning, which is deep learning technique, as it has the ability to learn, especially with the availability of a large amount of data.

A. Methods and Tools

The software and libraries needed to construct the LSTM model, which is utilized to identify SSRF assaults, will be detailed in this section. The programming language used is Python, which contains a huge number of external libraries. The diversity of these libraries and their easy handling of data has led to the widespread use of the Python language in all research Paper. Among these libraries used in the proposed Paper:

1) Numpy: This library is for scientific computing to work on the creation, editing and calculation of N-dimensional array objects.

2) Pandas: This library is used to handle large numerical tables with high performance. It can handle data that contains more than 2000 columns.

3) Scikit-learnrn: This library contains many methods that are used for data processing and evaluation of machine learning and deep learning models. It has been used in this Paper to evaluate the proposed deep learning model, using accuracy, precision, Recall, F1 score metrics.

4) Keras: It is an interface for programming neural networks and deep learning applications, by accessing the top of TensorFlow to create, train and test models. This interface supports the training of neural network models on the CPU or GPU. It also supports most types of deep learning networks such as CNN, RNN, and LSTM. In this Paper, keras was used to build an LSTM model for detecting SSRF attacks [18].

B. Development of the LSTM Model

In this part, a description of how to develop, train and evaluate an LSTM network be presented. At first, the data be read from the file that contains the dataset, then a figure be drawn showing the percentage of the data containing the attack as shown in Fig. 14.

After that, the data is scaled between 1 and -1 to enhance the accuracy of the deep learning model. To prepare dataset for training and testing process, it split to the training and testing parts using the following python code [19].

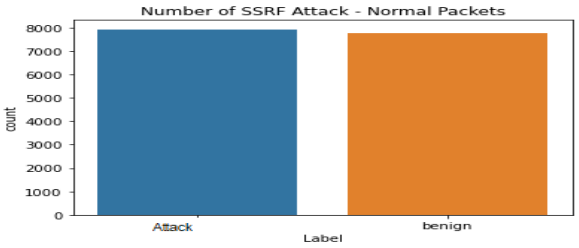


Fig. 14. Attack and Benign Percentage in the Dataset.

As a result of this code, the first 70% of the dataset is classified as training and the last 30% is used as testing data. The `X_train, X_test, Y_train, Y_test = train_test_split(X,y,test_size=0.3, random_state=6)` parameter values of LSTM network is optimized by testing its values for many times, and the final values are:

Batch size: 25

Epochs: 100

Hidden units: 20

Learning rate: 0. 005

Loss Function: MSE

A loss function quantifies how "good" or "bad" a given predictor is at classifying the input data points in a dataset. If the gap between training loss and validation loss is large its means that your model is overfitting and if training loss is large its means your model is underfitting. If your training loss and validation loss are overlapping or close to each other means your model is now good for prediction. Here the model is look good as illustrated in Fig. 15.

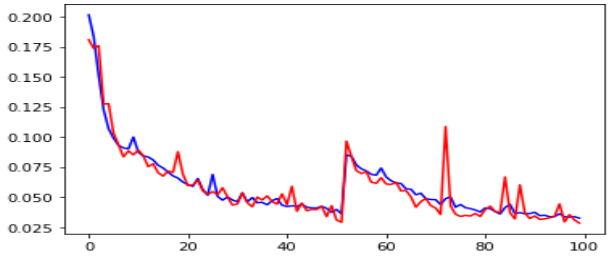


Fig. 15. Training and Validation Loss through Epoch Values.

C. Evaluation of the LSTM Model

The results of the attack detection on dataset are shown in Fig. 16 (Confusion matrix). The values of TP, FN, FP, TN is obtained from confusion matrix and can be summarized as follows:

1) True Positive (TP): 3015
2) False Negative (FN): 80
3) False Positive (FP): 112
4) True Negative (TN): 3078

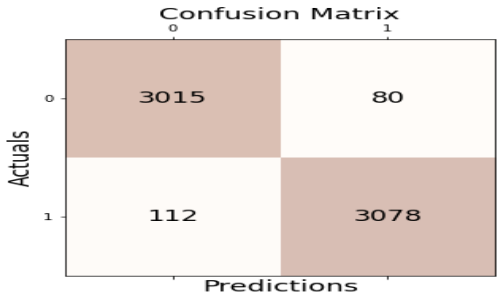


Fig. 16. Confusion Matrix for the Testing Results for LSTM Model.

From these values the metrics for evaluation the proposed model can be calculated as follows:

Precision: 0.975

Recall: 0.965

Accuracy: 0.969

F1 Score: 0.970

V. Conclusion

Server-side request forgery (SSRF) is a web application vulnerability that attackers exploit to access services on servers that an attacker is not allowed to use. In this paper a comprehensive review of the types of SSRF attacks and how they occur is presented. How to protect against these attacks is also explained with a review of literature reviews that provide various solutions to counter these attacks. Some machine learning and deep learning techniques have been demonstrated with an emphasis on the LSTM deep learning network. The LSTM network was used to design a deep learning model to detect SSRF attacks contained in URLs. This network was trained using a set of data after preprocessing operations were performed on it, and the data was scaled between values 1 and -1. This model was tested using several metrics such as accuracy to measure the accuracy of the proposed model in this Paper. The results of the deep learning model test were good, reaching 96%.

VI. Future Work

In the future, we will try to design other machine learning and deep learning models and compare them to get the most accurate model. Also, searching for another preprocessing operation to increase the accuracy of the deep learning model will be the focus of our next work.

References

[1] Michael Cross. 2007. Developer's Guide to Web Application Security. Syngress Publishing.

[2] Hoffman, A. (2020). Web Application Security: Exploitation and Countermeasures for Modern Web Applications.

[3] Web Application Security, accessed, March 2021, https://www.synopsys.com/glossary/what-is-web application-security.html.

[4] https://beaglesecurity.com/blog/article/server-side-request-forgery-attack.html.

[5] Server-Side Request Forgery (SSRF) & the Cloud Resurgence, (2020), accessed 2021, https://appcheck-ng.com/server-side-request-forgery-ssrf/.

[6] Palash Goyal, Sumit Pandey, and Karan Jain. Introduction to natural language processing and deep learning. In Deep Learning for Natural Language Processing, pages 1–74. Springer, 2018.

[7] Ng, A. (2016). Machine learning yearning: Technical strategy for AI Engineers, in the era of deep learning, draft version 0.5. Harvard Business Publishing.

[8] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85–117.

[9] Ghods, A., & Cook, D. (2019). A Survey of Techniques All Classifiers Can Learn from Deep Networks: Models, Optimizations, and Regularization. ArXiv, abs/1909.04791.

[10] Jabiyev, B., & Kharraz, A. (2020). Preventing Server-Side Request Forgery Attacks.

[11] Alma, T., & Das, M. (2020). Web Application Attack Detection using Deep Learning. ArXiv, abs/2011.03181.

[12] Pan, Y., Sun, F., Teng, Z., White, J., Schmidt, D., Staples, J., & Krause, L. (2019). Detecting web attacks with end-to-end deep learning. Journal of Internet Services and Applications, 10, 1-22.

[13] C. Hasegawa and H. Iyatomi, "One-dimensional convolutional neural networks for Android malware detection," 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 2018, pp. 99-102.

[14] Sean McElroy, Detecting Server-Side Request Forgery Attacks on Amazon Web Services ISSA Journal February 2020, volume 18, issue 2.

[15] Ahmad, Z., Khan, A., Cheah, W.S., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Trans. Emerg. Telecommun. Technol., 32.

[16] Williams, M., Barranco, R.C., Naim, S.M., Dey, S., Hossain, M.S., & Akbar, M. (2020). A vulnerability analysis and prediction framework. Comput. Secur., 92, 101751.

[17] Wichers, D.: OWASP Top Ten Project. https://owasp.org/www-project-top-ten/ Online; accessed March 2021].

[18] Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd.

[19] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

English Semantic Similarity based on Map Reduce Classification for Agricultural Complaints

Esraa Rslan[1], Rasha M.Badry[4]
Information Systems Department
Faculty of Computers and Information, Fayoum University
Fayoum, Egypt

Mohamed H.Khafagy[2]
Computer Science Department
Faculty of Computers and Information, Fayoum University
Fayoum, Egypt

Kamran Munir[3]
Computer Science and Creative Technologies Department
University of West of England
Bristol, United Kingdom

***Abstract*—Due to environmental changes, including global warming, climatic changes, ecological impact, and dangerous diseases like the Coronavirus epidemic. Since coronavirus is a hazardous disease that causes many deaths, government of Egypt undertook many strict regulations, including lockdowns and social distancing measures. These circumstances have affected agricultural experts' presence to help farmers or advise on solving agricultural problems. For helping this issue, this work focused on improving support for farmers on the major field crops in Egypt Retrieving solutions corresponding to farmer query. For our work, we have mainly focused on detecting the semantic similarity between large agriculture dataset and user queries using Latent Semantic Analysis (LSA) based on Term Frequency Weighting and Inverse Document Frequency (TF-IDF) method. In this research paper, we apply SVM MapReduce classifier as a framework for paralleling and distributing the work on the dataset to classify the dataset. Then we apply different approaches for computing the similarity of sentences. We presented a system based on semantic similarity methods and support vector machine algorithm to detect the similar complaints of the user query. Finally, we run different experiments to evaluate the performance and efficiency of the proposed system as the system performs approximately 77.8%~94.8% in F-score measure. The experimental results show that the accuracy of SVM classifier is approximately 88.68%~89.63% and noted the leverage of SVM classification to the semantic similarity measure between sentences.**

***Keywords*—Agricultural system; semantic textual similarity; text classification; latent semantic analysis; part of speech**

I. Introduction

The semantic similarity of sentences has many real applications like Intelligent Question Answering (IQA) system. When a question is asked, the existing answer can be returned if a similar question is found in the database. In this paper, we provided a solution for calculating semantic similarity between sentences that based on vectoring sentences using their syntactic and semantic features. Semantic Textual Similarity (STS) is focusing on finding the similarity between two sentences. Similarity between the sentences is based on the explicit or implicit semantic relationships between them[1]. These relationships can be identified or measured by finding semantic relations among them. Many algorithms are presented for textual similarity. We can group them based on the algorithm or method that we used to perform the semantic similarity process.

Agriculture has a huge impact in the economy of countries. Since over a huge number of the population in Egypt is dependent on agriculture. Moreover, it considers to be one the source national economy, foreign currency, Livelihood, and food supply [2]. Further, it creates job opportunities to a large scale of the population.

This paper uses the an English approach based on latent semantic analysis [3],[4] for measuring the semantic similarity between English sentences of agricultural data and user query to find the appropriate solution for the complaints of farmers. The proposed system used SVM classification in MapReduce Hadoop environment to classify the agricultural dataset complaints based on crop name to improve the efficiency of the semantic similarity process.

Therefore, the aim of the approach is providing the support for experts and farmers in the system in Egypt. The complaints' associations are distributed over around 4242 villages and 198' centers' across Egypt [5]. In Arabic script format, these complaints' are submitted to support for farmers in their agriculture problems. Storage all farmer agriculture problems stored on a public cloud which hosting analytics toolkits [6], [7].

In our approach, first; the farmer submits his agriculture problem in the Arabic language; then, Google machine translation is used to translate the problem from Arabic into English. Second, Analyses of the complaints through data analytics techniques to extract (most) term frequency and classify the query to which crop class using support vector machine in map/reduce model. The classification process might take some time to correctly classify the crops. Third, Building an automated support response by searching for similar complaints within the agriculture complaint datasets. We saved our dataset on the public cloud to store massive data or the number of complaints as big data. Our key focus has been on calculating the semantic similarity between Arabic and English cross-language sentences using LSA. We consider different methods like term frequency weighting and inverse document frequency to identify words in each complaint. The rest of the paper is presented as follows: Related Work in Section II describes a few Semantic Textual Similarity approaches. In Section III, Proposed System, we present our proposed LSA with SVM classification. Section IV, Discussion and Results describe the experimental results of

these methods. Finally, the Section V, Conclusion will be presented.

II. Related Work

Words can be similar in two ways lexical or semantical. Similar lexical words, if the words have the same sequence of character. Similar semantical words, if they have almost the same meaning, used in the same way, used in the same context. The String-Based algorithm is based on lexical similarity. Corpus-Based and Knowledge-Based algorithms are based on Semantic Similarity. String similarity measures operate on word sequences and character composition. It can be categorized into two sets: Character-Based Similarity, Term-based Similarity Measures. Character-Based Similarity like longest Common SubString (LCS) algorithm. N-gram algorithm Smith-Waterman [8]. Term-based Similarity Measures like Cosine similarity measure, Euclidean distance, Jaccard similarity, and Block Distance that also called Manhattan Distance.

Corpus-Based Similarity: Latent Semantic Analysis (LSA) [3] is the most popular technique of Corpus-Based Similarity. Hyperspace Analogue to Language (HAL), Generalized Latent Semantic Analysis (GLSA), Explicit Semantic Analysis (ESA), Normalized Google Distance (NGD). Knowledge-Based Similarity can be categorized into three groups like: (1) node-based/ information content (IC): like Resnik (res), and Conrath (jcn), (2) edged-based like Lesk, and vector pairs, and (3) hybrid where it combines both node and edge-based. We used LSA corpus-based algorithm in our work that depending on the corpus and word embedding to compute the semantic similarity degree between the sentences.

Nagoudi et al. [9] presented a word embedding representations for calculating the semantic similarity between Arabic and English sentences. This paper used machine translation and word embedding approach to get the properties of words like semantic and syntactic. Machine translation is used to translate English complaint into the Arabic one for applying a classical monolingual comparison. Word embedding methods are applied to measure the semantic similarity. The proposed method is used Bag-of-word alignment, IDF, and part of speech weighting to determine the most descriptive words in each sentence. The performance of this approach is evaluated on the four datasets of the shared task of SemEval in 2017. The results achieved the best accuracy rate compared to the other systems in the semantic text similarity in Arabic-English cross-language of SemEval 2017.

Wafa Wali et al. [4] proposed several methods for calculating the semantic similarity among two English sentences, which consider semantic and syntactic knowledge. It presented a technique for measuring sentence similarity, which combined the three components: lexical similarity, semantic similarity, and syntactic-semantic similarity. Lexical similarity included the common words, the semantic similarity used for finding the synonymy words, and the syntactic-semantic similarity based on common semantic arguments, thematic role, and semantic class. The word-based semantic similarity is measured for estimating the semantic degree among words by exploring the WordNet "[10] is a" taxonomy. Furthermore, the semantic argument is determined by the VerbNet database. The experiments are applied on the Microsoft Paraphrase Corpus and shown the metric F-score compared to other metrics. The results are shown that the proposed technique could support using several sentence features like semantic arguments and properties in measuring the sentence similarity. Therefore, this technique can be applied in many applications, such as plagiarism detection.

The author in [11] presented both the design and implementation of an evaluation system for English short answers. Handwritten Short Answer Evaluation System (HSAES) is an automated short answer system for determining the answer in answer papers and testing each short answer's marks depending on the model's knowledge during training. The proposed system was used the Optical Character Recognition (OCR) tools for extracting handwritten texts. Natural Language Processing is applied to retrieve the main feature from person tested datasets for answer keys and the handwritten of answer papers. The proposed system was used the cosine similarity approach for measuring the semantic similarity among sentences. Marks were given to each sentence in the evaluated answer paper. The developed model was applied for assessing the un-scored short answer marks.

Chandrathlake et al. [12] focused on providing an accuracy level for English news posts written on social media. The proposed system performed many functions: extract the news item's content, search the Internet for finding the similar posts in online articles sources, match the returned content with the online article sites' content and finally generate the accuracy level. Many Natural Language Processing techniques are used for developing this model like web scrolling, text summarization, URL ranking, and semantic similarity methods like Word2vec, part of speech, and cosine similarity. This system achieved an accuracy of 70% for the news posts on social media comparing with the trustable online news in the social media.

Taieb et al. [13] proposed a Features-based Measure of Sentences Semantic Similarity (FM3S) approach for computing the semantic similarity between English sentences. The proposed method combined three methods: the noun semantic similarity, the verb semantic similarity and the common word. This approach used the information content-based measure in computing similarity between keywords using the WordNet [14]. The experiments are performed and tested on the Microsoft Paraphrase Corpus (MPC) and scored the best results compared with other metrics for high similarity thresholds. The results showed that FM3S proved the importance of syntactic information, compound nouns, and verb tense in computing the semantic similarity.

Xiaolin Jin et al. [10] proposed a model based on Word2vec for measuring the semantic similarity between English sentences. This method was presented to solve the low universality problem and the contextual information's absence in calculating the word based on the dictionary. This method improved the approaches based on the Chinese dictionary, e.g., HowNet and Tongyici Cilin. It also used the word vector model as a weighing parameter for measuring the word

similarity after comparing the words' similarity by giving different weights to the three methods. The experiments were conducted on this algorithm and achieved a high Pearson coefficient. The proposed method could include most words that could effectively solve the word similarity calculation problem in the dictionary.

Many work related to SVM in the parallel environment (or distributed system) have introduced in Ngoc et al. [15], Wen et al [16], and Rao [17] . There are many researcher papers using Cloudera and Hadoop [18] Map/reduce. Studies using SVM [19] in parallel environment for semantic classification are proposed. However, there isn't work which combine them such as: Hadoop Map/ Reduce, SVM classification, parallel system, and semantic similarity. Our new system uses all of them.

III. Proposed System

This section presents the proposed system main steps as shown in Fig. 1. The proposed system has five steps: (1) translate the user query (farmer complaint) from Arabic into English language (2) preprocessing the farmer query; (3) classification method using SVM in map/reduce model (4) Finding the word vector, building the sentence vectors matrix using LSA; and computing the similarity between sentence vectors by using vector similarity methods like cosine similarity (5) The problems are ranked and then select the one with the highest semantic score.

A. Translation

In this step, the complaint text is translated into English language. We used Google Cloud translator API [20] to translate the Arabic sentence into English one.

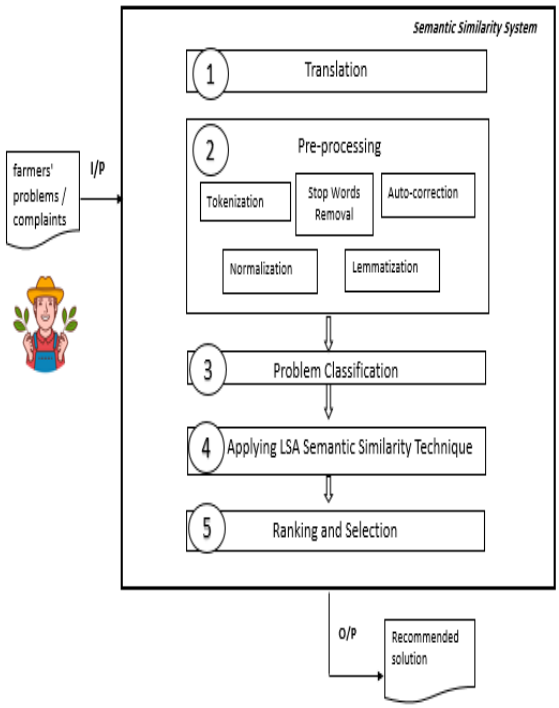


Fig. 1. The Proposed System.

B. Preprocessing

The farmer who describes the problem information like: the crop name, planting and watering method, and soil type. The farmer's query may contains useful words that effect the text processing phase. Pre-processing is important for removing the noise rows or data from historical complaint/response. It is focused on the historical agriculture dataset and farmer query to be used in this step [21]. Data pre-processing has many steps, like as: "tokenization", "stop words removal", "auto-correction", "normalization", and "lemmatization".

1) Tokenization : is the process of dividing written text into units (tokens) [21].White spaces, commas, semicolon and punctuations are used as a segment point in various languages especially in Arabic and English.

2) Stop words removal: is the process of removing unnecessary words. There are some words that are less importance, less useful, and less informative. These words are called stop words such as words in English complaints like "the", "is", "and", "an","a", etc. To enhance and generate a better solution, it is necessary to eliminate and remove these words using a predefined list. We also used the WordNet database that has a list of all English words. WorldNet is a huge lexical database of English verbs, nouns, pronouns, adjectives, and adverbs that are used for knowledge-based semantic similarity. WordNet's Relations make it a useful tool for natural language processing and computational linguistics.

3) Auto-correction: is used to correct errors made by the farmer when entering the complaint text. The complaints can also contain words written in a slang language. Auto-correction is used to solve such problem by replacing the incorrect word with the correct one.

4) Normalization: is the process of transforming the input text into a standard form. It focuses on removing inconsistent variations or unwanted data such as: "riice" is transformed to "rice".

5) Lemmatization: is the process of finding the base form of words, such as: "fruits" is transformed to "fruit".

C. Classification

In this phase, the classification is made semantically using SVM map reduce approach which is applied on the agriculture dataset and the farmer query in a parallel manner. The classification is paralleled between several machines using a Hadoop cluster with MapReduce [19], programming model for our work. Our approach is based on the English data set. The dataset is classified into a number of crop names like Wheat, Rice, Cotton, Local Bean, Tomato, Corn, Onion, and Beet Each group contains the SVM using Hadoop Map (M)/Reduce (R) is applied to classify the farmer query based on which crop class belongs to find the suitable solution.

D. LSA

Once the farmer complaint text and historical agriculture dataset are pre-processed and classified, and word vectors, the next step is to build a semantic model to compute the semantic similarity between the farmer query and the historical

agriculture dataset. LSA Algorithm builds in three main steps, Input Matrix Creation, Singular Value Decomposition (SVD), and Sentence Selection. Almost all previous works perform the first two steps of latent semantic similarity algorithm are in the same way. There is some difference in the word weighs like term-frequency and part of speech tagging which used to fill in the input matrix. Another difference is that they select words in the two sentence to measure the similarity [3] , [22]. The developed semantic model is based on LSA. LSA [23] is one of the most and important corpus-based techniques used for measuring semantic similarity. It consist of three steps are input matrix creation, singular value decomposition (SVD), and sentence selection.

A word co-occurrence matrix is calculated where the rows filling with the main words and columns filling with the sentences and the cells values have word occurrence counts. This matrix has an important underlying corpus so SVD dimensionality reduction is applied using a mathematical techniques. Such dimensionality reduction is highly used to: (i) minimize the output dimensionality and (ii) increase overall performance. Finally, the semantic score is calculated for each farmer complaint; then the sentences are ranked according to the semantic score to select the closest solution to the farmer query.

In this phase, an input matrix is computed for the farmer query and historical agriculture dataset. Each row in the matrix represents the word or term in the farmer query. Each column represents the problem. The cell value is the result of the intersection between term and problem. There are two ways that are used for filling the cell values, which are Term Frequency-Inverse Document Frequency (TF-IDF) or Term Frequency (TF). In TF-based LSA, the cells are filled with the term frequency (TF_i) of terms in the complaint query (C_j) according to Eq. 1.

$$W(t_{ij}) = tf_{ij} \quad (1)$$

Where $W(t_{ij})$ the weight of a term (i) in each problem (j), and tf_{ij} is the frequency of a term (i) in each problem text (j). In LSA bas based on TF, the cells are filled with the weight of term (i) in

problem statement (Cj) according to Eq. 2.

$$TF - IDF_{ij} = TF_{ij} * IDF_{ij} \quad (2)$$

Where TF-IDFij is TF is the frequency of a term (i) in each complaint statement (j), and IDF explain the importance of N terms between all problems.

E. Ranking and Selection

The semantic similarity is measured as the cosine value output between these sentences vectors. LSA system is generalized by changing rows with texts and columns with samples and can be used to compute the similarity between sentences, paragraphs, and documents. After applying the SVD matrix, the cosine similarity method will be calculated between the user complaint and each historical data problem to find the most suitable answer to identify the similarity among them. The cosine is calculated as Eq. 3:

$$\text{cosine similarity}(X, Y) = \frac{X.Y}{\|X\|.\|Y\|} \quad (3)$$

Ordering the agriculture problems based on to the semantic similarity result as shown in Fig. 2 decision function, and then select the problem (complaint) with the highest score based on the semantic similarity score.

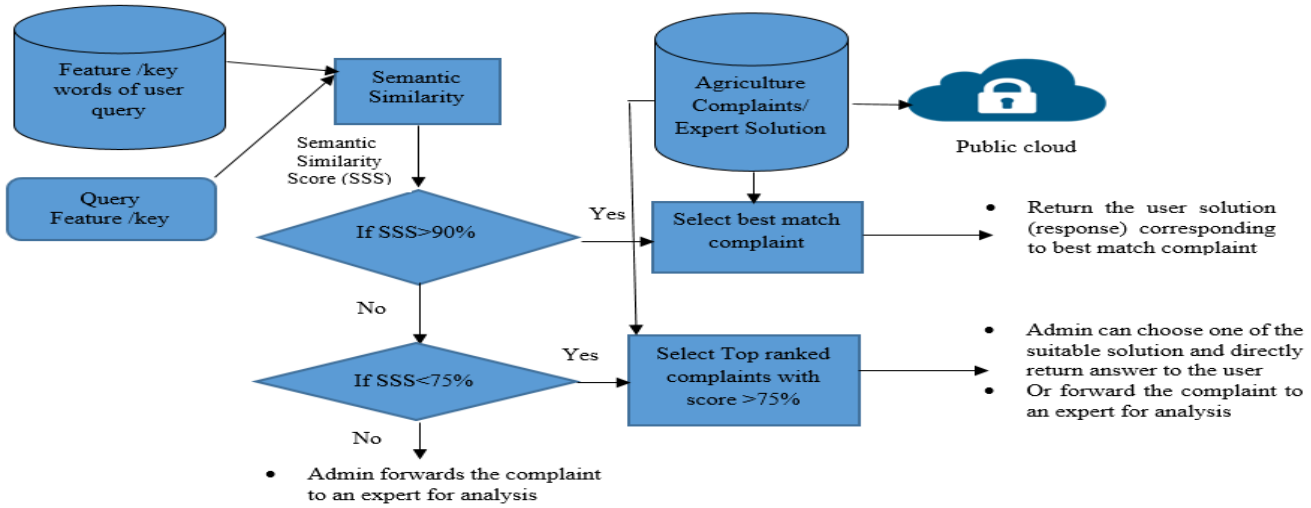


Fig. 2. Decision Function.

IV. Decision and Results

A. Environmental Setup

We used python programming language to implement our LSA model. The dataset is divided into 80% training and 20% testing with 10 experiments. We train the agriculture data set. The experiments are performed in devices with the following properties as shown in Table I.

B. Dataset

The dataset collected from Egypt's Virtual Extension Research Communication Network(VERCON) [5] and Agriculture Research Center (ARC), it contains historical complaints and solutions provided by the experts in text forms in English language complaints. The agricultural dataset was deployed on a public Cloud. The dataset is important because it has complaints/solutions from different agricultural problems that contain data for the main crop like: corn, cotton, wheat, and rice, also problem categories like environment, irrigation, pest, weed, diseases, and farming. Table II shows some examples of VERCON dataset. Table III presents the number of complaints in VERCON dataset in each crop.

C. Result Analysis

Consider the farmer query example: as presented in Table IV; firstly the farmer query is translated into English using google API. Secondly we apply preprocessing on the farmer query. Third, classify farmer query based on crop name by using SVM classifier in Hadoop Map/Reduce. Fourth create term frequency matrix. Fifth; compute semantic similarity score from the generating LSA matrix using TF-IDF or TF as shown in Tab. 3. The semantic similarity using TF-IDF achieved better result than the TF, because TF-IDF method shows the important features in each complaint however TF shows the number of term occurrence that appears in a complaint. Finally, ranking the complaints according to the semantic similarity score, and return solution of the farmer query with the highest similarity result.

We apply accuracy measure is to calculate the accuracy of SVM classification before semantic process. SVM is also used to predict the farmer query belongs to which crop class before being combined with semantic similarity process. The results show that the performance of classification with accuracy is approximately 88.68%~89.63%, as shown in Table V.

We proposed semantic similarity approach when using TF-based LSA and TF-IDF-based LSA. The work was evaluated using different measures like F-measure, precision and recall. F-measure expresses a trade-off between the two measures, precision, and recall as shown in Tables VI, VII and VIII. We compare our proposed results with the other models like POS (Part of Tagging) [24], [25]. The work was tested and evaluated on our agriculture data set. We test our dataset on different crop such as Wheat Rice, Cotton, Local Bean, Tomato, Corn, and other crops as shown in the following tables.

The F-measure in Table VI using TF-IDF weights scores the highest one about 0.939 in cotton crop, then the TF (term frequency) about 0.899 in the TF in Table VII while part of speech (POS) in Table VIII is 0.889. We run different kind of queries and get the average of F-measure, precision and recall. The F-measure in in Table VI using TF-IDF weights of F-score approximately 77.8%~94.8%, then the TF (term frequency) approximately 75.7%~92.3 while part of speech (POS) is approximately 73.3%~91.4%.

TABLE I. Environmental Setting

Item	Description
programming language	python
Processor	dual-core processor
CPU	Pentium CPU speed of 6.00 GHz
GPU(Graphics processing unit)	Tesla V100-SXM2-8GB
RAM (Random Access Memory)	8GB

TABLE II. Example of Dataset Complaints

Complaint	Solution
How to treat piercings with rice	Fyuridan is used at a rate of 6 kg per acre.
Yellowing of the lower leaves and the drying of the edges of the leaves in wheat.	Yellowing of the leaves at this time is normal, especially the lower one, to reach the ripening stage of the crop.
White spots on the leaf, spikes and stems feel cotton.	These symptoms of microflora disease, and if the infection requires severe chemotherapy, is spraying with a sumateite at a rate of 35 cm / 100 liters of water.

TABLE III. numer of complaints in historical dataset

Crop Name (English)	#of Complaints	Crop Name (English)	#of Complaints
Wheat	1073	Mango	435
Rice	1021	Citrus	254
Cotton	937	Grapes	247
Local Bean	783	Eggplant	227
Tomato	757	Green Pepper	199
Corn	648	Cucumber	178
Onion	546	Zucchini	168
Beet	461	Orange	153
Potato	440	Garlic	151
Clover	321	Guava	146

TABLE IV. An Example of System Process

Process	An example
Farmer query	يقع صفراء على أوراق نباتات البصل.
Translation	Yellow spots on the leaves of onion plants.
Tokenization	Yellow, spots, on, the, leaves, of, onion, plants
Stop word removal	Yellow, spots, leaves, onion, plants
Lemmatization	Yellow, spot, leaf, onion, plant
Classification	Onion class
Solution	These are the symptoms of onion thrips infection and it is treated with a 50% Acylac pesticide at a rate of 500 cm/100 liters of water, or a 72% silicron pesticide at a rate of 750 cm/f.

TABLE V. THE EVALUATION OF THE SVM CLASSIFICATION IN DATASET

Crop Name	# of records	Correct classification	Incorrect classification	Accuracy
Wheat	1073	955	118	89.19%
Rice	1021	909	112	89.03%
Cotton	937	834	103	88.98%
Local Bean	783	691	92	88.23%
Tomato	757	668	89	88.50%
Corn	648	572	76	88.43%
Onion	546	482	64	87.13%
Beet	461	407	54	89.23%
Potato	440	388	52	88.23%
Clover	321	282	39	88.58%
Mango	435	383	52	89.36%
Citrus	254	224	30	88.41%
Grapes	247	217	30	88.35%
Eggplant	227	203	24	89.64%
Green Pepper	199	178	21	89.11%
Cucumber	178	160	18	89.35%
Zucchini	168	151	17	89.06%
Orange	153	137	16	88.68%
Garlic	151	135	16	89.54%
Guava	146	131	15	89.14%
Summary	9145	8106	1039	88.63%

TABLE VI. EVALUATION MEASURES FOR USING TF-IDF FOR EACH CROP

crop name	TF-IDF		
	Precision	**Recall**	**F-score**
Wheat	0.849	0.855	0.852
Rice	0.841	0.856	0.848
Cotton	0.926	0.952	0.939
Local Bean	0.864	0.853	0.858
Tomato	0.893	0.866	0.879
Corn	0.867	0.899	0.883
Onion	0.955	0.941	0.948
Beet	0.902	0.895	0.898
Potato	0.925	0.915	0.921
Clover	0.822	0.856	0.839
Mango	0.785	0.795	0.793
Citrus	0.796	0.813	0.804
Grapes	0.773	0.784	0.778
Eggplant	0.899	0.875	0.887
Green Pepper	0.866	0.879	0.872
Cucumber	0.866	0.879	0.872
Zucchini	0.942	0.954	0.948
Orange	0.864	0.855	0.861
Garlic	0.941	0.948	0.938
Guava	0.796	0.813	0.804

The results show that the proposed text similarity LSA model using the TD-IDF method resolves the problem of the low recall of words in traditional semantic approaches well, and high the similarity performance of relevant words more than using only term frequency (TF) as shown in Tables VI, VII and VIII. The tables show the different measures for using TF-IDF, TF and POS for weeds, pests, diseases and irrigation categories.

TABLE VII. EVALUATION MEASURES FOR USING TF FOR EACH CROP

crop name	TF		
	Precision	**Recall**	**F-score**
Wheat	0.773	0.795	0.784
Rice	0.891	0.874	0.882
Cotton	0.902	0.896	0.899
Local Bean	0.881	0.854	0.867
Tomato	0.806	0.783	0.794
Corn	0.822	0.843	0.832
Onion	0.967	0.942	0.954
Beet	0.952	0.943	0.947
Potato	0.914	0.952	0.933
Clover	0.811	0.806	0.808
Mango	0.763	0.752	0.757
Citrus	0.752	0.767	0.759
Grapes	0.799	0.812	0.805
Eggplant	0.889	0.879	0.884
Green Pepper	0.853	0.861	0.857
Cucumber	0.856	0.879	0.789
Zucchini	0.921	0.926	0.923
Orange	0.864	0.831	0.847
Garlic	0.911	0.923	0.917
Guava	0.791	0.802	0.796

TABLE VIII. EVALUATION MEASURES FOR USING POS FOR EACH CROP

crop name	POS		
	Precision	**Recall**	**F-score**
Wheat	0.823	0.845	0.834
Rice	0.865	0.856	0.863
Cotton	0.897	0.881	0.889
Local Bean	0.832	0.889	0.862
Tomato	0.802	0.816	0.809
Corn	0.819	0.841	0.831
Onion	0.922	0.895	0.908
Beet	0.831	0.856	0.843
Potato	0.894	0.854	0.874
Clover	0.788	0.823	0.805
Mango	0.734	0.744	0.739
Citrus	0.744	0.723	0.733
Grapes	0.786	0.796	0.791
Eggplant	0.882	0.876	0.879
Green Pepper	0.869	0.856	0.862
Cucumber	0.855	0.879	0.788
Zucchini	0.909	0.911	0.914
Orange	0.823	0.822	0.822
Garlic	0.926	0.923	0.891
Guava	0.789	0.798	0.793

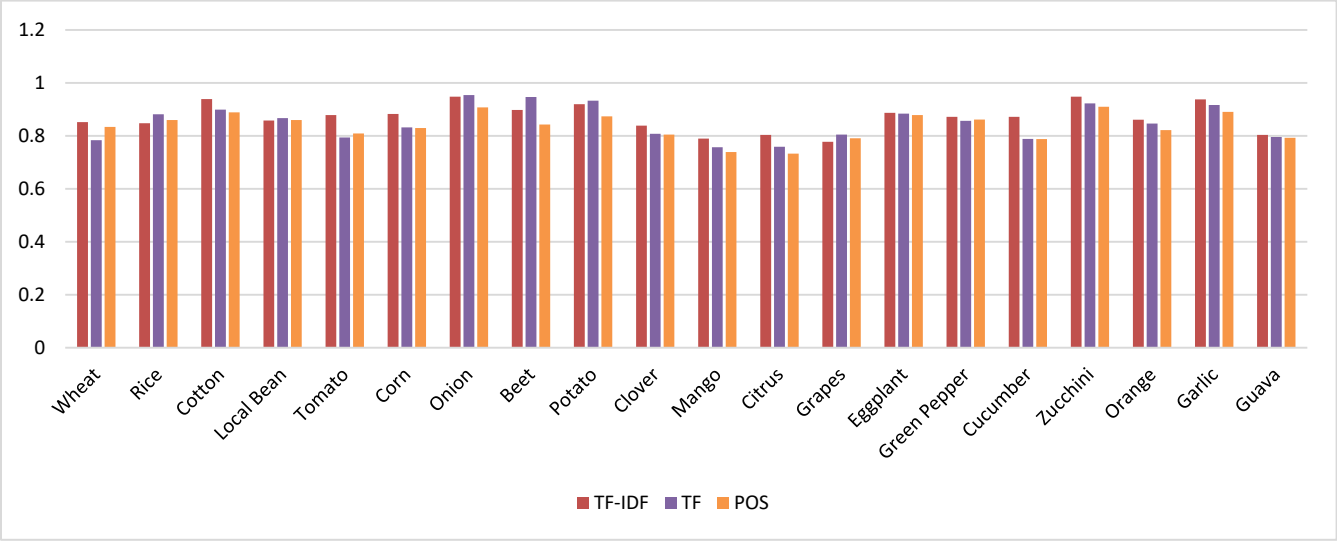


Fig. 3. F-score Values for different Metrics.

As a result, by evaluating the different experimental results of TF, TF-IDF and POS weight, we summarized that the results of LSA approach based on TF-IDF have the highest average F-measure as presented in Fig. 3.

V. Conclusion

Our work in this paper focused on building a semantic model for the available agricultural data, the design of interfaces and features for the system to ensure timely advice, easy access, consistency, and broadcasting service possible to farmers. We used MapReduce SVM classifier in Hadoop MapReduce to classify agricultural dataset into crops names. The performance of the system achieved better results than previous work. Also, we propose English semantic system for farmers' complaints that based on Latent Semantic Analysis depend on TF-IDF term to calculate similarity between user query and the complaints in the agriculture database. The results are tested on twenty different crops and also different complaint queries are applied on each crop. The system performed F-score with 0.939 using TF-IDF, then about 0.899 in the TF. The developed system with LSA based on TF-IDF achieved better results than the TF. The support provided by the system will be quickly and reliable not only for farmers but also for the 'research centers' and 'agricultural units' with minimal resources and training needs.

In the future work we will use different methods in semantic similarity process to enhance the system performance and also classify the dataset base on problem categories like pest, weed and irrigation.

Acknowledgment

This work has been supported by a Newton Institutional Links grant ID 347762518, under the Egypt Newton-Mosharafa Fund ID 30812 partnership. The grant is funded by the 'UK Department for Business, Energy and Industrial Strategy' and 'Science and Technology Development Fund (STDF)' and delivered by the British Council. For further information, please visit www.newtonfund.ac.UK.

References

[1] Chandrasekaran and V. Mago, "Evolution of Semantic Similarity - A Survey," arXiv, vol. 37, no. 4, 2020.

[2] G. Veeck, A. Veeck, and H. Yu, "Challenges of agriculture and food systems issues in China and the United States," Geogr. Sustain., vol. 1, no. 2, pp. 109–117, 2020, doi: 10.1016/j.geosus.2020.05.002.

[3] K. Al-Sabahi, Z. Zhang, J. Long, and K. Alwesabi, "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization," Arab. J. Sci. Eng., vol. 43, no. 12, pp. 8079–8094, 2018, doi: 10.1007/s13369-018-3286-z.

[4] W. Wali, B. Gargouri, and A. Ben Hamadou, "Sentence Similarity Computation based on WordNet and VerbNet," vol. 21, no. 4, pp. 627–635, 2017, doi: 10.13053/CyS-21-4-2853.

[5] "الهيئة العربية المصرية للاستشعار عن بعد." http://www.vercon.sci.eg/indexUI/uploaded/wheatinoldsoil/wheatinoldsoil.htm#r1 (accessed Sep. 01, 2021).

[6] "Apache Hadoop." http://hadoop.apache.org/ (accessed Sep. 01, 2021).

[7] V. N. Phu, V. T. N. Chau, and V. T. N. Tran, "SVM for English semantic classification in parallel environment," Int. J. Speech Technol., vol. 20, no. 3, pp. 487–508, 2017, doi: 10.1007/s10772-017-9421-5.

[8] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," Comput. y Sist., vol. 20, no. 4, pp. 647–665, 2016, doi: 10.13053/CyS-20-4-2506.

[9] E. Moatez et al., "Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences To cite this version : HAL Id : hal-01683494 Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences," 2018.

[10] X. Jin, S. Zhang, and J. Liu, "Word Semantic Similarity Calculation Based on Word2vec," ICCAIS 2018 - 7th Int. Conf. Control. Autom. Inf. Sci., pp. 12–16, 2018, doi: 10.1109/ICCAIS.2018.8570612.

[11] K. Al-Sabahi and Z. Zuping, "Document Summarization Using Sentence-Level Semantic Based on Word Embeddings," Int. J. Softw. Eng. Knowl. Eng., vol. 29, no. 2, pp. 177–196, 2019, doi: 10.1142/S0218194019500086.

[12] R. Chandrathlake, L. Ranathunga, S. Wijethunge, P. Wijerathne, and D. Ishara, "A Semantic Similarity Measure Based News Posts Validation on Social Media," 2018 3rd Int. Conf. Inf. Technol. Res. ICITR 2018, pp. 1–6, 2018, doi: 10.1109/ICITR.2018.8736136.

[13] B. Hassan, S. E. Abdelrahman, R. Bahgat, and I. Farag, "UESTS: An Unsupervised Ensemble Semantic Textual Similarity Method," IEEE Access, vol. 7, pp. 85462–85482, 2019, doi: 10.1109/ACCESS.2019.2925006.

[14] S. Zhang, Z. Liang, and J. Lin, "Sentence Similarity Measurement with Convolutional Neural Networks Using Semantic and Syntactic

Features," Comput. Mater. Contin., vol. 63, no. 2, pp. 943–957, 2020, doi: 10.32604/cmc.2020.08800.

[15] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4.5 algorithm for english emotional classification," Evol. Syst., vol. 10, no. 3, pp. 425–451, 2019, doi: 10.1007/s12530-017-9180-1.

[16] N. Yang et al., "Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification," Int. J. Speech Technol., vol. 20, no. 1, pp. 27–41, 2017, doi: 10.1007/s10772-016-9364-2.

[17] A. Haque and K. S. Rao, "Modification of energy spectra, epoch parameters and prosody for emotion conversion in speech," Int. J. Speech Technol., vol. 20, no. 1, pp. 15–25, 2017, doi: 10.1007/s10772-016-9386-9.

[18] G. S. Victor, P. Antonia, and S. Spyros, "CSMR: A scalable algorithm for text clustering with cosine similarity and MapReduce," IFIP Adv. Inf. Commun. Technol., vol. 437, pp. 211–220, 2014, doi: 10.1007/978-3-662-44722-2_23.

[19] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., vol. 12, no. 1, pp. 1–7, 2010.

[20] "Cloud Translation | Google Cloud." https://cloud.google.com/translate/ (accessed Sep. 01, 2021).

[21] D. A. Said, N. M. Wanas, N. M. Darwish, and N. H. Hegazy, "A Study of Text Preprocessing Tools for Arabic Text Categorization," Second Int. Conf. Arab. Lang. Resour. Tools, no. January 2009, pp. 230–236, 2009.

[22] Y. Liu, C. Sun, L. Lin, and X. Wang, "yiGou: A Semantic Text Similarity Computing System Based on SVM," no. SemEval, pp. 80–84, 2015, doi: 10.18653/v1/s15-2014.

[23] S. Alowaidi, M. Saleh, and O. Abulnaja, "Semantic Sentiment Analysis of Arabic Texts," vol. 8, no. 2, pp. 256–262, 2017.

[24] A. Voutilainen, "Part-of-Speech Tagging," Oxford Handb. Comput. Linguist., vol. 9780199276, no. June 2018, pp. 1–16, 2012, doi: 10.1093/oxfordhb/9780199276349.013.0011.

[25] V. Batanović and D. Bojić, "Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity," Comput. Sci. Inf. Syst., vol. 12, no. 1, pp. 1–31, 2015, doi: 10.2298/CSIS131127082B.

Multi-objective based Cloud Task Scheduling Model with Improved Particle Swarm Optimization

Chaitanya Udatha, Gondi Lakshmeeswari
Department of Computer Science and Engineering
GITAM (Deemed to be University)
Visakhapatnam, India

***Abstract*—Now-a-days, advanced technologies have emerged from the parallel, cluster, client-server, distributed, and grid computing paradigms. Cloud is one of the advanced technology paradigms that deliver services to users on demand by cost per usage over the internet. Nowadays, a number of cloud services have rapidly increased to facilitate the user requirements. The cloud is able to provide anything as a service over web networks from hardware to applications on demand. Due to the complex infrastructure of the cloud, it needs to manage resources efficiently, and constant monitoring is required from time to time. Task scheduling plays an integral role in improving cloud performance by reducing the number of resources used and efficiently allocating tasks to the requested resources. The paper's main idea attempts to assign and schedule the resources efficiently in the cloud environment by using proposed Multi-Objective based Hybrid Initialization of Particle Swarm Optimization (MOHIPSO) strategy by considering both sides of the cloud vendor and user. The proposed algorithm is a novel hybrid approach for initializing particles in PSO instead of random values. This strategy can obtain the minimum total task execution time for the benefit of the cloud user and maximum resource usage for the benefit of the cloud provider. The proposed strategy shows improvement over standard PSO and the other heuristic initialization of PSO approach to reduce the makespan, execution time, waiting time, and virtual machine imbalance parameters are considered for comparison results.**

Keywords*—Cloud computing; task scheduling; cloud service provider; virtual machines; PSO; multi-objective; cloud service broker

I. Introduction

The rapid growth of internet data processing prompted the creation of cloud computing systems. Cloud computing is critical for providing technology-based services through the use of the internet. It gives access to computing resources like storage, network and data without requiring active user control. Cloud environments can provide three distinct services: SaaS, PaaS, and IaaS. SaaS (Software as a Service) is the top layer service that distributes software to consumers. SaaS allows users to utilize software straight from the cloud without the need to install anything locally; you can access it immediately from the cloud.

The middle one is PaaS (Platform as a Service), it allows users to develop and deploy their own applications on top of the provided platform. Finally, IaaS (Infrastructure as a service) is the bottom layer service, the capability to deliver services as servers, storage and operating system and compute resources. Cloud providers use virtualization technology to provide consumers with computational resources virtually. Optimal task scheduling strategy is essential in the multi-tenant cloud computing model for enhancing the performance of a cloud environment. An efficient scheduling strategy enables the best virtual machine (VM) allocation to the required tasks in a way that to attain the required quality of service. The purpose of optimal allocation of tasks to VMs that fits certain criteria to obtain a specific objective as a result, the scheduling algorithm is a vital part of any cloud architecture.

A. Scheduling in Cloud Environment

Nowadays, everyone is trending towards advanced technology to save management efforts, time, and personnel. Cloud Computing is a new paradigm for hosting services and delivering those through the internet. Cloud dynamically provisions platform, infrastructure, and software applications as services to the cloud users based on the pay-as-you-go model, which means charging per usage. The cloud is a metaphor for the internet and virtualization technology is the key concept used to deliver services through the cloud by maintaining data integrity. Cloud environment is a pay as you go service model and it is an important aspect to business owners to compute huge amounts of data. Scheduling helps in better utilization of resources optimally. Thus, scheduling is the heart of cloud computing for the management of resources effectively.

Scheduling is categorized into two distinct levels:

1) The first level of Scheduling under IaaS is the Task or Workflow scheduling

2) The next level is the Virtual machine scheduling under IaaS.

Fig. 1 indicates that the primary type of scheduling technique used in cloud computing is further divided into two types: workflow task scheduling and independent task scheduling.

- Workflow task scheduling enables the tasks in a specific order because the tasks are interdependent, like a parent-child relationship.
- Independent task scheduling is converse to workflow scheduling in which all tasks are independent and there is no dependency among the set of tasks.

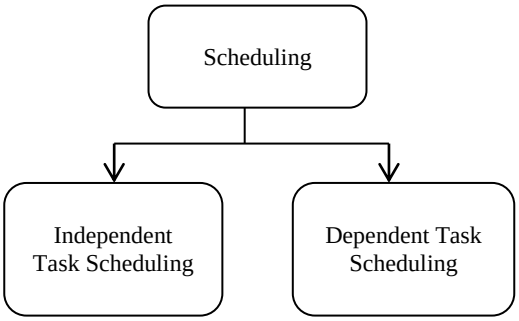


Fig. 1. Task Scheduling Categories.

Scheduling can be achieved with either static or dynamic scheduling approaches. The resources and scheduling strategy is pre-determined in static scheduling. Whereas in dynamic scheduling, the resources are allocated at the time of execution according to requirement and resource allocation can be modified during execution. Each scheduling technique can be achieved through different categories of heuristic, meta-heuristic, and combination of both approaches. Heuristic scheduling techniques are the most common type of scheduling methodology. Static scheduling is done by using heuristic methods, to give a single static solution. Cloud task allocation strategy is a type of NP-hard problem. Meta-heuristic algorithms are required to solve problems that are multidimensional in nature. They provide multiple solutions dynamically. Several researchers proposed multi-stage hybrid meta-heuristic algorithms to obtain better performance, combining heuristic and meta-heuristic approaches.

The following is a proposed work summary of the key contribution of task scheduling to the current literature:

- MOHIPSO solution for optimum task scheduling strategy employs a hybrid approach that combines two heuristic methodologies such as shortest job first and minimum execution time, to initialize the particles in PSO with a good starting point to explore the search space more efficiently instead of random values.
- Creating a multi-objective scheduling technique that reduces both task execution and waiting time for the benefit of the user in reducing the cost of application based on pay per usage policy and improving resource utilization, maximizing the profit for cloud provider by reducing makespan and degree of VM imbalance.
- The proposed method was implemented in CloudSim framework by extending JSwarn package and validating the proposed method with multi-objective-based standard PSO and SJF-PSO methods.

Section II focuses on the literature review, Section III focuses on the proposed MOHIPSO model for scheduling, the outcome of the proposed method is compared and analyzed in Section IV, and the conclusion in Section V.

II. Related Work and Background

Researchers sought to discover acceptable scheduling mappings in the cloud environment using various methodologies based on heuristic and meta-heuristic approaches. Many authors have improved scheduling strategy by novel variants in nature-inspired algorithms to enhance the global search capability of traditional standard techniques to avoid premature convergence. However, the focus on scheduling using multi-objective-based algorithms was minimal.

Bangyal, Waqas Haider, et al. [1] survey provides a complete overview of the different PSO and DE initialization procedures based on the Sobol, Halton, and random distribution families of quasi-random sequences. The fundamental purpose of the proposal was applied to various meta-heuristic approaches. It provides future work directions for the researchers.

Alsaidy et al. [2] proposed heuristic initialized PSO [3] [4] [5] outperforms among other approaches by considering of convergence and load balance, but it is a single objective based solution not able to satisfy the targets of multiple objectives by considering both cloud provider and user.

Bangyal et al. [6] proposed an enhanced version of bat algorithm using torus walk instead of uniform walk for improving local search and chaotic mapping [7] introduced for inertia weight to explore more in global search space for hyper dimensional global optimization problems. TW-BA is useful for the researchers to propose a new variant to all traditional nature-inspired algorithms.

Zhou, Zhou, et al. [8] introduced a unique variation of GA using a greedy approach to optimize scheduling strategy, which converges solution with very few iterations. This approach had considered only the makespan as a fitness function.

Ngatman et al. [9] survey on modified PSO compared to traditional PSO to solve issues of random initialization of population for convergence of best solution by exploring the search space effectively. This survey is useful for many authors to propose the advancement of PSO by considering the study.

Zhang et al. [10], [11] survey provided a thorough examination of PSO. PSO advancements by initializing with chaotic and quantum behavior, analyzed PSO with different population topologies, hybridization and extensions by discussing multiple objectives [12] and theoretical analysis of PSO were considered in various computing environments for targeting researchers from all engineering fields.

MOPSO [12] [13] [15] based new task scheduling model by E. S. Alkayal et al. [14] on a ranking strategy to achieve minimum waiting time and maximum throughput for only heterogeneous tasks but there is a chance of performance degradation for homogeneous tasks.

The proposed scheduling technique makes three main assumptions:

- The first assumption is that each task is an independent task.
- The second assumption is that users can submit n tasks, and are executed on m virtual machines, mapping tasks to VMs.

- The final one is that there is no task migration across virtual machines. That is a task cannot be assigned to multiple virtual machines at the same time.

A major challenging issues of scheduling is to schedule, distribute varying number of tasks to multiple virtual machines (VMs) and minimizing the turnaround time of a task. In the cloud environment which consists of number of data centers.

Each datacenter (DC) consists of hosts , i = 1, 2, . . ., N are presented and it can represented as:

$$DC = [H1, H2, \dots Hi \dots . HN]$$

Host(Hi) consists of VMs , i = 1, 2, . . ., N are presented in each host and it can represented as:

$$Hi = [VM1, VM2, \dots, VMi, \dots . VMN]$$

Similarly each VMi consists of tasks, i = 1, 2, . . ., N are executed on each virtual machine based on task scheduling algorithm as shown in Fig. 2.

$$VMi = [T1, T2, \dots, Ti, \dots, TN]$$

Each cloud user submitted job is considered as a task. The cloud broker acts on behalf of the user in a data center environment and abstracts VM management functions like VM creation, cloudlet assignment to these VMs, and VM destruction. CIS(Cloud Information Service) is one of the cloud entities which performs cloud resource registry and indexing. The datacenter informs CIS that they are ready to process the cloudlets. The cloud broker can communicate with this entity, which returns a list of all the VM IDs that have been registered and allocated to the tasks.

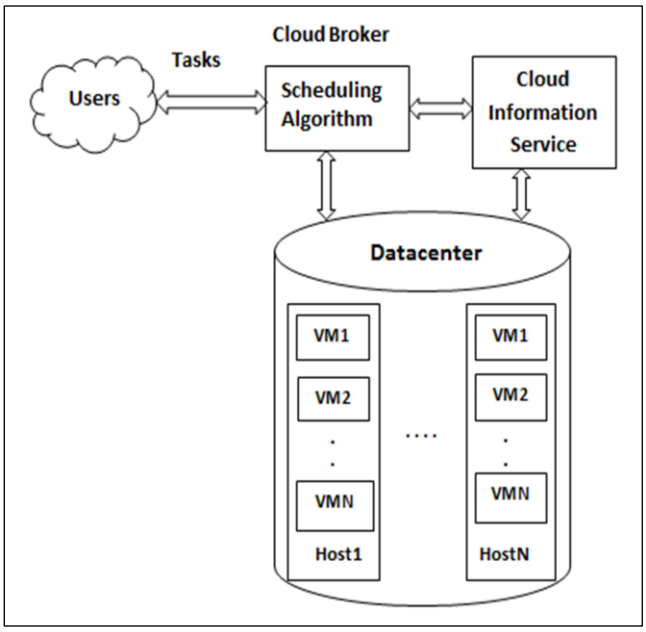


Fig. 2. Cloud Task Scheduling Model.

III. Problem Description for Proposed Methodolgy

In this section, the primary discussion is about the standard PSO method and MOHIPSO algorithm.

TABLE I. MOHIPSO Parameters

PSO Parameter	Value
Population	50
Iterations	100
α	0.05
β	0.7
γ	0.25
ω	0.9
$k1, k2$	2.0
$r1, r2$	[0,1]

Table I lists the PSO parameters considered for implementing MOHIPSO model. The PSO algorithm is a meta-heuristic population-based strategy for finding food sources in an optimal way influenced by the social behavior of birds flocking. PSO quickly gained popularity as a general purpose global optimizer. In this method, particles are moved across a multidimensional solution search space to find their destination. Each particle position changes in response to its own experiences as well as the experiences of others around it based on particle fitness value denoted by fit(X). The symbol $X(p)_i^{(t)}$ represents the location of particle position *i* at iteration *t*. It is possible to change the location of particle position $X(p)_i^{(t)}$ by adding the new random velocity $U(p)_i^{(t+1)}$ to the current position, as shown in the below equation:

$$X(p)_i^{(t+1)} = X(p)_i^{(t)} + U(p)_i^{(t+1)} \quad (1)$$

The particle velocity reflects the movement of particle position that is socially exchanged is given by the equation:

$$U(p)_i^{(t+1)} = \omega * U(p)_i^{(t)} + k1r1 * \left(X(p)_{pbesti}^{(t)} - X(p)_i^{(t)} \right) + k2r2 \left(X(p)_{gbesti}^{(t)} X(p)_i^{(t)} \right) \quad (2)$$

In (2), k1 and k2 are used as constant factors of a particle for personal and global influence, respectively. In this equation, ω denotes the inertia weight of a particle, which is used to control the movement of the particle velocity.

$$X(p)_{pbesti}^{(t+1)} = \begin{cases} X(p)_i^{(t+1)}, if\ f\left(X(p)_i^{(t+1)}\right) < f(X(p)_{pbesti}^{(t)}) \\ X(p)_{pbesti}^{(t)}\ otherwise \end{cases} \quad (3)$$

$X(p)_{pbesti}^{(t)}$ denotes local best position of the particle, $X(p)_{gbesti}^{(t)}$ denotes the particle best position of the entire swarm globally and $r1 \& r2$ denotes pseudo random values within the range between 0 and 1 at each iteration i.

B. Proposed MOHIPSO Model

One of the criteria used to classify the meta-heuristics algorithms for optimization problems by considering the no. of objectives, they are single objective, multi-objective [12] [13] and many objectives [11].

The propsed Multi-Objective based Hybrid Initialization of Particle Swarm Optimization (MOHIPSO) algorithm determines which virtual machines are most ideal for scheduling tasks and finds the most efficient task scheduling

schema. Total n tasks t1, t2, . ., tn are defined in the task scheduling model, and they must be allocated to m virtual machines (vm1, vm2,...,vmm) in order for them to be executed.

Algorithm 1: Pseudocode for MOHIPSO

Input: Tasks, Task length, VMs ,VM Processing rate
Output: Task scheduled on VM
Start:
Initialize particles Xi with SJFP-MET algorithm
For each iteration i
For each particle Xi
calculate TET , MS and DI applying Eqs. 3, 4 and 5.
calculate the fitness function fit (X) applying Eq. 6.
Find X(p)pbesti and update the velocity value by 2
Update the position of the particle according to Eq. 1.
End
Find X(p)pbesti
End
Output the optimal task scheduled with X(p)gbesti

In the proposed MOHIPSO, particles are initialized with a good starting point with the help of hybrid strategy of using minimum execution time (MET) [16] and shortest job to fastest processor (SJFP) [2] to explore the search space effectively instead of random selection. In each iteration multi-objective fitness value is used to find the particle local and global best values by updating velocity randomly and tasks are scheduled on VMs based on $X(p)^{(t)}_{pbesti}$.

C. Task Scheduling Problem Description

The scheduler determines which task should be assigned to which machine.

Cloud Task Scheduling helps in:

- Reducing operational cost.
- Reducing waiting time.
- Increasing resource utilization.

Table II lists cloud simulation parameters considered for the proposed task scheduling strategy implemented on the CloudSim framework. The simulation initially starts by initializing the CloudSim clock instance and creating a data center and datacenter broker. VM and cloudlet specifications as per Table II were created and submitted to the cloud broker. MOHIPSO model is used to schedule the tasks to specific VMs based on resource availability.

The proposed work improves the PSO algorithm for scheduling tasks in a cloud environment with a multi-objective decision problem. This context mainly considers the three objectives: execution time, makespan, and degree of imbalance. It is expressed as follows:

Execution Time (ET): The time required for processing a task on a particular virtual machine.

$$ET_{ij} = \frac{TLeni}{VMj}$$

TABLE II. CLOUD SIMULATION PARAMETERS

Cloud Parameter	Value
No.of Tasks/ Clouldets	10-50
Task Length (MI)	1100-2000
Cloudlet file size	300
Cloudlet output size	300
No.of VMs	5
VM Processing rate (MIPS)	500-900
VM RAM (MB)	512
VM Bandwidth (Mbps)	1000
Vmm Name	Xen
VM Pes Number	1
Data Center	1
Host	1

ET_i denotes Task i execution time on VM_j and T_{Leni} denotes ith task length specified in MI (Million Instructions).

Total Execution Time (TET): It is the summation of all tasks processing time.

$$TET = \sum ET_{ij} \quad (3)$$

Makespan: It is the last task finishing time on virtual machine.

$$MS = Max\ \{ ET_{ij} \} \quad (4)$$

MS is the maximum make span of all VMs.

Degree of Imbalance (DI): The difference between the maximum and minimum execution time and the total execution time.

$$DI = \frac{Max\{ ETij \} - Min\{ ETij \}}{TET} \quad (5)$$

The fitness function in (6) is calculated based on TET, MS and DI using weighted sum method for MOHIPSO method as follows:

$$fit(X) = Min\{\alpha * TET + \beta * MS + \gamma * DI \} \quad (6)$$

Where parameter α refers to the weight of total execution time, β refers to the weight of makespan and γ refers to the weight of degree imbalance. fit(X) function is considering three parameters specified in the equations (3),(4) &(5) and three control parameters such as α, β and γ with in the range of [0,1] and sum of α+β+γ=1 and these parameters values are specified in Table I.

IV. RESULTS AND COMPARISION

MOHIPSO algorithm is compared with two other variants of PSO algorithm. The first variant of PSO is a multi-objective based PSO algorithm by considering MS, TET and DI objectives for finding the fitness value of a particle and the second variant of PSO is called as SJF-PSO, in which initialization of particles with SJFP algorithm and multi-objective based fitness value is calculated to map the tasks to VMs. Finally proposed MOHIPSO is a combination of Standard PSO with MET and SJF-PSO gives better performance compared with existing algorithms.

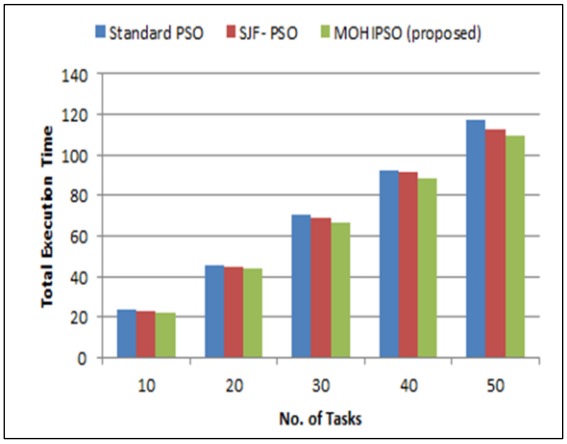


Fig. 3. Comparision of Total Execution time between PSO Varients.

Fig. 3 indicates the number of tasks vs. total execution time, in which the proposed algorithm MOHIPSO shortens the execution time compared other standard PSO and variant of PSO.

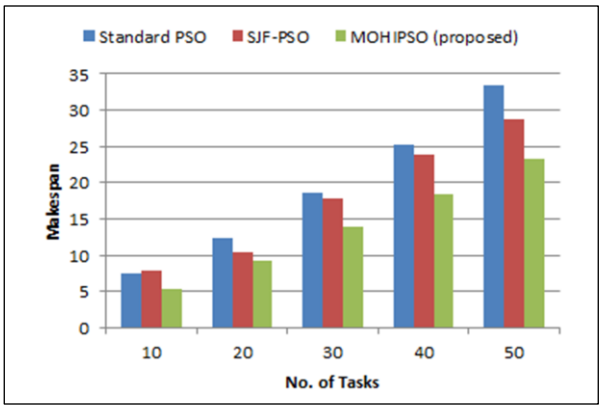


Fig. 4. Comparision of Makespan between PSO Varients.

In Fig. 4, the comparison between nuber of tasks and makespan, the proposed algorithm reduces the makespan compared to the standard PSO and SJF-PSO

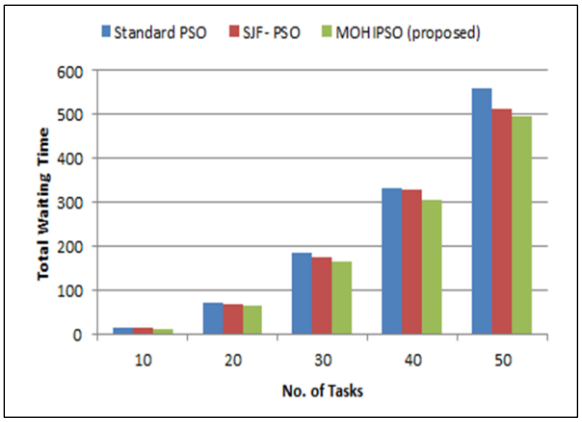


Fig. 5. Comparision of Total Waiting Time between PSO Varients.

Fig. 5 indicates the number of tasks vs. total waiting time of all tasks. The proposed algorithm reduces the waiting time compared to the standard PSO and SJF-PSO.

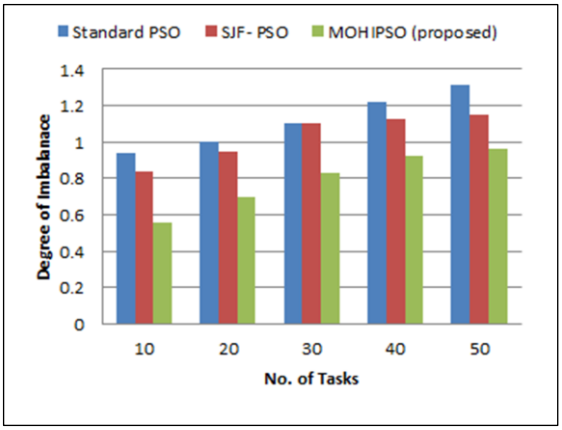


Fig. 6. Comparision of Degree of VM Imbalance between PSO Varients.

Fig. 6 shows the graph between tasks and degree of VM imbalance, it clearly shows that as the number of tasks increases then DI is increased. In order to reduce this, MOHIPSO algorithm considers the best way of scheduling tasks to the given resources on efficient way.

V. Conclusion and Future Work

Scheduling plays a crucial role in the cloud environment for effective distribution of tasks to enhance the quality of service. The MOHIPSO satisfies multiple objectives and provides obvious improvements in terms of makespan, execution time, waiting time, and degree of VM imbalance compared to traditional PSO and SJF-PSO.The simulation results show that the MOHIPSO has improved.

Task scheduling is not a multi-objective solution but it is to be a many-objective based solution by considering multiple objectives on both sides of the cloud provider and user. In future, the proposed algorithm can be extended to appraise other quality parameters like energy consumption and cost apart from makespan, execution time, waiting time, and degree of imbalance.

References

[1] Bangyal, Waqas Haider, et al. "Comparative Analysis of Low Discrepancy Sequence-Based Initialization Approaches Using Population-Based Algorithms for Solving the Global Optimization Problems." *Applied Sciences* 11.16 (2021): 7591.

[2] Alsaidy, Seema A., Amenah D. Abbood, and Mouayad A. Sahib. "Heuristic initialization of PSO task scheduling algorithm in cloud computing." *Journal of King Saud University Computer and Information Sciences* (2020).

[3] Madni, Syed Hamid Hussain, et al. "Performance comparison of heuristic algorithms for task scheduling in IaaS cloud computing environment." *PloS one* 12.5 (2017): e0176321.

[4] Al-Qerem, Ahmad, and Ala Hamarsheh. "Statistical-Based Heuristic for Tasks Scheduling in Cloud Computing Environment." *International Journal of Communication Networks and Information Security* 10.2 (2018): 358-365.

[5] Mansouri, Najme, Behnam Mohammad Hasani Zade, and Mohammad Masoud Javidi. "Hybrid task scheduling strategy for cloud computing by modified particle swarm optimization and fuzzy theory." *Computers & Industrial Engineering* 130 (2019): 597-633.

[6] Bangyal, Waqas Haider, Jamil Ahmed, and Hafiz Tayyab Rauf. "A modified bat algorithm with torus walk for solving global optimisation problems." *International Journal of Bio-Inspired Computation* 15.1 (2020): 1-13.

[7] Abdullahi, Mohammed, et al. "An efficient symbiotic organisms search algorithm with chaotic optimization strategy for multi-objective task scheduling problems in cloud computing environment." *Journal of Network and Computer Applications* 133 (2019): 60-74.

[8] Zhou Zhou, Fangmin Li, Huaxi Zhu3, Houliang Xie, Jemal H. Abawajy, and Morshed U.Chowdhury "An improved genetic algorithm using greedy strategy toward task scheduling optimization in cloud environments." *Neural Computing and Applications* 32.6 (2020): 1531-1541.

[9] Ngatman, Mohd Farhan, Johan Mohd Sharif, and Md Asri Ngadi. "A study on modified PSO algorithm in cloud computing." *2017 6th ICT international student project conference(ICT-ISPC)*. IEEE, 2017.

[10] Zhang, Yudong, Shuihua Wang, and Genlin Ji. "A comprehensive survey on particle swarm optimization algorithm and its applications." *Mathematical problems in engineering* 2015 (2015).

[11] Geng, Shaojin, et al. "Many-objective cloud task scheduling." *IEEE Access* 8 (2020): 79079-79088

[12] Gobalakrishnan, N., and C. Arun. "A new multi-objective optimal programming model for task scheduling using genetic gray wolf optimization in cloud computing." *The Computer Journal* 61.10 (2018): 1523-1536.

[13] Langhnoja, Himani K., and Hetal A. Joshiyara. "Multi-Objective Based Integrated Task Scheduling In Cloud Computing." *2019 3rd International conference on Electronics,Communication and Aerospace Technology (ICECA)*. IEEE, 2019

[14] Alkayal, Entisar S., Nicholas R. Jennings, and Maysoon F. Abulkhair. "Efficient task scheduling multi-objective particle swarm optimization in cloud computing." *2016 IEEE 41st Conference on Local Computer Networks Workshops (LCN Workshops)*. IEEE, 2016.

[15] Alguliyev, Rasim M., Yadigar N. Imamverdiyev and Fargana Jabbar Abdullayeva."PSO-based load balancing method in cloud computing." *Automatic Control and Computer Sciences* 53.1 (2019): 45-55.

[16] Almazok, Salem A., and Bülent 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* 2020.1 (2020): 1-26.

GML_DT: A Novel Graded Multi-label Decision Tree Classifier

Wissal Farsal, Mohammed Ramdani, Samir Anter
Computing Laboratory of Mohammedia (LIM)
FSTM, HassanII University of Casablanca
Morocco

***Abstract*—The goal of Graded Multi-label Classification (GMLC) is to assign a degree of membership or relevance of a class label to each data point. As opposed to multi-label classification tasks which can only predict whether a class label is relevant or not. The graded multi-label setting generalizes the multi-label paradigm to allow a prediction on a gradual scale. This is in agreement with practical real-world applications where the labels differ in matter of level relevance. In this paper, we propose a novel decision tree classifier (GML_DT) that is adapted to the graded multi-label setting. It fully models the label dependencies, which sets it apart from the transformation-based approaches in the literature, and increases its performance. Furthermore, our approach yields comprehensive and interpretable rules that efficiently predict all the degrees of memberships of the class labels at once. To demonstrate the model's effectiveness, we tested it on real-world graded multi-label datasets and compared it against a baseline transformation-based decision tree classifier. To assess its predictive performance, we conducted an experimental study with different evaluation metrics from the literature. Analysis of the results shows that our approach has a clear advantage across the utilized performance measures.**

Keywords*—Graded multi-label classification; algorithm adaptation; decision tree classifier; label dependencies

I. Introduction

Multi-label classification (MLC) has become an extensively researched and prominent field in machine learning. This is attributed to various real world applications that the traditional task of classification could simply not cover. Instead of predicting one class at a time, MLC predicts multiple classes at the same time. The classes are predicted based on a relevance/non relevance paradigm, while this task has proven to be useful, it remains limited as to the information it provides. Hence, an extension of MLC called Graded Multi-label Classification (GMLC) was proposed in [1].

GMLC assigns a degree of relevance or membership for each label to an instance. The degrees of relevance are gradual memberships in the sense of fuzzy set theory. A Covid-19 article, for example, may belong to three classes {health, economy, society} at the same time. However, the degree of membership to each class differs. The article can fully belong to the class health while it remains somewhat socio-economical.

In this light, all multi-label problems are graded multi-label problems, where the membership degrees are reduced to two binary values, relevant/non relevant. However, the reverse is not true, and reducing the graded multi-label problem to a standard multi-label problem was shown to decrease the predictive performance [1]. Hence the need of graded multi-label classifiers that can generalize the multi-label learning to encompass graded multi-label learning tasks.

Research on multi-label learning in recent years provided solutions in a variety of real-word problems where the traditional learning paradigms were not applicable, ranging from text categorization [2], automatic video and image annotation [3] [4] [5], web mining [6], information retrieval [7] to medical research and bioinformatics [8] [9] [10]. The different algorithms and approaches proposed exploited the transformation and the adaptation methods [11]. The diversity of these approaches is necessary to answer various real-world applications. In bioinformatics, for example, and more specifically genomics, the adaptation of the decision tree classifiers was proven to be very important in deducing comprehensible and readable rules predicting the functional classes of the genes.

Similarly, the ongoing research on Graded multi-label classification aims at developing solutions for real-world problems where the multi-label learning paradigm is not applicable or not optimal. For this purpose, some graded multi-label classifiers were proposed [1] [12] [13] [14]. However, to the best of our knowledge, there are no adaptation-based classifiers for GMLC in the literature.

In this paper, we propose a novel adapted decision tree classifier (GML-DT) that is suited for the graded multi label setting. The main advantages of this approach are its ability to fully model the label dependencies which improves the quality of its predictions. Furthermore, this algorithm is the first adapted tree-based model which makes it the most interpretable existing approach in GML. It is the only model in the literature that constructs a single decision tree from which a set of intelligible and accurate rules can be easily extracted.

The rest of the paper is organized as follows: Section II reviews previous works on GMLC and adapted decision trees in MLC. Section III presents the GML-DT algorithm. Section IV displays the experimental results on real-world graded multi-label data. Finally, Section V concludes this work and introduces future perspectives.

II. Related Work

In this section, we go over related work in both graded multi-label classification and multi-label classification.

A. Graded Multi-label Classifiers

Graded multi-label classification [1] was formalized as an extension of multi-label classification [15] [16], to predict the degrees of relevance of the labels rather than the set of relevant labels. By extension, graded multi-label classifiers fall within two main categories, problem transformation and problem adaptation. The former transforms the multi-label problem into a combination of regular classification tasks for each class label, whereas the later modifies directly the classifier to deal with graded multi-label data.

Cheng et al. [1] proposed a solution by decomposing the problem into an ordinal classification problem and a multi-label classification problem. They introduced two transformation methods, namely the vertical reduction, which predicts the membership degree for each label, and the horizontal reduction which predicts a subset of labels on each grade level. The authors also proceeded to prove the usefulness of graded multi-label classification, by deploying and comparing their approach on both GML data and ML data. Although the model proved to be effective in this setting, it does not model label dependencies. Brinker et al. [12] applied pairwise decomposition using three variants of Calibrated Label Ranking [17], to model the preferences between labels. While these approaches outperformed the predictive model developed in [1], they can only model pairwise dependencies. Lastra et al. [13] proposed a non-deterministic learner based on binary relevance that returns an interval whenever the classification is uncertain for a label. This method relies on a tradeoff between the size of the interval and the improvement of the accuracy.

Laghmari et al. [14] introduced an approach for learning label dependencies and label preferences. This is achieved by using the horizontal decomposition to reduce the problem into a combination of multi-label learning tasks, and then combining pairwise comparisons and classifier chains [18], which is an extension of binary relevance consisting of adding the labels as descriptive attributes.

While transformation methods can be easily implemented with the existing algorithms, their inability to fully model label dependencies and their run time can render them inefficient. This is especially true in cases where an interpretable model is needed, specifically a tree-based one capable of inferring accurate rules, which is the focus of this article. In fact, if one is interested in a model that produces rules identifying the features relevant for the prediction, these approaches would be inefficient and even inapplicable. The approaches in [1] can only identify the features relevant for one class label. Pairwise comparisons are not sufficient to fully model label dependencies. Classifier chains method is not applicable in this setting since it includes the labels as features, which would result in unintelligible rules. Furthermore, these approaches have to build a number of learners, proportional to the number of class labels, which affects their run time and interpretability.

B. Multi-label Decision Tree Classifiers

Clare et al. [19] adapted the c4.5 algorithm to handle multi-label data. The authors modified the formula of the entropy to account for the existence and non-existence of each label, and thus producing a decision tree capable of predicting all the class labels at once. The multi-label entropy is calculated as follows:

$$Entropy(S) = -\sum_{i=1}^{N} p(c_i)\log p(c_i) + q(c_i)\log q(c_i) \quad (1)$$

Where $p(c_i)$ is the probability of the class label c_i

$$q(c_i) = 1 - p(c_i)$$

Blockeel et al. [20] proposed a hierarchical multi-label decision tree, based on predictive clustering trees [21]. The tree is built by recursively partitioning the data into smaller clusters. This is achieved by finding the best attribute-value that reduces the intra-cluster variance. Where the variance is calculated based on the weighted Euclidean distance. Following this work, Vens et al. [22] presented an empirical study confirming the findings in [19] [20] and thus proving the ineffectiveness of transformation-based decision tree learners in comparison to adapted multi-label decision trees.

III. Graded Multi-Label Decision Tree

A. Formal Task Description

In graded multi-label classification, we have a number of training examples from which we build a classifier that assigns a grade or membership degree to each class label. An instance is represented as a vector x of d attribute values $x = [x_1, \dots, x_d]$ drawn for an input domain $A_1 \times \dots \times A_d$. Given $L = \{\lambda_1, \dots, \lambda_n\}$ a finite set of predefined labels and $M = \{\mu_1, \dots, \mu_m\}$ a finite set of predefined ordered membership degrees such that $\mu_1 < \mu_2 < \cdots < \mu_m$ ranging from complete irrelevance to full relevance. An instance x is assigned a vector of membership degrees $y_x = [y_x^1, \dots, y_x^n]$, where y_x^i corresponds to the degree of relevance of the ith label λ_i for the instance x.

We define a graded multi-label classifier H: $A_1 \times \dots \times A_d \to M^n$ as $(x) = \hat{y}_x$, where $\hat{y}_x = [\hat{y}_x^1, \dots, \hat{y}_x^n]$ corresponds to the set of predicted membership degrees for each label $\lambda_i \in$ L and an instance x.

B. GML_DT: Graded Multi-label Decision Tree

To deal with graded multi-label learning tasks, we propose a novel graded multi-label decision tree algorithm (GML-DT), capable of predicting the membership degrees of all target labels simultaneously.

The GML_DT, given in Algorithm 1, is a greedy model that follows a top-down induction approach for building decision trees. The algorithm takes as input the training set. It starts by searching for the best attribute-value test for the root node. It proceeds to splitting the training set based on the selected test into two partitions, one for which the test succeeds and one for which the test fails, and then calls itself recursively on each partition to construct the left and right subtrees.

Algorithm 1 GML_DT

Input: an attribute-valued training set S
If stopping criterion is True then terminate
End if
For each attribute A do
For each split value v do
Compute overall entropy for splitting on (A, v)
End for
End for
$(A, v)_{best}$ = Best attribute-value that reduces the overall entropy
Create a node in the Tree with the best test $(A, v)_{best}$
S_1, S_2 = Induced sub-datasets from S based on the test $(A, v)_{best}$
Sub_Tree_1 = GML_DT(S_1)
Sub_Tree_2 = GML_DT(S_2)
Add Sub_Tree_1 and Sub_Tree_2 to the corresponding branches of the Tree
Output: Tree

The best attribute-value test is selected by considering all possible split values for each attribute. If the attribute is categorical, the algorithm constructs a test of the form $a_i = v_j$, if it is continuous the test takes the form $a_i \leq v_j$. For each node, the algorithm computes the heuristic values of all the possible attribute-value tests. The heuristic calculates the overall entropy induced by splitting the node on an attribute-value test. The overall entropy as defined in equation (2) is the sum of the weighted entropy of the two partitions created by the split according to their size.

The algorithm then selects the test that reduces this heuristic to put in an internal node. It splits the instances based on the test into two partitions and constructs the subtrees as explained above.

$$Overall\_Entopy = \sum_{i\in\{1,2\}} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

$$Entropy(S_i) = \frac{1}{n}\sum_{j=1}^{n} Entropy(S_i, \lambda_j) \quad (3)$$

Where $Entropy(S_i, \lambda_j)$ is defined as follows:

$$Entropy(S_i, \lambda_j) = -\sum_{k=1}^{m} p(\mu_k)\log p(\mu_k) \quad (4)$$

We modified the formula of the entropy in order to handle graded multi-label classification tasks. We propose a graded multi-label entropy that is computed as the averaged sum of the entropies of class labels. This definition ensures that instances with similar degrees of relevance to the set of labels go in the same subset and thus allowing the prediction of a set of membership degrees for the set of labels in the leaves. We use the majority vote on each class label, to predict its corresponding relevance degree.

The algorithm builds the tree until a stopping criterion is triggered. The stopping conditions are:

- The partition is pure, meaning that all instances have the same degree of relevance for each class label.
- The number of instances in the node are less than a predefined threshold.
- The tree reaches a maximum depth.

C. A Toy Example

To demonstrate the process of building a graded multi-label decision tree, we use a toy example. Table I displays 10 samples of the toy dataset, which originally contains a total of 35 instances described with 3 attributes. a_1 is a categorical feature, a_2 and a_3 are continuous features. The set of class labels is constituted by $\{c_1, c_2, c_3\}$ and the set of degrees is $\{0, 1, 2, 3\}$. This set is equivalent to a set of descriptive nominal counterpart for each degree {not at all, somewhat, almost, fully} characterizing the ordinal levels of relevance of the class labels.

TABLE I. GRADE MULTI-LABEL TOY DATASET

Instances	a_1	a_2	a_3	c_1	c_2	c_3
x_1	A	63	7	3	0	0
x_2	C	29	5	3	3	1
x_3	A	69	13	1	2	3
x_4	B	49	11	1	0	2
x_5	C	51	7	2	1	0
x_6	B	61	5	1	0	2
x_7	C	43	17	2	1	0
x_8	A	69	9	3	0	0
x_9	C	27	13	3	3	1
x_{10}	B	10	14	1	0	2

First, we find the best split by iterating over the attribute columns to get potential split values. The potential split values of a continuous attribute column being the middle values between each consecutive values of the attribute. Then we calculate the overall entropy induced by splitting on an attribute a and a potential split value v.

For example, according the samples in table 1, the potential splits for the attribute a_1 are $(a_1 = A)$, $(a_1 = B)$ and $(a_1 = C)$

Based on these samples, the overall entropy induced by the test $(a_1 = A)$ is calculated as follows:

By applying Equations (2) and (3), we obtain:

$$Overall\_Entropy = \frac{3}{10} \times Entropy(S_1) + \frac{7}{10} \times Entropy(S_2)$$

$$Entropy(S_1) = \frac{1}{3} \times [Entropy(S_1, c_1) + Entropy(S_1, c_2) + Entropy(S_1, c_3)]$$

Where S_1 is the set of instances for which $a_1 = A$, and S_2 is the set of instances for which $a_1 \neq A$.

By applying Equation (4), the entropy of the subset S_1 for the class label c_1 is calculated as follows:

$$Entropy(S_1, c_1) = -\,p(0) \times \log p(0) - p(1) \times \log p(1) - p(2) \times \log p(2) - p(3) \times \log p(3)$$

$$= 0 - \frac{1}{3} \times \log\frac{1}{3} - 0 - \frac{2}{3} \times \log\frac{2}{3}$$

$$= 0.918$$

After computing $Entropy(S_1, c_2)$ and $Entropy(S_1, c_3)$ following the same process, we get the entropy of the subset S_1:

$$Entropy(S_1) = 0.918$$

In the same way, we calculate the entropy of the subset S_2 and we obtain:

$$Entropy(S_2) = 1.556$$

The overall entropy for splitting on the test $(a_1 = A)$ is 1.364

The overall entropies for the remaining potential splits of the attributes a_1, a_2 and a_3 can be computed following the same process. The potential tests for the continuous features are determined by considering the middle values between each consecutive values.

Fig. 1 displays the decision tree built by GML_DT based on the toy dataset (35 instances). We can infer the five following rules from this decision tree:

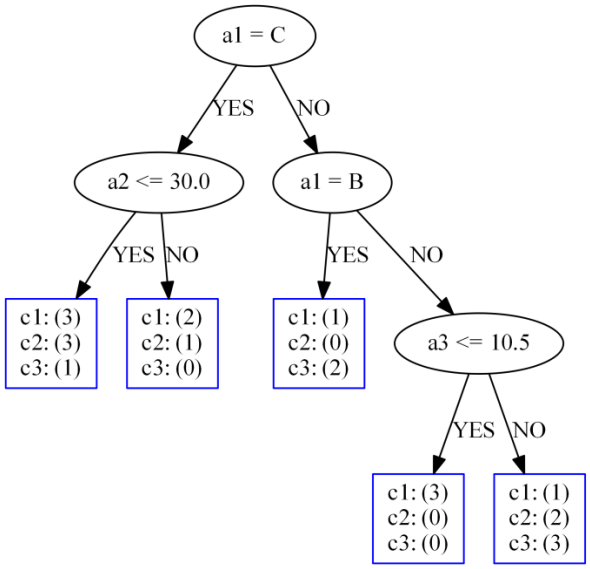


Fig. 1. A Graded Multi-label Decision Tree Constructed by GML_DT.

R_1: IF $a_1 = C$ AND $a_2 \leq 30$ THEN $<Degree(c_1) = 3, Degree(c_2) = 3, Degree(c_3) = 1>$

R_2: IF $a_1 = C$ AND $a_2 > 30$ THEN $<Degree(c_1) = 2, Degree(c_2) = 1, Degree(c_3) = 0>$

R_3: IF $a_1 = B$ THEN $<Degree(c_1) = 1, Degree(c_2) = 0, Degree(c_3) = 2>$

R_4: IF $a_1 \notin \{C, B\}$ AND $a_3 \leq 10.5$ THEN $<Degree(c_1) = 3, Degree(c_2) = 0, Degree(c_3) = 0>$

In this toy dataset the attribute a_1 has three possible values A, B or C. Hence, the condition $a_1 \notin \{C, B\}$ is equivalent to $a_1 = A$. Therefore, the fourth rule becomes:

R_4: IF $a_1 = A$ AND $a_3 \leq 10.5$ THEN $<Degree(c_1) = 3, Degree(c_2) = 0, Degree(c_3) = 0>$

R_5: IF $a_1 = A$ AND $a_3 > 10.5$ THEN $<Degree(c_1) = 1, Degree(c_2) = 2, Degree(c_3) = 3>$

As demonstrated above, the new developed graded multi-label model yields rules that are intelligible and interpretable.

IV. Experimental Study

We conducted an experimental study on real-world datasets, comparing our approach with binary relevance (BR) applied with a state of the art decision tree classifier, under the evaluation metrics from the literature [12] [1]. Cheng et al. [1] generalized some of the common loss functions used in multi-label classification for the graded multi-label setting and introduced the benchmark dataset BeLaE. These performance metrics were then used in the experimental study in [12], which compares the previous work in [1] and three new implemented approaches. This study was carried out on the benchmark dataset BeLaE and two additional real-world graded multi-label datasets curated by the authors, which makes it the most extensive work on GMLC compared to the rest of the work in the literature. For the purpose of conformity, we used the same datasets in our experimental study, along with three of the evaluation metrics from this previous work.

The experimental study is obtained by carrying out 10-fold cross validation on each single dataset. The same folds were used for both experiments on GML_DT and a baseline decision tree classifier applied with binary relevance approach (BR_DT).

We developed the GML_DT algorithm proposed in this paper from scratch using Python. The baseline BR_DT is implemented using the Scikit-learn library [23].

A. Evaluation Metrics

We evaluated the predictive performance of the algorithm based on three metrics; the hamming loss, which corresponds to the mean deviation of the predicted membership degrees to the true membership grades:

$$E_H(\hat{y}_x, y_x) = \frac{\sum_{i=1}^{n} AE(\hat{y}_x^i, y_x^i)}{(m-1)n} \quad (5)$$

Where AE is the absolute error of the predicted membership degree and it is defined as:

$$AE: M \times M \to L, AE(\mu_i, \mu_j) = |i - j|$$

The vertical 0-1 loss measures the percentage of class labels with incorrectly predicted degrees of relevance:

$$E_{0/1}(\hat{y}_x, y_x) = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_x^i \neq y_x^i) \quad (6)$$

Where I is the indicator function.

The C-index measures the pairwise ranking errors between the true membership set and the predicted membership set.

$$E_{CI} = \frac{\sum_{i<j}\sum_{(\lambda,\lambda') \in M_i \times M_j} S(h_x(\lambda), h_x(\lambda'))}{\sum_{i<j} |M_i| \times |M_j|} \quad (7)$$

Where $M_i = \{\lambda \in L | L_x(\lambda) = \mu_i\}$

$L_x(\lambda)$ is a function returning the degree of membership of the label λ for an instance x.

$h_x(\lambda)$ is the predicted degree of membership of the label λ for an instance x.

and $S(u, v) = I(u > v) + \frac{1}{2} I(u = v)$

B. Datasets

The datasets used for the experimental study are the BeLa-E which is a benchmark dataset introduced in [1] consisting of 100 variants, 50 datasets predicting 5 labels and 50 datasets predicting 10 labels. BeLaE was constructed based on a survey conducted with 1930 students, in order to grade on a finite ordinal scale the importance of different job properties.

Movies [12] is a dataset collected from 1967 movies where each movie is graded on its level of membership to five descriptive categories, e.g. its level of humor, action, suspense.

Medical [12], based on 1953 radiology reports, annotated with a set of ICD-9-CM disease/diagnosis classification codes. This dataset was adapted from the multi-label dataset by taking into account the level of agreement of the annotators.

Table II summarized the different properties of the aforementioned datasets.

C. Results

Table III displays the experimental results of GML_DT in comparison to BR_DT. We averaged the evaluations across the 10-fold cross validation for each single dataset. For the benchmark datasets, BeLaE (n=5) and BeLaE (n=10), we averaged the performance over the 50 variants of each. We summarize their predictive measures in terms of the mean and the standard deviation.

The experimental study conducted reaches the following conclusions:

GML-DT outperforms the baseline classifier BR_DT in terms of predictive performance according to all three evaluation metrics used in this experiment for all four datasets.

The performance metrics used in these experiments evaluate the results along three dimensions depicting the disparity, accuracy and pairwise ranking errors between the true membership set and the predicted membership set. Hence, GML_DT yields more accurate rules across all these different dimensions.

TABLE II. OVERVIEW OF THE GRADED MULTI-LABEL DATASETS, AND THEIR PROPERTIES: NUMBER OF INSTANCES, NUMBER OF ATTRIBUTES, NUMBER OF CLASS LABELS AND THE NUMBER OF GRADES

Datasets	Instances	Attributes	Labels	Grades
BeLaE n=5	1930	45	5	5
BeLaE n=10	1930	40	10	5
Movies	1967	27002	5	4
Medical	1953	1602	204	4

TABLE III. EXPERIMENTAL RESULTS FOR EACH DATASET ACCORDING TO THE HAMMING LOSS, THE VERTICAL 0/1 LOSS AND THE C-INDEX

Datasets	Evaluation Measures	GML_DT	BR_DT
BeLa-E n=5	Hamming Loss	0.168 ±0.018	0.257 ±0.026
	Vertical 0-1 Loss	0.526 ±0.039	0.689 ±0.032
	C-Index	0.264 ±0.047	0.374 ±0.055
BeLa-E n=10	Hamming Loss	0.174 ±0.011	0.259 ±0.017
	Vertical 0-1 Loss	0.540 ±0.020	0.690 ±0.023
	C-Index	0.263 ±0.030	0.361 ±0.040
Movies	Hamming Loss	0.172	0.253
	Vertical 0-1 Loss	0.424	0.536
	C-Index	0.247	0.368
Medical	Hamming Loss	0.002	0.010
	Vertical 0-1 Loss	0.006	0.017
	C-Index	0.135	0.448

Furthermore, GML_DT has a smaller model size compared to BR_DT. In fact, GML_DT is a single model that predicts the membership degrees relative to the set of class labels simultaneously. It builds a single decision tree that identifies the attribute-value conditions relevant for the prediction of the complete set of degrees associated to the label set. On the other hand, BR_DT runs $|L|$ times, and results in $|L|$ constructed decision trees, one for each class label, which not only affects its execution time and complexity but also its interpretability. The higher the number of labels, the more complicated the model gets and therefore the less effective it becomes for retrieving useful and comprehensible rules.

Moreover, if we were to compare GML_DT to the state of the art approaches solely based on the results reported in [1] and [12], we can deduce that GML_DT outperforms the model in [1] across all three metrics. Furthermore, it has better results for the hamming loss and the vertical 0-1 loss compared to the full CLR and Joined CLR while it remains very competitive against the Horizontal CLR [12].

V. CONCLUSION

We present a graded multi-label decision tree classifier, GML_DT, which generalizes the multi-label setting by predicting the membership degrees of the target labels instead of the binary relevance/non relevance. This approach utilizes the interpretability of decision tree classifiers and produces readable and comprehensive trees, which can be translated into useful, homogenous rules. GML_DT is also the first adaptation algorithm in the literature that fully models label dependencies, resulting in an increase of the predictive performance and the quality of the deduced rules.

The proposed algorithm is based on a new adapted graded multi-label heuristic that allows the algorithm to split based on the homogeneity of the combined set of labels, and ultimately retuning a vector containing the majority grade in each class label. We carried out an experimental study on real-world graded multi-label datasets, and evaluated our approach against a state of the art transformation-based decision tree classifier.

Our model is more interpretable and has the best predictive quality according to a variety of performance measures from the GMLC literature.

This paper constitutes a preliminary presentation of GML_DT, we are currently investigating further adaptations of the heuristic to the GML setting in order to improve the predictive performance of the model. Moreover, we are working on reducing the complexity of the generated tree via an adapted post pruning method.

REFERENCES

[1] CHENG, Weiwei, DEMBCZYNSKI, Krzysztof, et HÜLLERMEIER, Eyke. Graded multilabel classification: The ordinal case. In : ICML. 2010.

[2] CHEN, Guibin, YE, Deheng, XING, Zhenchang, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In : 2017 international joint conference on neural networks (IJCNN). IEEE, 2017. p. 2377-2383.

[3] LIU, Yang, WEN, Kaiwen, GAO, Quanxue, et al. SVM based multi-label learning with missing labels for image annotation. Pattern Recognition, 2018, vol. 78, p. 307-317.

[4] ZHU, Feng, LI, Hongsheng, OUYANG, Wanli, et al. Learning spatial regularization with image-level supervisions for multi-label image classification. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 5513-5522.

[5] MARKATOPOULOU, Foteini, MEZARIS, Vasileios, et PATRAS, Ioannis. Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. IEEE transactions on circuits and systems for video technology, 2018, vol. 29, no 6, p. 1631-1644.

[6] PRABHU, Yashoteja, KAG, Anil, GOPINATH, Shilpa, et al. Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation. In : Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018. p. 441-449.

[7] ZHANG, Zheng, ZOU, Qin, LIN, Yuewei, et al. Improved deep hashing with soft pairwise similarity for multi-label image retrieval. IEEE Transactions on Multimedia, 2019, vol. 22, no 2, p. 540-553.

[8] BUSTOS, Aurelia, PERTUSA, Antonio, SALINAS, Jose-Maria, et al. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 2020, vol. 66, p. 101797.

[9] ZHOU, Jian-Peng, CHEN, Lei, et GUO, Zi-Han. iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. Bioinformatics, 2020, vol. 36, no 5, p. 1391-1396.

[10] ZHANG, Jingpu, ZHANG, Zuping, WANG, Zixiang, et al. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. Bioinformatics, 2018, vol. 34, no 10, p. 1750-1757.

[11] ZHOU, Zhi-Hua et ZHANG, Min-Ling. Multi-label Learning. 2017.

[12] BRINKER, Christian, MENCÍA, Eneldo Loza, et FÜRNKRANZ, Johannes. Graded multilabel classification by pairwise comparisons. In : 2014 IEEE International Conference on Data Mining. IEEE, 2014. p. 731-736.

[13] LASTRA, Gerardo, LUACES, Oscar, et BAHAMONDE, Antonio. Interval prediction for graded multi-label classification. Pattern Recognition Letters, 2014, vol. 49, p. 171-176.

[14] LAGHMARI, Khalil, MARSALA, Christophe, et RAMDANI, Mohammed. Learning Label Dependency and Label Preference Relations in Graded Multi-label Classification. Computational Intelligence for Pattern Recognition, 2018, p. 115-164.

[15] TSOUMAKAS, Grigorios, KATAKIS, Ioannis, et VLAHAVAS, Ioannis. Mining multi-label data. In : Data mining and knowledge discovery handbook. Springer, Boston, MA, 2009. p. 667-685.

[16] TSOUMAKAS, Grigorios et KATAKIS, Ioannis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 2007, vol. 3, no 3, p. 1-13.

[17] FÜRNKRANZ, Johannes, HÜLLERMEIER, Eyke, MENCÍA, Eneldo Loza, et al. Multilabel classification via calibrated label ranking. Machine learning, 2008, vol. 73, no 2, p. 133-153.

[18] READ, Jesse, PFAHRINGER, Bernhard, HOLMES, Geoff, et al. Classifier chains for multi-label classification. Machine learning, 2011, vol. 85, no 3, p. 333-359.

[19] CLARE, Amanda et KING, Ross D. Knowledge discovery in multi-label phenotype data. In : European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2001. p. 42-53.

[20] BLOCKEEL, Hendrik, SCHIETGAT, Leander, STRUYF, Jan, et al. Decision trees for hierarchical multilabel classification: A case study in functional genomics. In : European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2006. p. 18-29.

[21] BLOCKEEL, Hendrik, DE RAEDT, Luc, et RAMON, Jan. Top-down induction of clustering trees. arXiv preprint cs/0011032, 2000.

[22] VENS, Celine, STRUYF, Jan, SCHIETGAT, Leander, et al. Decision trees for hierarchical multi-label classification. Machine learning, 2008, vol. 73, no 2, p. 185.

[23] PEDREGOSA, Fabian, VAROQUAUX, Gaël, GRAMFORT, Alexandre, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011, vol. 12, p. 2825-2830.

A Recognition Method for Cassava Phytoplasma Disease (CPD) Real-Time Detection based on Transfer Learning Neural Networks

Irma T. Plata[1], Edward B. Panganiban[2]
Darios B. Alado[3]
Faculty, CCSICT, Isabela State University
San Fabian, Echague, Isabela, Philippines

Allan C. Taracatac[4], Bryan B. Bartolome[5]
Freddie Rick E. Labuanan[6]
Technical Staff, CCSICT, Isabela State University
San Fabian, Echague, Isabela, Philippines

***Abstract*—Object detection technology aims to detect the target objects with the theories and methods of image processing and pattern recognition, determine the semantic categories of these objects, and mark the specific position of the target object in the image. This study generally aims to establish a recognition method for Cassava Phytoplasma Disease (CPD) real-time detection based on transfer learning neural networks. Several methods and procedures were conducted, such as the testing of two methods in transmitting long-distance high definition (HD) video capture; establishment of a compact setup for a long-range wireless video transmission system; the development, testing of the real-time CPD detection and quantification monitoring system, providing the comparative performance analysis of the three models used. We have successfully custom-trained three artificial neural networks using transfer learning: Faster Regions with Convolutional Neural Networks (R-CNN) Inception v2, Single Shot Detector (SSD) Mobilenet v2, and You Only Look Once (YOLO) v4. These deep learning models can detect and recognize CPD in actual environment settings. Overall, the developed real-time CPD detection and quantification monitoring system was successfully integrated into the wireless video receiver and seamlessly visualized all the incoming data using the three different CNN models. If the consideration is the image processing speed, YOLOv4 is better compared to other models. But, if accuracy is the priority, Faster R-CNN inception v2 performs better. However, since CPD detection is the main purpose of this study, the Faster R-CNN model is recommended for adoption to detect CPD in a real-time environment.**

***Keywords*—Cassava phytoplasma disease; faster regions with convolutional neural networks (R-CNN) inception v2; you only look once (YOLO) v4; object detection; precision agriculture**

I. Introduction

The agricultural industry plays an important role in the economy. Plant illness is also caused by climatic circumstances, exacerbated by the exponential trend of population growth. The major challenges of sustainable development include reducing pesticide use, the expense of preserving the environment, and the cost of building quality. Exact, exact, and timely decisions may reduce pesticide use [1]. Cassava is a key crop that has been produced in the Philippines. Nowadays, innovation is commonly used for plant disease prediction. The concept of image processing combined with information mining improvements aids in identifying plant diseases. With the development of intelligent devices, the data bulk on Internet has grown with high speed. As an important aspect of image processing [2][3][4], object detection has become one of the popular international research fields. In recent years, the powerful ability with feature learning and transfer learning of Convolutional Neural Network (CNN) has received growing interest within the computer vision community, thus making a series of important breakthroughs in object detection [5][6]. So, it is a significant application to apply CNN to object detection for better performance.

Real-time object detection is the task of doing object detection in real-time with fast inference while maintaining a base level of accuracy [7]. Object detection technology aims to detect the target objects with the theories and methods of image processing and pattern recognition [8][9], determine the semantic categories of these objects, and mark the specific position of the target object in the image. It is a very challenging task in the actual application to use computer technology to detect objects automatically. Complex background, noise disturbance, occlusion, low-resolution, scale, and attitude changes, and other factors will seriously affect the object detection performance. The conventional object detection method was based on the hand-crafted feature. It is not robust to illumination change, lacking good generalization ability. Using Google's TensorFlow and YOLOv4 as well as transfer learning, we were able to custom-trained three separate artificial neural networks [10]. These deep learning models can detect and recognize CPD in actual environment settings.

This study established a recognition method for Cassava Phytoplasma Disease (CPD) real-time detection based on transfer learning neural networks. Specifically, this paper is presented to test the two methods in transmitting long-distance high definition (HD) video capture. Likewise, it established a compact setup for a long-range wireless video transmission system. Furthermore, the project developed a real-time CPD detection and quantification monitoring system, performed a wireless video transmission test, and compared the test results in each model used.

As compared with the recent papers and technologies related to plant disease detection, this paper presented a disease detection technology specifically for Cassava

Phytoplasma Disease. It demonstrated which real-time detection technology is the most appropriate to use, which is more practical and can have a great impact on society. This research provides a visual object identification framework capable of processing pictures at high detection rates while processing images at a fast pace.

Hence, the authors conducted a thorough experiment on what innovative technology will be used. To befittingly implement the selected technology to our partner agency, the EDCOR, a cassava development cooperative located in San Guillermo, Isabela. Their cassava farms are situated in remote locations where internet connectivity is scarce. We developed a remote image processing server system utilizing digital First-Person View (FPV) [11]. This remote server can perform real-time image processing for CPD detection and recognition with a dedicated ground station for long-range high-definition video transmission without using the Internet.

The next sections discussed the related works, methods, and results that transpired during the conduct of the study. It explained the details on how the entire project was done and the results of the undertakings.

II. Literature Review

Cassava plants exhibiting characteristic phytoplasma signs, such as witches' broom, general stunt, chlorosis, distortion, and decreased size [12], were seen in farms near San Guillermo, Isabela. CPD is a severe danger to cassava producers' food security [13].

Object detection is a critical computer vision problem that involves detecting visual objects in digital photos of a specific type (such as humans, animals, or autos). Object detection aims to create computational models and approaches that give one of the most fundamental bits of data required by computer vision applications [14]. Object detection is essential in identifying the disease type in the specific video sequence performed in agricultural farming. Object detection reduces computing time while increasing detection accuracy [15].

Several popular deep learning-based object identification algorithms are available that can accurately identify which section of a farm field is infected with Cassava Phytoplasma Disease. Some frameworks will need a lot of computing power, while some will provide less accuracy. By surveying several neural network frameworks, the authors identified the following frameworks with the best performance in object detecting technology.

Transfer learning is the enhancement of learning in a new task by transferring knowledge from a previously learned related activity [16]. While most machine learning algorithms are designed to solve particular tasks, the development of algorithms that promote transfer learning is a continuing area of study in the machine-learning field. The following are the transfer learning neural networks considered in this paper:

A. Faster Regions with Convolutional Neural Networks (R-CNN) Inception v2

The Region Proposal Network (RPN) is a fully convolutional network that predicts object limits and objectness scores at each place simultaneously. RPNs are trained from start to finish to create high-quality region suggestions, which Fast R-CNN uses for detection [17]. RPN and Fast R-CNN may be taught to share convolutional features using a simple alternating optimization. RPN was added to the design of faster R-CNN. It means that it uses a quick neural network to handle a sluggish search selection process.

RPN comes after the last convolution layer of a Convolutional Neural Network. Inception V2 was a module created to minimize the complexity of a convolution network. This module causes the convolution network to be broader rather than deeper. There are three types of modules in Inception V2 [18].

B. You Only Look Once (YOLO) v4

YOLO is a cutting-edge, real-time object detecting technology. It is a real-time object identification system that recognizes several items in a single frame. YOLO takes a completely new approach to earlier detection methods. It uses a single neural network to process the entire picture. This network separates the picture into regions and predicts each bounding boxes and probabilities. The projected probabilities are used to weigh these bounding boxes [19]. YOLO offers excellent real-time performance in multi-scale object recognition [20].

C. Single Shot Detection (SSD)

Single Shot Detection is a technique for detecting objects in a single shot. By capturing a single photograph, you can analyze many items. A single frame is used to recognize and analyze several items in a picture. Compared to Convoluted Neural Networks, this is a considerably faster analysis. For p channel analysis, a feature layer of m*n is obtained. For each of the k areas, a bounding box is generated. SSD is sometimes called a Multibox detector since it computes each bounding box and offsets relative to the initial bounding boxes [21].

III. Methods

A. Conceptual Framework

The different processes and methodologies in the wireless video transmission system, its establishment, the development of a real-time CPD detection monitoring system, the performance testing were realized based on the conceptual framework of the study as presented in Fig. 1. The Real-time CPD detection and quantification Framework Model.

B. Methods in Transmitting Long-distance High Definition Video Capture using the Long-Range Wireless Video Transmission System

We have used two methods in transmitting long-distance high definition (HD) video capture. These methods involve using 2.4Ghz and 5Ghz frequencies. Both had their advantages and disadvantages. 2.4GHz offers a wide network coverage and is superior at bypassing substantial impediments like trees. However, it has a smaller data rate and is more susceptible to interference; more devices normally utilize this frequency. 5Ghz, on the other hand, offers a greater data rate and is less susceptible to interference; it is typically used by

fewer devices. However, it has a smaller coverage area and is not performing well when penetrating solid/thick layered obstacles.

To implement the 2.4Ghz concept, we adapted the method called Wi-Fi Broadcast illustrated in Fig. 2. Wi-Fi-broadcast activates the monitor mode on the Wi-Fi cards. This mode allows you to transmit and receive arbitrary packets without having to associate them with anything. Furthermore, it is possible to get incorrect frames (where the checksum does not match). In this method, a real unidirectional connection is produced, simulating the benefits of an analog link. They comprise the following: the transmitter broadcasts data regardless of whether or not there are any receivers nearby. As a result, there's no possibility of video stalling due to a loss of association; the receiver gets video as long as it's within range of the transmitter. The video quality declines as it moves out of range, but it does not stall. Even if frames are incorrect, they will be shown rather than discarded; the classic "one broadcaster – numerous receivers" method will function right out of the box. Bystanders who wish to view the video stream on their devices only need to "shift to the correct channels"; Wi-Fi-broadcast permits the simultaneous use of numerous low-cost receivers and the combining of their data to enhance the likelihood of correct data reception. This so-called software diversity enables the use of identity and complementary receivers to increase dependability (imagine one reception with a 360° omnidirectional antenna and multiple directional antennas for long distances, all functioning in tandem). To archive high dependability at minimal bandwidth needs, Wi-Fi-broadcast employs Forward Error Correction. It can recover packets that have been lost or corrupted at the receiver.

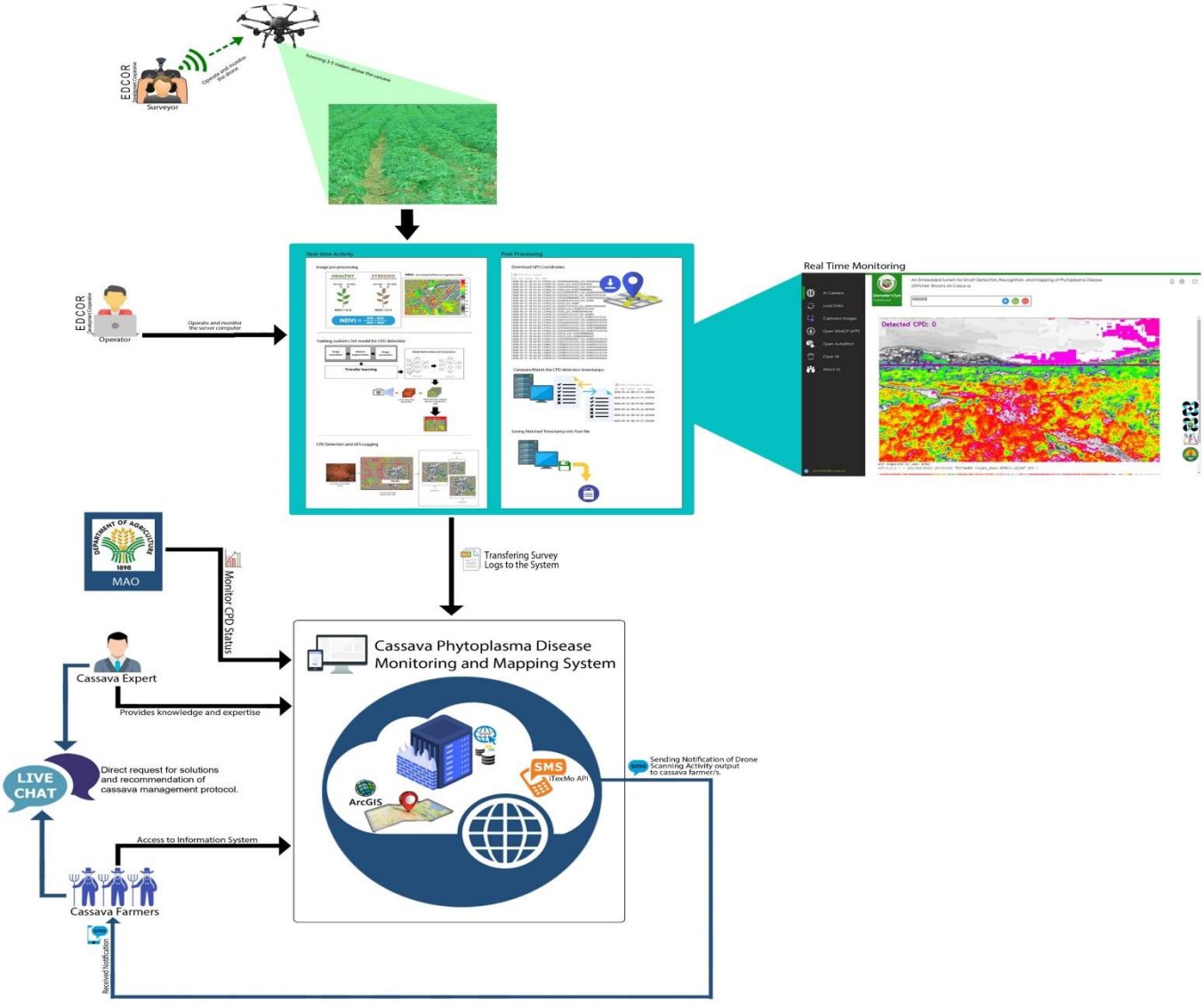


Fig. 1. The Structural Framework of Faster R-CNN(a), YOLOv4(b), SSD Mobilenet v2(c).

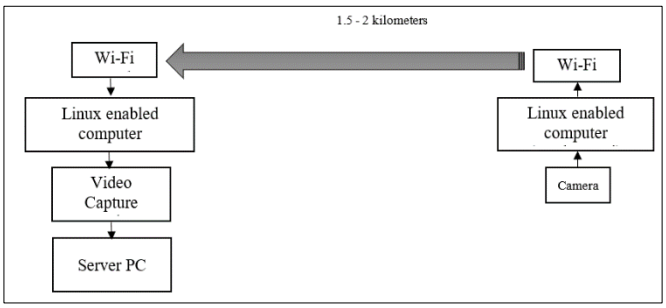


Fig. 2. Wi-Fi Broadcast Block Diagram.

Using standard Wi-Fi dongles with Atheros AR9271 802.11 chip for the transmitter and receiver, and a 2.4Ghz Yagi antenna for our ground station presented in Fig. 3 and 4. This setup can process camera feed from either USB Video or through RTSP/TCP URLs.

The ground station uses a sturdy tripod with a height of 2.5 meters. The Wi-Fi dongle, Yagi antenna, and a single board computer are securely installed on it, as shown in Fig. 5 and 6.

Fig. 3. Atheros AR9271 802.11 Wi-Fi Dongle.

Fig. 4. 2.4Ghz Yagi Antenna.

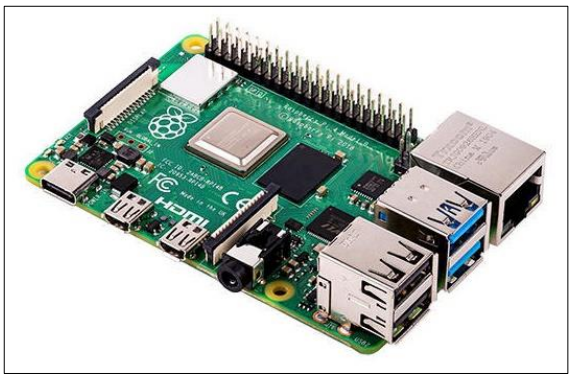


Fig. 5. Single Board Computer; Raspberry pi 4.

Fig. 6. A 2.5 Meters Tripod Stand where the Wi-Fi Dongle, Yagi.

On the other hand, we used a commercially available FPV high-definition video transmitter using 5Ghz presented in Fig. 7, which also supports HDMI all format video input up to 1920*1080@60fps, and output is up to 1920*1080@30fps. A 5dB omnidirectional antenna is installed on the ground unit (standard). According to the manufacturer, the effective transmission distance is over 2km with an output delay of 80ms/0.08s.

Fig. 7. Commercially Available full HD Digital Video System.

The receiver/ground unit is also installed on the tripod, just beside the 2.4Ghz Yagi antenna. This method takes advantage of the commercially available HD Video TX/RX and its dedicated windows application. The commercially available video TX/RX works perfectly fine using its dedicated application (Insight.exe) on a windows computer. However, the downside is that we cannot tap on the video feed from the transmitter into our developed system. Given that its dedicated application is not open-source and there is no available software development kit (SDK) for this device, it is hard to analyze and reverse engineer. On the other hand, the RTSP URL that supposedly works based on its documentation turns out to be incorrect (or, if not, maybe they no longer support this feature).

The alternative concept uses the "HD Video Capture" method, as illustrated in Fig. 8, which works by connecting the NDVI camera through HDMI into the Video TX and connected to the Video RX wirelessly (5Ghz band). The Video RX will also act as a Wi-Fi access point using the 5Ghz band. A lightweight, small single-board computer (SBC) running Windows 10 Pro x64 with Insight.exe configured will be linked to the Video RX over Wi-Fi (5Ghz) and display the live camera feed from the TX side. Then a video capture device will be used to mirror the preview display on the SBC and send it to our server computer via wired connection in digital HD resolution (720P - 1080P) in real-time (or near real-time).

After which, our developed Demeter's Eyes Monitoring System will then be able to adapt to this alternative method and process each frame from the incoming digital HD resolution camera pass through our custom-trained CNN model for CPD detection and quantification. These two different videos transmission configurations work independently of each other, and only one setup can be used at a time since there is only one camera installed to the transmitter.

C. Establishment of a Compact Setup for a Long-Range Wireless Video Transmission System

The long-range wireless video transmission system is equipped with 16 megapixels NDVI camera, GPS module, a single-board computer, and a swappable video transmitter (2.4Ghz and 5Ghz), as shown in Fig. 9. This compact setup also includes a portable power supply, mini-fan, and LED light indicators. The LED lights indicate the status of each component on this embedded system. It is very useful when conducting pre-flight checkups and during troubleshooting.

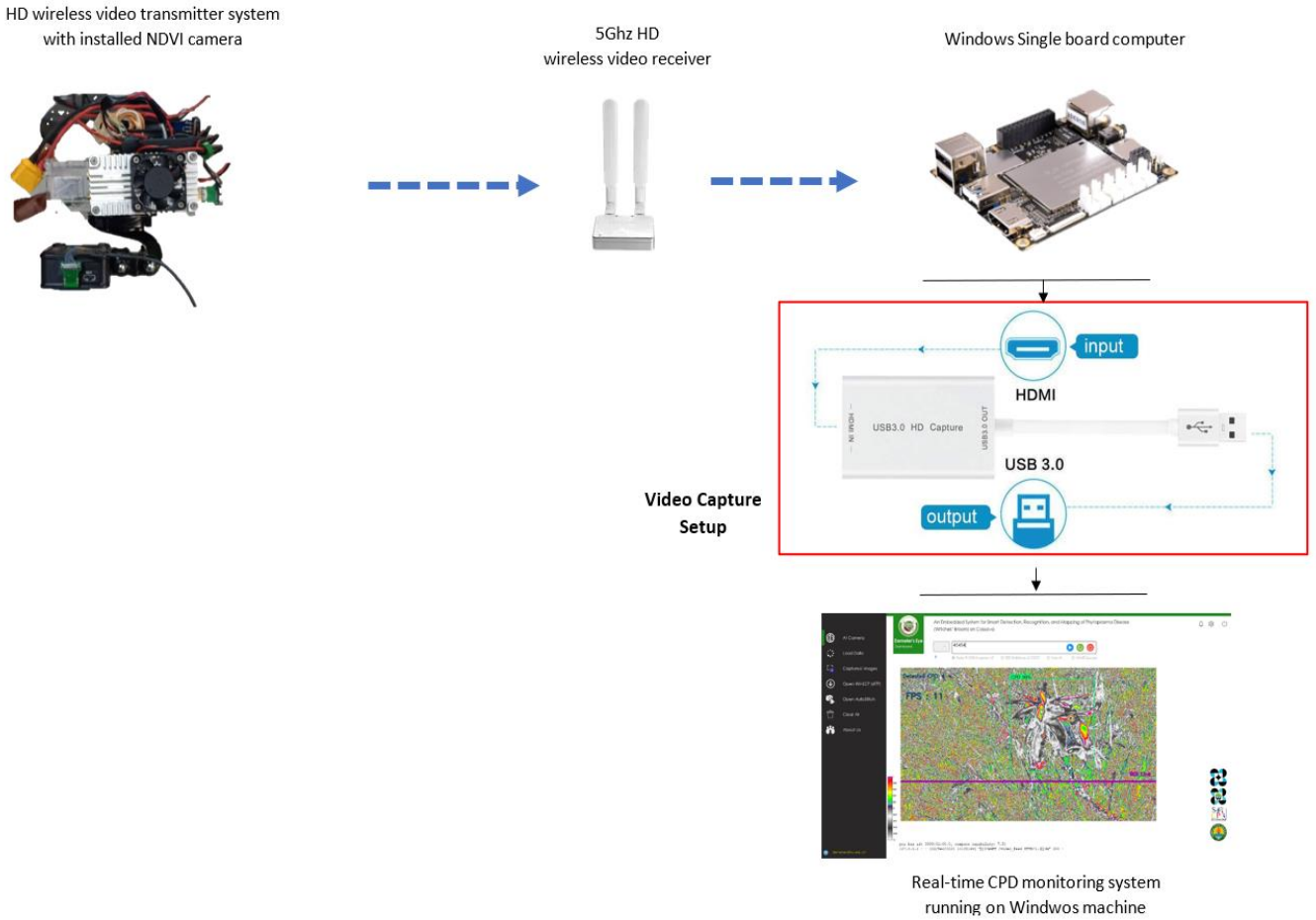


Fig. 8. Video Capture System Layout.

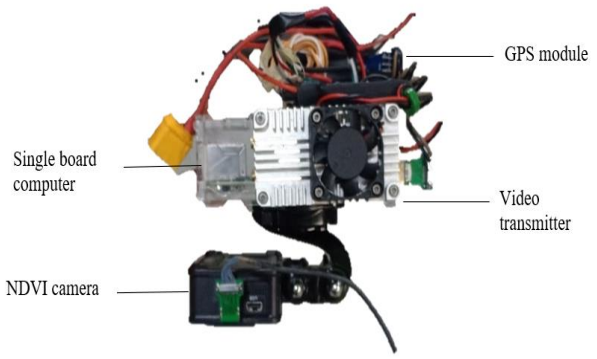


Fig. 9. HD Wireless Video Transmitter System with Installed NDVI Camera and GPS Module.

D. Development and Testing of the Real-Time CPD Detection and Quantification Monitoring System

The real-time CPD detection [22][23] and quantification system work by switching on the NDVI camera, video transmitter and receiver (including the single-board computer from the ground station), GPS logging module (will automatically start logging time stamps and GPS coordinates once switched on and successfully linked to a satellite) and, the drone itself. After which, the video transmitter and receiver will automatically link each other (the green light indicator for data on the video receiver must be steady).

On the server-side, the CPD real-time monitoring system should now be able to detect and preview incoming live camera feed from the video transmitter, as shown in Fig. 10.

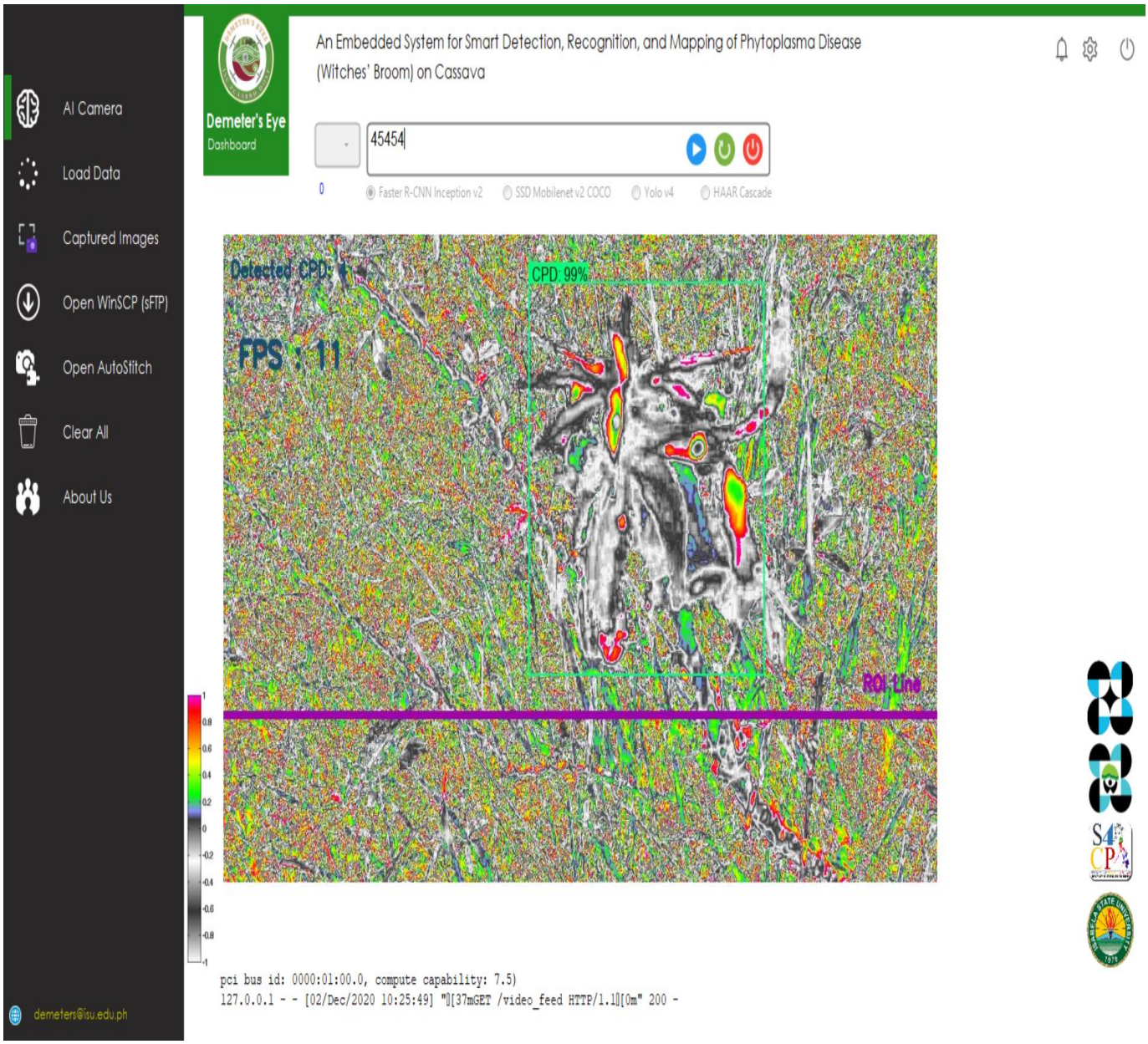


Fig. 10. Real-Time CPD Detection and Quantification System Dashboard.

Once the video link is established, the surveyor can now start hovering above the cassava field (3-5 meters above the cassava). While hovering over the cassava field, the CPD real-time monitoring system performs image pre-processing procedures on each incoming frame. This includes applying a Fastie color map to visualize the NDVI values in the image. The pre-processed frame(s) is passed through our CNN model (custom trained CPD detection model using transfer learning with Faster RCNN for prototype 2, SSD MobileNet V2 [1] COCO for prototype 3, and YOLO v4 for prototype 4) for CPD detection and quantification based on the TensorFlow object detection API. The real-time CPD detection is displayed on the system preview window with an overlaid running count of CPD detected. The system also automatically records each CPD detected timestamp (the latency between the transmitter and receiver was considered; thus, an adjustment with the time delay was done).

After completing the survey and the drone has returned, the GPS logging module will automatically connect to the access point located on the operator side. This will allow the operator to download GPS coordinates and corresponding timestamp Logs from the drone onboard GPS logging module through sFTP and upload it into the server computer. The system will now compare/match the CPD detection timestamps from the server computer and the timestamps from the GPS logging module. For each timestamp that matches, save its corresponding GPS coordinate into a .txt file. This .txt file will be uploaded into the web-based information system for visualization. These processes are illustrated in our conceptual framework shown in Fig. 1.

E. Testing the Long-range Wireless Video Transmission System

To verify the actual performance of the video transmission system, we conducted a stress test to determine the following: working time per battery charge of the NDVI camera; working time per battery charge of the video receiver; the actual working distance of the video TX/RX; actual latency of the video transmission; actual latency of the video transmission when adapted into the monitoring system.

We conducted the test at an open field inside the Isabela State University located at San Fabian, Echague, Isabela, as shown in Fig. 11 and 12. The test location is ideal since it has an almost identical topography with the actual cassava farms. It has a combination of wide-open fields with surrounding trees in various heights and densities.

F. Portable Handheld Prototype Field Testing

With favorable weather, we conducted the onsite pilot testing of our prototype at Villa Teresita, San Guillermo Isabela, on a cassava farm owned by one of the EDCOR members between 10:00 in the morning up until 3:00 in the afternoon. The planted cassava was at around three months old at the time of testing. We used PVC pipes to hold the prototype steady, facing the camera down at a height of 2-3 meters from the top-most part of the cassava. Thus, mimics the drone altitude and flight movement.

Fig. 13 shows the actual field testing of the prototype. This was performed in the cassava farm in Villa Sanchez, San Guillermo, Isabela. This testing makes sure that the prototype works well as per its functionalities.

G. Drone and Prototype Payload Field Testing

After the initial test, we embedded the prototype as a payload into our custom-built drone and went back to the same site to conduct another series of tests. This milestone is presented in Fig. 14 and 15.

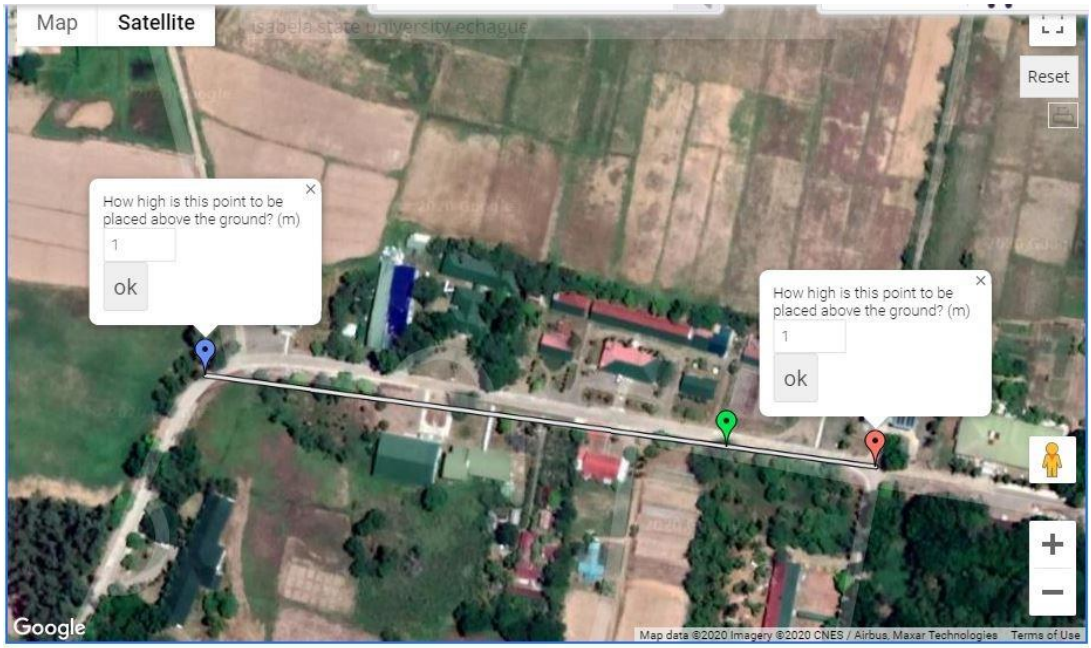


Fig. 11. Testing Site.

Fig. 12. Actual Wireless Video Transmission Test Conducted.

Fig. 13. Actual Field Testing of the Prototype.

Fig. 14. Installed Prototype as Payload underneath the Drone.

Fig. 15. Actual Field Testing using the Drone with the Prototype Payload on it

H. Comparison of the Test Results in each Model used

The ground-truthing [24] procedure was done on a selected area from a cassava farm with identified CPD infection. Our plant pathologist selected two lanes of planted cassava by which our prototypes camera can cover at one pass at a height of 2-2.5 meters. The cassava planted on the selected test area was counted manually. We got a total of 40 cassava stalks, wherein our plant pathologist identified 18 as CPD infected, and 22 were healthy/normal/or with a different disease. As the prototype passes through each cassava plant, our plant pathologist verifies if it can detect CPD infected or has no detection.

IV. Results and Discussion

For objective one, on testing two methods in transmitting long-distance high definition (HD) video capture. The wireless video transmission results show that the working time per battery charge of the NDVI camera (while installed in the video TX/RX) is capable of reaching three hours on preview mode with a remaining one bar on the battery indicator. Actual working distance of the Video TX/RX. Both 2.4Ghz and 5Ghz setups reached 900 meters in the actual ground test with few trees in between. Also, both setups lose connection after a total loss of line of sight between the TX and RX. The actual latency of the video transmission is at 90-100 milliseconds.

The second aim is to construct a small configuration for a long-range wireless video transmission system. The tests conducted between the two wireless configurations showed promising results with the given environmental condition during the field testing. However, we opt to use the 5Ghz setup moving forward simply because it is very straightforward to deploy both its transmitter and receiver. Also, given the fact that the end-users are not technically knowledgeable on setting up the 2.4Ghz configuration, which may be too complex for non-technical users, using it may pose problems during the turnover and maintenance of the system.

The creation and testing of a real-time CPD detection and quantification monitoring system and wireless video transmission test performance are goals 3 and 4. The developed real-time CPD detection and quantification monitoring system was successfully integrated into the wireless video receiver and seamlessly visualized all the incoming data using the three different CNN models, as shown in Fig. 16.

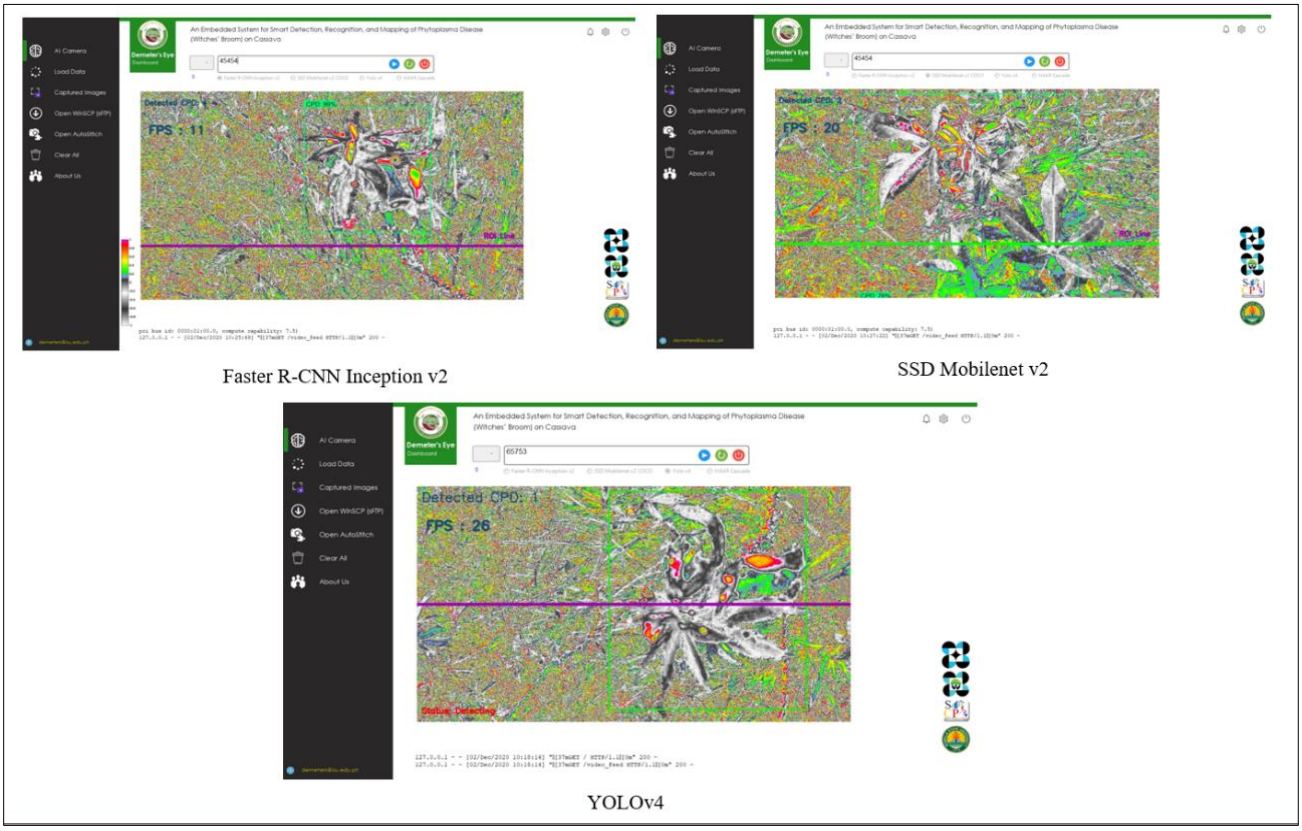


Fig. 16. Deploying the Three Custom Trained CNN Models (Faster R-CNN, SSD Mobilenet v2, and YOLOv4.

For objective 5, the comparison of the test results in each model used, the tabulated results using each custom-trained CNN model at different frame resolutions presented in Tables I to VI show that the Faster R-CNN model has the highest detection accuracy across different frame resolutions with a top detection rate of 13 out of 18 CPD samples. However, this has the lowest processing frame count among the three CNN models used. On the other hand, the YOLOv4 model has a top processing speed of 25 frames per second and a 9 out of 18 detection rate on CPD samples. In contrast, the SSD Mobilenet v2 model has a top processing frame count of 20 frames per second, which is second to the YOLOv4 model in terms of processing speed. However, it has the lowest detection rate, with just 6 out of 18 CPD samples being detected.

Table I shows the detection results using Faster R-CNN Inception v2. The video resolution was 1080 x 720 at eight frames per second (FPS). Negative samples mean that during the detection process of the system, these are samples that are not infected with CPD. In contrast, those positive samples are samples that were infected with CPD. This particular table showed that from the 25 negative samples, it recognized 20 True Negative (TN) from the samples while only 5 were False Negative (FN). Furthermore, the Faster R-CNN framework recognized 13 True Positive (TP) and 2 False Positive (FP) samples from 20 samples infected with CPD.

Table II is the detection tests using Faster R-CNN Inception v2 960 x 640 at 11FPS. The 24 negative samples recognized 17 as TN and seven as FN. From the 16 positive samples, the result is 11 as TP and five as FP.

Table III shows the result of tests using YOLOv4 1080 x 720 @22FPS. From this table, 18 were detected as TN while nine were detected as FN from the 27 negative samples. For the positive samples, the system saw nine as a TP, while 4 were FP.

Table IV also presented the results using YOLOv4 960 x 640 @25FPS. The result is not much desirable since from the 26 negative samples. It only detects 14 as TN, which is close to detection of FN, which is 12. The same is through with its detection on positive samples. It detects 8 FP and 7 FP from 15 positive samples.

Table V presents the results using SSD Mobilenet v2 1080 x 720 @18FPS. From this table, the detection accuracy can be seen as low. It has 13 TN and 12 FN from 25 negative samples, while 9 FP and 6 TP from 25 positive samples.

Table VI shows the result using SSD Mobilenet v2 960 x 640 @20FPS. Its accuracy is less good than the previous methods. From the 25 negative samples it detects 11 TN and 14 FN. While from the 15 positive samples, it detects 11 FP and 4 TP.

Based on these results, if the image processing speed is considered, YOLOv4 is better than other models. Faster R-CNN inception v2 performs better if accuracy is a priority. Hence, these two models can be used depending on the purpose of the detection of the CPD. However, the most important factor to be considered must be its accuracy since CPD detection is the main objective of this study.

TABLE I. FASTER R-CNN INCEPTION V2 1080 X 720 @8FPS

	Positive	Negative
Negative	FN=5	TN=20
Positive	TP=13	FP=2

TABLE II. FASTER R-CNN INCEPTION V2 960 X 640 @11FPS

	Positive	Negative
Negative	FN=7	TN=17
Positive	TP=11	FP=5

TABLE III. YOLOV4 1080 X 720 @22FPS

	Positive	Negative
Negative	FN=9	TN=18
Positive	TP=9	FP=4

TABLE IV. YOLOV4 960 X 640 @25FPS

	Positive	Negative
Negative	FN=12	TN=14
Positive	TP=7	FP=8

TABLE V. SSD MOBILENET V2 1080 X 720 @18FPS

	Positive	Negative
Negative	FN=12	TN=13
Positive	TP=6	FP=9

TABLE VI. SSD MOBILENET V2 960 X 640 @20FPS

	Positive	Negative
Negative	FN=14	TN=11
Positive	TP=4	FP=11

V. CONCLUSION

During the field testing/actual testing, the 5Ghz set-up was used because it is straightforward to deploy both its transmitter and receiver. The developed real-time CPD detection and quantification monitoring system was successfully integrated into the wireless video receiver and seamlessly visualized all the incoming data using the three different CNN models. If the consideration is the image processing speed, YOLOv4 is better compared to other models. Faster R-CNN inception v2 performs better if accuracy is a top requirement. Hence, these two models can be used depending on the purpose of the detection of the CPD. However, the most important factor to be considered must be its accuracy since CPD detection is the main objective of this study.

VI. RECOMMENDATION

The test's discussed results using the three methods - SSD Mobilenet v2, Faster R-CNN Inception v2, and YOLOv4, shows that Faster R-CNN Inception v2 has the highest accuracy. However, the accuracy rate needs to be improved to achieve optimal accuracy. The suggestion is to increase the training datasets and modify the hyperparameters to achieve maximum accuracy.

ACKNOWLEDGMENT

We want to extend our sincerest gratitude to the Department of Science and Technology (DOST) for funding this project under the Science for Change Program, particularly under the CRADLE PROGRAM. The Philippine Council for Agriculture, Aquatic and Natural Resources Research and Development (PCAARRD) as the monitoring body, particularly the Agricultural Resources Management Research Division (ARMRD). To the Isabela State University, Quirino State University, EDCOR Cassava Development Cooperative, and the LGU San Guillermo, Isabela.

REFERENCES

[1] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. K. Behera, "Image Processing Techniques for Diagnosing Rice Plant Disease: A Survey," Procedia Comput. Sci., vol. 167, no. 2019, pp. 516–530, 2020, doi: 10.1016/j.procs.2020.03.308.

[2] S. Jiddi, P. Robert, and E. Marchand, "Reflectance and Illumination Estimation for Realistic Augmentations of Real Scenes," Adjun. Proc. 2016 IEEE Int. Symp. Mix. Augment. Reality, ISMAR-Adjunct 2016, pp. 244–249, 2017, doi: 10.1109/ISMAR-Adjunct.2016.0085.

[3] A. Miyatra, D. Bosamiya, and N. Kamariya, "A Survey on Disease and Nutrient Deficiency Detection in Cotton Plant," Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 11, pp. 812–815, 2013.

[4] W. Zhu, H. Chen, I. Ciechanowska, and D. Spaner, "Application of infrared thermal imaging for the rapid diagnosis of crop disease," IFAC-PapersOnLine, vol. 51, no. 17, pp. 424–430, 2018, doi: 10.1016/j.ifacol.2018.08.184.

[5] S. Lange, T. Gabel, and M. Riedmiller, "Transfer Learning for Reinforcement Learning Domains: A Survey," J. of Machine Learn. Res., vol. 10, no. 2009, pp. 1633–1685, 2009, doi: 10.1007/978-3-642-27645-3_2.

[6] S. Panigrahi, A. Nanda, and T. Swarnkar, "A Survey on Transfer Learning," Smart Innov. Syst. Technol., vol. 194, pp. 781–789, 2021, doi: 10.1007/978-981-15-5971-6_83.

[7] D. Qiao and F. Zulkernine, "Vision-based Vehicle Detection and Distance Estimation," 2020 IEEE Symp. Ser. Comput. Intell. SSCI 2020, no. December, pp. 2836–2842, 2020, doi: 10.1109/SSCI47803.2020.9308364.

[8] Y. Lu, S. Yi, N. Zeng, Y. Liu, and Y. Zhang, "Identification of rice diseases using deep convolutional neural networks," Neurocomputing, vol. 267, pp. 378–384, 2017, doi: 10.1016/j.neucom.2017.06.023.

[9] H. Gassoumi, N. R. Prasad, and J. J. Ellington, "Neural Network-Based Approach For Insect Classification In Cotton Ecosystems," Int. Conf. Intell. Technol., no. January 2000, pp. 1–7, 1994.

[10] J. Lyu et al., "Extracting the tailings ponds from high spatial resolution remote sensing images by integrating a deep learning-based model," Remote Sens., vol. 13, no. 4, pp. 1–17, 2021, doi: 10.3390/rs13040743.

[11] A. C. Taracatac and R. Q. Camungao, "Rice insect classification and quantification (RICQ) using portable neural network model with EFPV-image processing algorithm," J. Adv. Res. Dyn. Control Syst., vol. 11, no. 11 Special Issue, pp. 979–985, 2019, doi: 10.5373/JARDCS/V11SP11/20193124.

[12] D. Flôres, I. C. Haas, M. C. Canale, and I. P. Bedendo, "Molecular identification of a 16SrIII-B phytoplasma associated with cassava witches' broom disease," Eur. J. Plant Pathol., vol. 137, no. 2, pp. 237–242, 2013, doi: 10.1007/s10658-013-0250-3.

[13] K. D. Kra, Y. M. N. Toualy, A. C. Kouamé, H. A. Diallo, and Y. A. Rosete, "First report of a phytoplasma affecting cassava orchards in Cote d'Ivoire.," New Dis. Reports, vol. 35, no. 1, pp. 21–21, 2017, doi: 10.5197/j.2044-0588.2017.035.021.

[14] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," arXiv Prepr. arXiv1905.05055, pp. 1–39, 2019, [Online]. Available: http://arxiv.org/abs/1905.05055.

[15] M. Xu, "Robust object detection with real-time fusion of multiview foreground silhouettes," Opt. Eng., vol. 51, no. 4, p. 047202, 2012, doi: 10.1117/1.oe.51.4.047202.

[16] C. C. Aggarwal, "Transfer Learning," Data Classif. Algorithms Appl., pp. 657–665, 2014, doi: 10.1201/b17320.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[18] D. Alamsyah and M. Fachrurrozi, "Faster R-CNN with inception v2 for fingertip detection in homogenous background image," J. Phys. Conf. Ser., vol. 1196, no. 1, 2019, doi: 10.1088/1742-6596/1196/1/012017.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, [Online]. Available: http://arxiv.org/abs/2004.10934.

[20] J. Yu and W. Zhang, "Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4," Sensors, vol. 21, no. 3263, pp. 1–21, 2021.

[21] R. Deepa, E. Tamilselvan, E. S. Abrar, and S. Sampath, "Comparison of Yolo, SSD, Faster RCNN for Real Time Tennis Ball Tracking for Action Decision Networks," Proc. 2019 Int. Conf. Adv. Comput. Commun. Eng. ICACCE 2019, pp. 2019–2022, 2019, doi: 10.1109/ICACCE46606.2019.9079965.

[22] I. T. Plata, A. C. Taracatac, and E. B. Panganiban, "Development and testing of embedded system for smart detection and recognition of witches' broom disease on cassava plants using enhanced viola-jones and template matching algorithm," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 5, pp. 2613–2621, 2019, doi: 10.30534/ijatcse/2019/113852019.

[23] I. T. Plata, E. B. Panganiban, B. B. Bartolome, F. E. R. Labuanan, and A. C. Taracatac, "A concept of cassava phytoplasma disease monitoring and mapping system using GIS and SMS technology," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 6, pp. 3357–3361, 2019, doi: 10.30534/ijatcse/2019/108862019.

[24] M. G. Selvaraj et al., "Machine learning for high-throughput field phenotyping and image processing provides insight into the association of above and below-ground traits in cassava (Manihot esculenta Crantz)," Plant Methods, vol. 16, no. 1, pp. 1–26, 2020, doi: 10.1186/s13007-020-00625-1.

Optimizing Smartphone Recommendation System through Adaptation of Genetic Algorithm and Progressive Web Application

Khyrina Airin Fariza Abu Samah[1], Nursalsabiela Affendy Azam[2]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Melaka Kampus
Jasin, Melaka, Malaysia

Raseeda Hamzah[3]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Shah Alam
Selangor, Malaysia

Chiou Sheng Chew[4]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Melaka Kampus
Jasin, Melaka, Malaysia

Lala Septem Riza[5]
Department of Computer Science Education
Universitas Pendidikan Indonesia
Indonesia

***Abstract*—The ubiquity of smartphone use nowadays is undeniable exponentially growing, replaced cell phones, and a host of other gadgets replaced personal computers to a certain degree. Different smartphones specifications and overwhelmed smartphone advertisements have caused broader choices for the customer. Many qualitative and quantitative criteria need to consider, and customers want to select the most suitable smartphones. They face difficulties deciding the best smartphone according to their budget and desire. Thus, a new method is needed to recommend the customer according to their preferences and budget. This study proposed a method for optimizing the recommendation system of the smartphone using the genetic algorithm (GA). Moreover, it is implemented with a progressive web application (PWA) platform to ensure the customer can use it on multiple platforms. They can choose the platform to input any specification of smartphone preferences besides the budget. Functional testing results showed the achievement of the study's objectives, and usability testing using UEQ managed to receive feedback of 93.64%, with an overall average mean of 4.682. Therefore, according to the outcome, it can be concluded that optimizing the smartphone recommendations through GA enables the customer to ease the comparison based on the obtained optimum result.**

Keywords*—Genetic algorithm; progressive web application; recommendation; smartphone introduction

I. Introduction

In this modern era, smartphones are one of today's most necessary and personal technologies [1]. Smartphones can drastically alter how humans communicate, consume information consumption, and use their time. Smartphones are used to make calls, read or send emails, view and upload images and videos, play games, and listen to music [2]. Besides, it reacts as a personal diary to record reminders or schedules and contacts, browses the internet, speech searching, verify the latest news update and the current or predicted climate. It also uses text chatting applications such as Facebook, Twitter, and WhatsApp and connects on social networks [3].

The selection of smartphones compatible with consumers' needs has motivated our research topic. Molinera et al. [4] claimed that whenever it comes to purchasing a new smartphone model, customers can easily get lost in the midst of hundreds of advertisements from different companies. We have surveyed 100 respondents, and it shows that 89% of respondents have trouble choosing the right smartphone. Also, 94% of respondents of the same survey claimed that it is more convenient to recommend a list of smartphones within their budget and preferences. Furthermore, they were having difficulties comparing the preferences of the smartphone features within the budget. They have to search manually through dozens of reviews on the overwhelming information on the internet. 77% of respondents claimed that they read reviews or comment on the smartphone before making a decision. It is time-consuming and puts much effort into cognitive while manually searching the rating and feedback based on user preferences [5]. The consequences of relying on reviews will be a combination of negative and positive feedback, and consumers feel difficult to seek a cogent response. Besides, the unauthorized review may be a scam too.

Nonetheless, most comparison approaches use the specific meaning of the attribute. Sometimes, customers may not accurately describe the artifacts in which they are interested. They do not entirely understand the level or degree of specific attributes and hard to locate a precise analysis for a specific service feature [6]. The recommendation of the product's latest features in an adaptive manner is not successful due to the high-end products' short life cycles. It has caused inappropriate reviews and obsolete scores rated by other users [7].

Sponsored by Indonesian Ministry of Research and Technology and Universitas Pendidikan Indonesia.

Cha & Seo [8] claimed an average of 54% estimated that 21 developing and emerging countries such as Malaysia, Brazil, and China had at least one internet or smartphone by 2015. The ubiquity of smartphone use nowadays is undeniable exponentially growing. Smartphones have become a host of other gadgets, replaced cell phones, to a certain degree, replaced personal computers [9]. In 2017, 5 billion individuals possessed smartphones, and by 2025, this number is predicted to rise to 5.9 billion. 95% of the population in the United States owns a smartphone. Through new features, smartphones are constantly evolving, becoming cheaper and faster each year at the same time. Therefore, it is essential to consider the quality and quantitative requirements to select the most preferred smartphones. For example, pixel density, the camera resolution, RAM, battery power, stand-by time, memory built-in, weight, thickening, scaling, type of the processor, processor, and costs are quantitative parameters. In the meantime, consistency requirements include longevity, reliability, aesthetics, and branding. Thus, Chen et al. [10] indicated that customers' purchasing decisions are different because their perceptions and desires vary. Customers would feel satisfied with the criteria leading to an informed decision to make one-hand purchases and meet their expectations.

Now-a-days, humans have been very dependent on mobile phones since developing robust mobile applications. Besides, mobile applications thrived originally aimed for productivity assistance and information retrievals such as emails, calendars, and contact databases. Due to the rapid advancement of technology and public demands, it is essential to implement an effective development of mobile applications as there is a need to overcome many challenges [11]. Unfortunately, every invention comes with limitations where mobile applications need to be compatible with the platform of the devices to work.

As a solution, in this study, the system will be a progressive web application (PWA) which is application software that can cross any platform. This enhancement of PWA does not force users to download the application to experience the features. However, the functionality remains the same [12]. Besides, this recommended system research optimization technique focuses on adapting a GA. The selection method works by using each member's fitness function according to the fitness value calculated. Resulting the fittest member is more likely to be selected based on the selection likelihood. This study outcome, SRcS, which stands for Smartphone Recommendation System, helps recommend which smartphone is affordable for the customer to purchase that follows customer preferences within the budget. This paper's organization begins with a brief introduction in Section 1. Section 2 explained the literature review and followed by methodology in Section 3. Section 4 elaborates on results and discussion. Finally, Section 5 concludes the study and briefly mentions future enhancement.

II. Literature Review

This section describes the smartphone preferences, recommendation technique, and progressive web application on the related issue.

A. Smartphone Preferences

A smartphone merged some electronic devices and became a miniature of a computer. It supports mobile or portable computing technology and applications with efficient operating systems [13]. The opportunities provided by the internet eventually makes smartphones often provide qualitatively different service. Smartphone has become a part of human life basic needs nowadays. Rotondi et al. [14] claimed that the smartphone's advent has significantly changed how information is accessed, allocated time, and interacted with others. The consumer's decision-making process depends on the product attributes. Price is the most obvious concerning the attribute of smartphones [15]. Due to that, a smartphone's price plays a vital role in a company's market strategy. Customers will also compare their needs and want between various products to buy their products inside their budget fit [16]. Therefore, the product quality must match the price to find that it is worth investing in a smartphone.

Next, consumers' need for multi-function cell phones drives the smartphone's development [17]. There are plenty of platforms for development, but the two famous and excellent platforms are iPhone Operating System (iOS) from Apple and Android from Google. Up until now, Android and iOS remain to dominate the market share of smartphones worldwide. Despite that, the Android operating system is considerably newer than iOS. Android utilizes iOS weaknesses and promotes a tangible cross-platform development operating system [18].

Furthermore, [19] agreed that organizations will always find ways to be different, especially in the smartphone industry, which continuously changes technology. The brand name can be an organization's brand and exclusivity. The brand name can be a title, word, logo, and design to differentiate its rivals such as Acer, Amazon, Apple, Samsung, BlackBerry, Nokia, Huawei, Lenovo, Microsoft, One Plus, Oppo, ZTE and Sony, and. Marketers were trying to create brand equity to improve customer response to win consumer preference and loyalty. Brand equity represents how the brand thinks, feels, and behaves. Thus, it becomes the products and services value-added [20].

Smartphones will only become more and more popular. Most people depend on their mobile devices to run their lives nowadays. Thus, smartphone brands need to thoroughly understand the current use and future adoption. The brand presence is essential, as it ensures that the business has a specific role in the markets and has established its reputation in the consumer's view [21]. However, [22] claimed that choice depends on the consumer's different variables, calculated by the utility. This proposed project focuses on the top five smartphone model brands in Malaysia: Samsung, Vivo, Oppo, Huawei, and Xiaomi. Purwanto [23] revealed a study outcome during the covid 19 pandemic where sales promotion and brand image influenced smartphone purchasing.

Besides, there are many high technology smartphones features available in the market today. Therefore, different individuals can choose a specific smartphone to meet their needs and desires—the smartphone features, including software and hardware. Hardware is a system concept that can

be physically touched; meanwhile, software, for instance, computer programs, procedures, and documentation are the general terms. Hardware is the smartphone's size, design, color, body, and weight, whereas software consists of the documentation and application. Rahman et al. [24] alleged that many consumer choices could be rational, such as time management, communication, and emotional, such as camera, games, music, and application features.

Cost, reliability, battery's lifespan, special promotions, camera resolution, size, storage offered, networking, or connectivity options affect customers' features when purchasing any smartphone. People believed that the smartphone's size connects with the screen's resolution and is inversely linked, such as the bigger the phone, the higher the resolution, and the harder it to carry. Therefore, with the enormous open doors within a short period in the smartphone showcase, smartphone suppliers need to understand factors that satisfy the customer decisions on which model to buy [25]. All of these demands in preferences become the input. Then, the system processed the algorithm and produced a list of smartphones that matched the most input preferences.

B. Recommendation Technique

Artificial intelligence (AI) approaches have become more prevalent in a variety of fields. For instance, recommender systems provide consumers with recommendations for selecting different items from a massive pool of items [26]. Consequently, it creates a program that can allow people to select requirements and remove the dilemma. Numerous options allow humans to be uncertain about what is best for them or fulfil their needs. The recommendation helps customers reduce the time and difficulty of searching for the information required. The methods promote customers towards the product by collecting and evaluating feedback from other buyers, implying reviews from specific establishments and even the customer [27].

Consequently, many new researchers have embarked on this study to develop more recommended research and techniques. Several techniques have been evaluated based on the accuracy, ability to receive multiple inputs, and simplicity—for instance, fuzzy logic, content-based filtering, and genetic algorithm. Table I explains the details of the comparison.

TABLE I. FEATURES COMPARISON BETWEEN RECOMMENDATION TECHNIQUES

Technique / Features	**Content-Based Filtering**	**Fuzzy Logic**	**Genetic Algorithm**
Accuracy	Medium	Low	High
Receive many inputs and run in a single run	Yes	No	Yes
Simplicity	No	No	Yes

In conclusion, GA has been chosen because simple programmability and efficiency features offered. GAs is a robust optimization system widely applicable since it requires users to give many inputs to run in a single run [28]. The GAs maintained the population of an individual's chromosomes along with their fitness scores. It gave more opportunities for individuals with better fitness scores to reproduce than others. Thus, GA can give the best optimization solution to the smartphone buyer. No matter what the user may input into the system, GA will always provide one recommendation instead of null.

C. Progressive Web Application

PWA is an abbreviation for Progressive Web Application. It is also a cross-platform with a new approach that modern web capabilities provide a user experience. PWA uses the most recent technology to incorporate the best of web and mobile apps. PWA hence unifies the browsing web experience on mobile and other devices of various pixel sizes, including laptops, tablets, and other devices [29]. The web-based framework is designed using HTML, CSS, and JavaScript standards. It is compatible with any platform that supports standards-compliant browsers. Besides, the PWA development and evolution is not a new framework or technology. It allowed the mobile expansion externally for cross-platform [30]. With the advantage of a mobile app's features, PWA enhances user retention and execution without complicating a mobile application's maintenance. Biørn-Hansen et al. [31] declared that the service worker sits at the heart of a PWA because, without a service worker, support will cause PWA not to work correctly. A service worker helps give the consumers of a web application an offline experience. A service worker is a client-side script that operates on a different JavaScript thread and is independent of the web application. It helps developers programmatically store and preload data so that the code can be loaded from the user cache if the network connection fails.

Furthermore, PWA requires a manifest file. The JSON file is the manifest file for the web application that applies to the user-installable home screen. A manifest file configures the application includes name, short name, icons, background color, view, width, and theme color. It manages to change the behavior and design of PWA. The PWA platform adapts in developing this recommendation system because it is understandable, reliable, and faster to access. Besides, PWA is a regular application on a computer. The ability to run it from a uniform resource locator (URL) makes it easy for users with a browser to use the program [32]. Therefore, it is unnecessary to maintain an application programming interface (API) with backward compatibility. Each user uses the same website version of the code, unlike the version fragmentation of native apps, making it easier to deploy and manage the software. Meanwhile, web-based information systems offer easy and cost-efficient resources to facilitate usability, effective delivery, efficient administration, and cross-platform versatility.

III. METHODS

Four subsections describe the flow in implementing the proposed idea: system use case, system flowchart, the phases of GA implementation, and PWA implementation.

A. System use case for SRcS

The overall system use case illustrated in Fig. 1 demonstrates using the UML on users' interaction. We identify ten use cases for this system: seven use cases handled by the admin and three by the user.

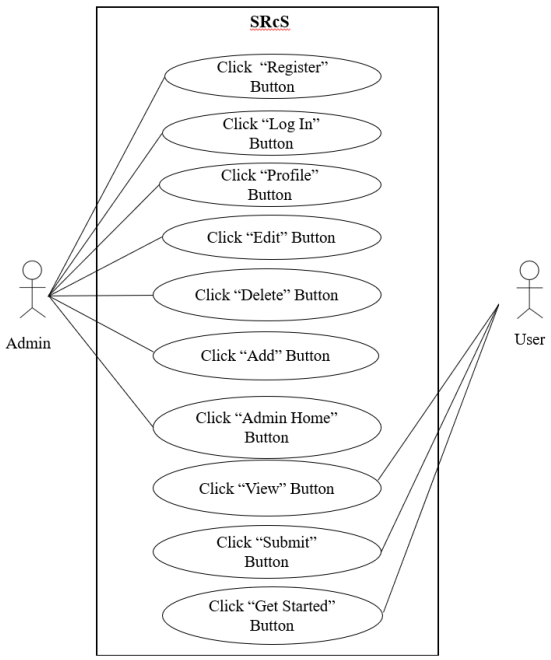


Fig. 1. SRcS use Case Diagram.

B. Flow Chart for SRcS

A visual representation of the series of steps and decisions or a flowchart requires a system using different symbols containing information. It is essential in design phases to avoid any obstacle and clearly describe the system. The flow of the recommendation process for SRcS shown in Fig. 2. The user must provide 16 specifications of their preferences into the system, including the budget. The chosen specification will then go through the five GA processes to get the smartphone's highest match with the user's input. Lastly, resulting in the top three smartphone recommendation lists.

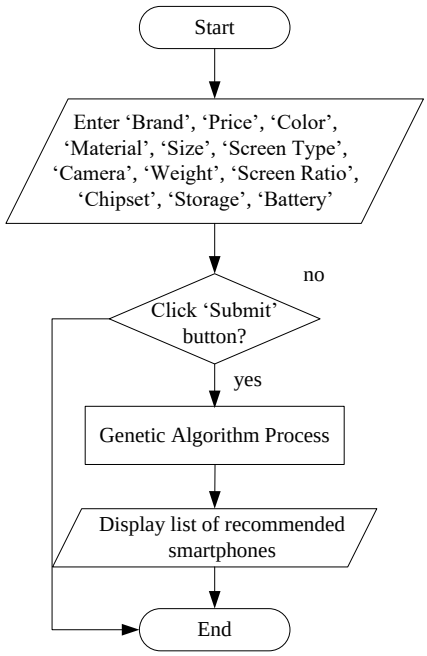


Fig. 2. Process flow for SRcS.

C. Genetic Algorithm Implementation

As mentioned, GA involves five processes that start with initializing the population, followed by fitness calculation, crossover, mutation, and convergence. All user information is stored in the system as input. This sub-section presented a detailed description of each process involved and how GA produces the final result. Fig. 3 shows the basic process of GA, which consists of six main steps: 1) initialize population, evaluate fitness, 2) create a new population through the selection of the individuals, 3) process the crossover and mutation, 4) test the condition and if satisfied, return the best individual of the current population. Else, repeat the process.

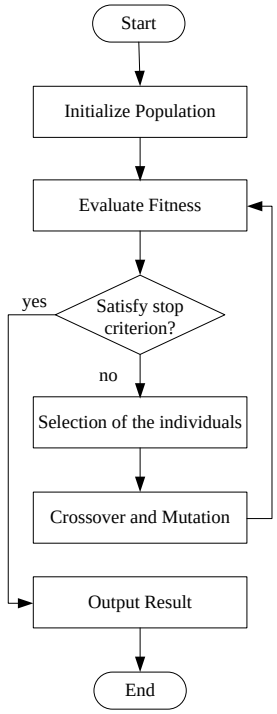


Fig. 3. Process for Genetic algorithm

1) Step 1: Initialize population: GA begins with an initial population of typically randomly formed phenotypes. The GA needs to continue to evolve new genotypes from the population and evaluate each genotype's fitness at each iteration. The population will create chromosomes up to 150 generations as their stopping condition is defined in for loop. Each chromosome encodes three types of smartphones. Each smartphone contains information like the brand, price, colour, material, size, year release, display, camera, weight, chipset, CPU, GPU, RAM, memory, and battery, as in Fig. 4.

2) Step 2: Evaluate fitness: The next phase is calculating each chromosome's fitness by comparing the user with database chromosomes. Each smartphone holds an equal percentage of totalling 100%. The fitness function is the inverse of the input given, for example, using three variables: a, b, and c. Fitness means the best result for the input given for a, b and c, so we can assume the value will be d as in (1).

$$a + b + c = d \quad (1)$$

Smartphone23	Smartphone67	Smartphone09	Fitness
Brand	Brand	Brand	
Price	Price	Price	
Color	Color	Color	
Material	Material	Material	
Size	Size	Size	
Release Year	Release Year	Release Year	
Screen Type	Screen Type	Screen Type	
Camera Number	Camera Number	Camera Number	
Rear Camera	Rear Camera	Rear Camera	
Front Camera	Front Camera	Front Camera	
Weight	Weight	Weight	
Resolution	Resolution	Resolution	
Chipset	Chipset	Chipset	
GPU	GPU	GPU	
RAM	RAM	RAM	
Storage	Storage	Storage	
Battery Capacity	Battery Capacity	Battery Capacity	
Battery Type	Battery Type	Battery Type	

Fig. 4. Chromosome Encoding.

The process of fitness function declares as the inverse of | a + b + c – d | because of the need to reduce the sum of the three variables from deviating from d. Thus, the fitness function identifies as in (2).

$$\text{Fitness Function} = 1/ | a + b + c - d \quad (2)$$

3) Step 3: Crossover and mutation: After calculating the value of fitness, the best fitness value is chosen and arranged to descend from the highest fitness-to-lowest. The crossover and mutation operation uses the first three highest fitness values for chromosomes. Then, it follows by sorting out the fitness value. Fig. 5 shows the crossover example between chromosome X and Y. The GA process chooses and displays the highest fitness value data to the user.

D. Progressive Web Application Implementation

A PWA requires a web manifest and service worker file. The manifest file allows the system to execute the full-screen web application as a standalone application. It can assign an icon to show when finishing the application and assign a theme and background colour app on the computer. Furthermore, this application also has implemented an installation banner that makes it easier to be download on any device.

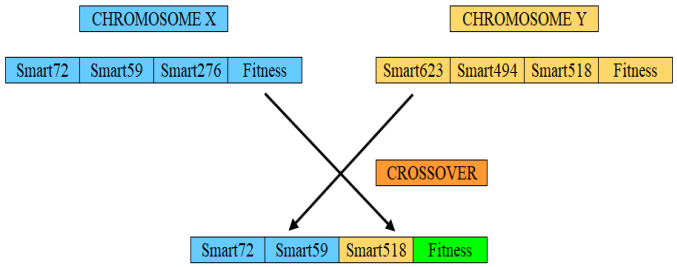


Fig. 5. Example of Crossover Process.

Next, service workers are the mastermind of PWA, in which it reacts as middleware by intercepting each request. It responds instantly to the cached request or performs the channel recovery. There are two caches implemented in this system, which are dynamic and static. In static stores, every single asset while in dynamic fetches all previously requested assets while users online limit to 20 requests to be stored. Inside service worker also implements install and activate event code. An event code fires when the service worker is mounted and occurs once. If the service worker is installed and activated, the device will use the currently installed service worker. The caches are deleted whenever there are changes to cache the latest version of code. Therefore, every declared asset will be cached automatically.

E. Evaluation and Acceptance

In this study, two types of tests were performed, which are testing on functionality and usability. Functionality evaluation is testing to verify the outcome for each use case module. Every module is evaluating whether it could generate the predicted result. Usability testing is about bringing actual people to connect with the system and watch their behavior and reactions. The key benefit of usability testing is to detect usability problems with a design as early as possible before the design is adopted. This step ensures that the program built is convenient for someone with no computer science experience to use. Therefore, we do the evaluation using the User Experience Questionnaire (UEQ). UEQ is a quantitative survey proposed by [33] [34], and we test it according to the SRcS functionality. UEQ consists of 26 dimensions, but we chose five dimensions related to the study as in Table II.

TABLE II. FIVE UEQ DIMENSION AND DESCRIPTION FOR USABILITY TEST

Dimension	Description
Attractiveness	Overall impression of the product. Do users like or dislike the product?
Perspicuity	Is it easy to get familiar with the product? Is it easy to learn how to use the product?
Efficiency	Can users solve their tasks without unnecessary effort?
Usefulness	Is it useful? Helpful? Beneficial? Rewarding using the application?
Novelty	Is the product innovative and creative? Does the product catch the interest of users?

Our research respondents consisted of 30 public participants who randomly took part in the application testing. Firstly, we briefed the participants on project details and what they were required to do with the application. Then they tried the application until they were satisfied with the recommendation given by the system. Once they finished it, we issued the UEQ using the Google Form.

IV. RESULTS AND DISCUSSION

A. Functionality Testing

Functionality evaluation is testing to verify the outcome for each use case module. Every module is evaluating whether it could generate the predicted result. Fig. 6 and Fig. 7 indicate the SRcS snapshot, the user filling up the form with the questions that began with the brand, price, and specification preferences question. Then, SRcS shows the user's smartphone recommendation result, as in Fig. 8.

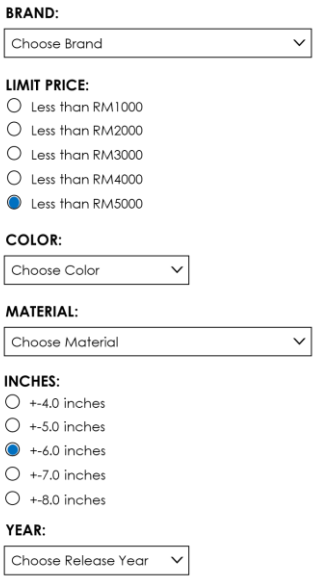


Fig. 6. Snapshot of SRcS submenu 1

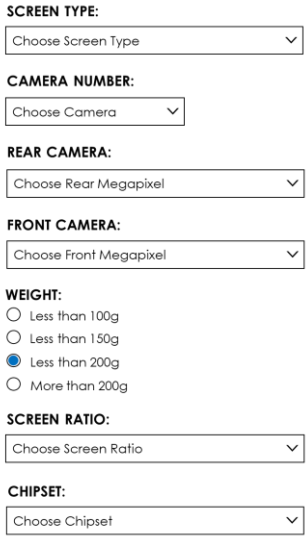


Fig. 7. Snapshot of SRcS submenu 2

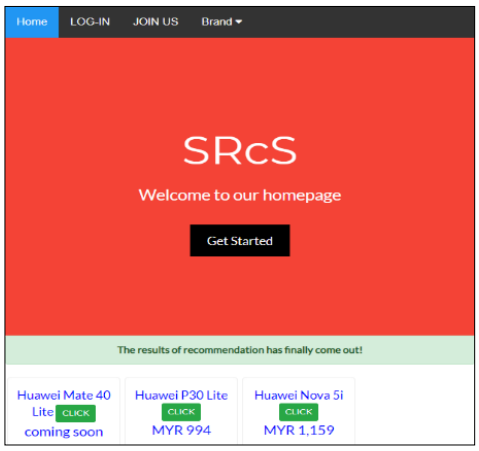


Fig. 8. SRcS recommendation result

The evaluation of the functionality test follows according to the use case of SRcS, and Table III displays the SRcS outcome on the functionality test to ensure that it works according to the proposed.

TABLE III. SRCS FUNCTIONALITY TEST RESULT

Use Case	Description	Remark
Register Button	Allows a new admin to register into the system	Successful
Log In Button	Allows admin to log in to the profile page	Successful
Profile Button	Allows admin to view their account information	Successful
Edit Button	Allows admin to update their account information	Successful
Delete Button	Allows admin to delete smartphone details system	Successful
Home Button	Allows both user and admin to view all smartphones available in the database	Successful
Find Button	Allows both users to find a smartphone that matches with user's preferences	Successful
Profile Button	Allows admin to view their account information	Successful

B. *Usability Testing*

We evaluate the feedback given by the 30 respondents and summarize the UEQ results for the five items. Each of the dimensions has a related and specific questionnaire to get a quantified value. Table IV shows the overall average value of the UEQ with the specific result for mean and average mean.

For the first dimension attractiveness, feedback shows that respondents felt that the SRcS shows the highest average mean of 4.800 for four questionnaires that asked whether the application is enjoyable, good, pleasant to use and user friendly. Item A2 get the highest average mean among the rest dimension with 4.933, and we get direct feedback that the application is good. Item A4 get the lowest average mean of 4.667 for the first dimension due to not all specification being well-known by some respondents. We further analyze the second dimension, perspicuity, which is related to the ease of using the application with an average mean score of 4.711. We guest the same issue for item P2 with A4, where not all users have deep knowledge about the smartphone specification. Dimension three is dependability asked on the application's reaction to the user input, whether predictable and meets expectations. Item D1 gets the lowest mean among the rest with 4.433, but item D2 shows a contra result that the input and command meet the user's expectations. The fourth dimension is related to the usefulness of the application. Item U1 until U3 managed to get the result more than 4.500 mean with the average mean of 4.750. We can assume that the system is really useful to the respondents. The last dimension is novelty involved in the idea behind the application into four different criteria: creative, inventive, leading edge and innovative. Although we got the lowest mean for item N3, we managed to get an average mean of 4.533, which is more than 4.500. In summary, the average overall mean score for SRcS is 4.682 or 93.64% conclude that the system considers received a 'High' level of usability acceptance.

TABLE IV. MEANS, STANDARD DEVIATION AND CONFIDENCE INTERVALS UEQ FOR SRCS

Dimension	Item	Question	Mean	Average Mean
Attractiveness	A1	In your opinion, the application is enjoyable	4.767	**4.800**
	A2	In your opinion, the application is good	4.933	
	A3	In your opinion, the application is pleasant to use	4.833	
	A4	In your opinion, the application is user friendly	4.667	
Perspicuity	P1	In your opinion, the application is easy to understand	4.833	**4.711**
	P2	In your opinion, the application is easy to learn	4.533	
	P3	In your opinion, using the application is easy	4.767	
Dependability	D1	In your opinion, the reactions of the application to your input and command is predictable	4.433	**4.617**
	D2	In your opinion, the reactions of the application to your input and command meets expectations	4.800	
Usefulness	U1	You consider using the application as useful	4.533	**4.750**
	U2	You consider using the application as helpful	4.767	
	U3	You consider using the application as beneficial	4.867	
	U4	You consider using the application as rewarding	4.833	
Novelty	N1	In your opinion, the idea behind the application and the designs are creative	4.567	**4.533**
	N2	In your opinion, the idea behind the application and the designs are inventive	4.500	
	N3	In your opinion, the idea behind the application and the designs are leading edge	4.433	
	N4	In your opinion, the idea behind the application and the designs are innovative	4.633	
Average Overall Mean Score			**4.682**	
Average Percentage of Mean Score			**93.64%**	

V. CONCLUSION

This study aims to develop a smartphone recommendations system (SRcS) using a GA adaptation with innovative PWA. With the GA advantages, SRcS helps users seek out and purchase a smartphone according to specification preferences, needs, and allocated budget. Tacitly, it helps to ease the time-consuming manual survey and comparison via websites. The outcome performed from the functionality testing by assessing and testing the use case function proves the SRcS functions work correctly. The usability testing using the five scale in UEQ shows a good result with a positive evaluation value mean scores that indicate the majority of the respondents preferred using the SRcS. The benchmark result also shows an excellent trend and prove the acceptance of SRcS. For the next improvement, SRcS can expand the fitness of the brand's choices, view the smartphone's picture in a 3D rotation image, and recommends the authorized seller.

ACKNOWLEDGMENT

The authors would like to acknowledge the Indonesian Ministry of Research and Technology and Universitas Pendidikan Indonesia for funding this work.

REFERENCES

[1] S. V. Manikanthan, T. Padmapriya, A. Hussain, and E. Thamizharasi, "Artificial intelligence techniques for enhancing smartphone application development on mobile computing," International Journal of Interactive Mobile Technologies, vol. 14, no. 17, pp. 1–16, 2020.

[2] A. B. Mohammed, "Selling smartphones to generation Z: Understanding factors influencing the purchasing intention of smartphone," International Journal of Applied Engineering Research, vol. 13, no. 6, pp. 3220–3227, 2018.

[3] M. Samaha and N. S. Hawi, "Relationships among smartphone addiction, stress, academic performance, and satisfaction with life," Computers in Human Behavior, vol. 57, pp. 321–325, 2016.

[4] J. A. M. Molinera, I. J. P. Gálvez, R. Wikström, E. H. Viedma, and C. Carlsson, "Designing a decision support system for recommending smartphones using fuzzy ontologies," Advances in Intelligent Systems and Computing, vol. 323, pp. 323–334, 2014.

[5] J. Feuerbach, B. Loepp, C. M. Barbu, and J. Ziegler, "Enhancing an interactive recommendation system with review-based information filtering," in Interfaces and Human Decision Making for Recommender Systems IntRS@ RecSys, 2017, pp. 2–9.

[6] D. Kamalapurkar, N. Bagwe, R. Harikrishnan, S. Shahane, and G. Manisha, "Phone recommender: sentiment analysis of phone reviews," International Journal of Engineering Sciences & Research Technology, vol. 6, no. 5, pp. 212–217, 2017.

[7] K. K. F. Yuen, "The fuzzy cognitive pairwise comparisons for ranking and grade clustering to build a recommender system: An application of smartphone recommendation," Engineering Applications of Artificial Intelligence, vol. 61, pp. 136–151, 2017.

[8] S. S. Cha and B. K. Seo, "Smartphone use and smartphone addiction in middle school students in Korea: Prevalence, social networking service, and game use," Health Psychology Open, vol. 5, no. 1, pp. 1–15, 2019.

[9] R. Trivedi and R. Raval, "Consumer buying intentions towards smartphones : A conceptual framework," International Journal of Applied Research, vol. 2, no. 12, pp. 736– 742, 2016.

[10] Y. S. Chen, T. J. Chen, and C. C. Lin, "The analyses of purchasing decisions and brand loyalty for smartphone consumers," Open Journal of Social Sciences, vol. 4, no. 7, pp. 108–116, 2016.

[11] N. A. Kumar, K. T. H. Krishna, and R. Manjula, "Challenges and best practices for mobile application development," Imperial Journal of Interdisciplinary Research, vol. 2, no. 12, pp. 1607–1611, 2016.

[12] V. Sharma, R. Verma, V. Pathak, M. Paliwal, and P. Jain, "Progressive web app (PWA) - one stop solution for all application development across all platforms," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 2, pp. 1120–1122, 2019.

[13] D. Wang, Z. Xiang, and D. R. Fesenmaier, "Smartphone use in everyday life and travel," Journal of Travel Research, vol. 55, no. 1, pp. 52–63, 2016, doi: 10.1177/0047287514535847.

[14] V. Rotondi, L. Stanca, and M. Tomasuolo, "Connecting alone: smartphone use, quality of social interactions and well-being," Journal of Economic Psychology, vol. 63, pp. 17–26, 2017, doi: 10.1016/j.joep.2017.09.001.

[15] J. Fölting, S. Daurer, and M. Spann, "Consumer preferences for product information and price comparison apps," in 13th International Conference on Wirtschaftsinformatik, 2017, pp. 1081–1095.

[16] P. Y. Satriawan, K. A., & Setiawan, "The role of purchase intention in mediating the effect of perceived price and perceived quality on purchase decision," International Research Journal of Management, IT and Social Sciences, vol. 7, no. 3, pp. 38–49, 2020.

[17] H. A. Watson, R. M. Tribe, and A. H. Shennan, "The role of medical smartphone apps in clinical decision-support," Artificial intelligence in medicine, vol. 100, pp. 1–11, 2019.

[18] M. H. Goadrich and M. P. Rogers, "Smart smartphone development: iOS versus android," in Proceedings of the 42nd ACM Technical Symposium on Computer Science Education, 2011, pp. 607–612.

[19] P. K. A. Ladipo, M. A. Awoniyi, and O. S. Akeke, "Influence of smartphone attributes on student's buying decision in Lagos State tertiary institutions," Jurnal Manajemen Dan Kewirausahaan, vol. 6, no. 1, pp. 70–81, 2018, doi: 10.26905/jmdk.v6i1.1938.

[20] R. P. Bringula, S. D. Moraga, A. E. Catacutan, M. N. Jamis, and D. F. Mangao, "Factors influencing online purchase intention of smartphones: A hierarchical regression analysis," Cogent Business & Management, vol. 5, no. 1, pp. 1–18, 2018.

[21] B. Chen, H., Zhang, L., Chu, X., & Yan, "Smartphone customer segmentation based on the usage pattern," Advanced Engineering Informatics, vol. 42, pp. 1–13, 2019.

[22] Y. W. Sullivan and D. J. Kim, "Assessing the effects of consumers' product evaluations and trust on repurchase intention in e-commerce environments," International Journal of Information Management, vol. 39, pp. 199–219, 2018.

[23] A. Purwanto, "Exploring factors affecting buying interest of smartphones during the Covid 19 pandemic.," Journal of Industrial Engineering & Management Research, vol. 2, no. 4, pp. 124–130, 2021.

[24] M. Rahman, Y. Ismail, M. Albaity, and C. R. Isa, "Brands and competing factors in purchasing hand phones in the Malaysian market," The Journal of Asian Finance, Economics, and Business, vol. 4, no. 2, pp. 75–80, 2017.

[25] S. Jain and B. Singh, "Consumer behavior toward mobile phone handsets," in International Conference on Innovative Computing and Communications, 2019, pp. 61–69.

[26] M. Kuanr, B. K. Rath, and S. N. Mohanty, "Crop recommender system for the farmers using mamdani fuzzy inference model," International Journal of Engineering & Technology, vol. 7, no. 2, pp. 277–280, 2018.

[27] Patil, A. E., S. Patil, K. Singh, P. Saraiya, and A. Sheregar, "Onlinebook recommendation system using association rule mining and collaborative filtering," International Journal of Computer Science and Mobile Computing, vol. 8, no. 4, pp. 83–87, 2019.

[28] K. A. F. A. Samah et al., "Optimization of house purchase recommendation system (HPRS) using genetic algorithm," Indonesian Journal of Electrical Engineering and Computer Science, vol. 16, no. 3, pp. 1530–1538, 2019.

[29] O. Adetunji, C. Ajaegbu, N. Otuneme, and O. J. Omotosho, "Dawning of progressive web applications (pwa): Edging out the pitfalls of traditional mobile development," American Scientific Research Journal for Engineering, Technology, and Sciences, vol. 68, no. 1, pp. 85–99, 2020.

[30] T. A. Majchrzak, A. Biørn-Hansen, and T.-M. Grønli, "Progressive Web Apps: the definite approach to cross-platform development?," Proceedings of the 51st Hawaii International Conference on System Sciences, pp. 5735–5744, 2018, doi: 10.24251/hicss.2018.718.

[31] B.-H. A., T. A. Majchrzak, and T. M. Grønli, "Progressive web apps: The possible web-native unifier for mobile development," in 13th International Conference on Web Information Systems and Technologies, 2017, pp. 344–351.

[32] B. Frankston, "Progressive web apps [bits versus electrons]," IEEE Consumer Electronics Magazine, vol. 7, no. 2, pp. 106–117, 2018.

[33] B. Laugwitz, T. Held, and S. Martin, "Construction and evaluation of a user experience questionnaire," in Symposium of the Austrian HCI and Usability Engineering Group, 2008, pp. 63–76.

[34] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Applying the user experience questionnaire (UEQ) in different evaluation scenarios," in International Conference of Design, User Experience, and Usability, 2014, pp. 383–392.

SG-TSE: Segment-based Geographic Routing and Traffic Light Scheduling for EV Preemption based Negative Impact Reduction on Normal Traffic

Shridevi Jeevan Kamble, Manjunath R Kounte
School of Electronics and Communication
REVA University, Bengaluru-560064, India

***Abstract*—Emergency Vehicles (EVs) play a significant role in giving timely assistance to the general public by saving lives and avoiding property damages. The EV preemption models help the EVs to maintain their speed along their path by pre-clearing the normal vehicles from the path. However, few preemption models are designed in literature, and they lack in minimizing the negative impacts of EV preemption on normal vehicle traffic and also negative impacts of normal vehicle traffic on EV speed. To accomplish such goals, the work proposes a Segment-based Geographic routing and Traffic light Scheduling based EV preemption (SG-TSE) that incorporates two mechanisms: Segment based Geographic Routing (SGR) and Dynamic Traffic Light Scheduling and EV Preemption (DTSE) for efficient EV preemption. Firstly, the SGR utilized a geographic routing model through the Segment Heads (SHs) along the selected route and passed the EV arrival messages to the traffic light controller to pre-clear the normal traffic. Secondly, the DTSE designs effective scheduling at traffic lights by dynamically adjusting the green time phase based on the minimum detection distance of EVs to the intersections. Thus, the EVs are passed through the intersections quickly without negatively impacting normal traffic, even the signal head in the red phase. Moreover, the proposed SG-TSE activates the green phase time at the correct time and minimizes the negative impacts on the EV preemption model. Finally, the performance of SG-TSE is evaluated using Network Simulator-2 (NS-2) with different performance metrics and various network traffic scenarios.**

Keywords*—Emergency vehicle (EV) preemption; Segment-based Geographic routing and Traffic light Scheduling based EV preemption (SG-TSE); geographic routing; Segment based Geographic Routing (SGR); dynamic traffic light scheduling; Dynamic Traffic light Scheduling and EV preemption (DTSE); green phase adjustment

I. Introduction

The Vehicular ad hoc networks (VANETs) enable real-time communication among the roadside vehicles with the support of roadside infrastructure [1]. The VANETs receive high popularity among researchers owing to the application diversity [2]. The VANET applications are mainly categorized into safety and infotainment. In infotainment applications, the vehicles exchange messages about parking areas and hotels and make the journey very comfortable. In contrast to infotainment applications, the safety applications alert the drivers about hazardous situations such as crash warning, accident warning, EV preemption, and others. Hence, the safety applications require strict delay bounds compared with comfort applications. The Emergency Vehicle (EV) preemption is one of the prime VANET applications in which the emergency vehicles are quickly navigated from the approaching lane and intersections [3]. The EVs such as ambulances, fire fighting vehicles, police vehicles, and other defense fighting vehicles receive high priority on roads, as they have to reach their destination on time to save human lives and property losses. The traffic lights integrate various preemption methods and assure desired speed to EVs along its selected path to the incident location to benefit such EVs [4].

The Emergency vehicle preemption system (EVP) interrupts the signal timings of normal traffic at the signalized intersections and provides a green band to the EVs along its routes [5]. Thus, the preemption assists the EVs to pass without stopping or waiting at intersections. It potentially minimizes the travel time and shrinks conflicts with other vehicles in the traveling route [6]. However, it may also negatively impact the general vehicle traffic in the approaching lane. It suffers the vehicles not only in the corresponding intersection but also on other neighboring intersections of coordinated signal control. Hence, it is essential to activate the traffic light green phase at the correct time to reduce the negative impact of EV on normal traffic and also pre-clear the roads in an efficient way to minimize the negative impacts of normal traffic on EVs. Owing to the high dynamic nature and frequent link failure, the geographic routing protocols are highly fit for the VANET environment [7]. Therefore, this work proposes a novel EV preemption model in which segment-based geographic routing and effective traffic light scheduling pass the EVs at intersections quickly with the desired speed.

By designing efficient EV preemption with timely green phase activation, the proposed SG-TSE diminishes both negative impacts, such as the negative impact on normal traffic due to EV preemption and the negative impact on EV due to normal traffic. The conventional methods handle only negative impact issues, resulting in inappropriate EV preemption and traffic light control. Thus, it leads to losses of human life and property damages. Hence, crucial green phase activation is required with optimal routing strategies. The SE-TSE solves such an issue significantly by splitting the vehicle density of highly congested scenarios into multiple segments and organizing the vehicles with accurate green phase activation. For that the SG-TSE utilizes a geographic routing

method. Compared to existing preemption methods, the performance of the proposed model is highly superior in terms of EV preemption speed, especially under a high vehicle density scenario. By navigating the emergency vehicles quickly along its path even the road is congested, the SG-TSE saves human lives and prevents property losses from a hazardous situation.

A. Contribution

The main contributions of the proposed work are as follows.

- To guarantee the desired speed of EVs at intersections under feasible traffic conditions, this work proposes an SG-TSE protocol that includes two different mechanisms, SGR and DTSE, to achieve its objective.
- To announces the EVs arrival to the traffic light controller, the SGR divides the EV approaching lane into many segments and elects an SH in each segment for geographic message routing.
- The DTSE uses an effective traffic light scheduling model in which the green phase time is adjusted based on the minimum detection distance of EVs from the intersection that effectively diminishes the normal traffic impacts on EV speed and also EV preemption impact on normal vehicles.
- Finally, the effectiveness of the proposed SG-TSE is evaluated using NS-2. The performance is analyzed with various metrics like packet delivery ratio, overhead, throughput, delay, and EV preemption speed under different network traffic conditions.

B. Paper Organization

The remaining part of the SG-TSE is organized as follows. Section 2 survey the paper related to EV preemption models and analyzes the gaps in the existing works. Section 3 provides an overview and network model of SG-TSE. Further, it clearly describes the two mechanisms such as SGR and DTSE. Section 4 describes the performance evaluation by applying the simulation parameters and performance metrics for SG-TSE performance analysis. Finally, Section 5 concludes this paper.

II. Literature Survey

For a clear view, the survey is categorized into two types that are geographic routing methods and EV preemption methods.

A. Geographic Routing Methods

A Predictive Geographic Routing Protocol (PGRP) in [8] maximizes the connectivity to deal with VANET dynamicity. The PGRP instructs the vehicles to assign a weight to their neighboring vehicles based on the direction and the angle of the corresponding vehicles. The PGRP can predict the position information of the vehicles at the time with the help of hello packets according to the acceleration information of vehicles. The work in [9] proposes a Maxduration-Minangle GPSR (MM-GPSR) routing protocol that defines a cumulative communication duration in greedy forwarding to obtain the node stability of neighbor nodes. Further, it selects the nodes with maximum cumulative communication duration as next-hop nodes for communication. If the greedy routing is failed, the MM-GPSR utilizes the perimeter mode with the minimum angle method. The node location information is used to estimate the angles. Moreover, the MM-GPSR successfully transmits the packets to the destination with optimal forwarder nodes. The work in [10] proposes a novel geographic routing protocol named Geo-LU to enhance the VANET routing performance. It elaborates the local view of the network topology at the current forwarder by incorporating two-hop neighbor information. It exploits a link utility (LU) measure to measure the utility of a two-hop neighbor link. Further, it takes into account the minimum residual bandwidth on that link and its packet loss rate. Moreover, the Geo-LU effectively reacts to high network traffic and frequent link disconnections by including the two-hop neighbor information with LU measurement. The work in [11] proposes a dissemination mechanism with reroute planning for exchanging the emergency vehicle information.

B. EV Preemption Methods

Several research works have been designed for emergency vehicle route selection and pre-clearing by integrating the real-time traffic and travel time information [12] [13]. An emergency vehicle pre-emption strategy has been proposed in [14]. Such a preemption model can reduce the delay of emergency vehicle arrival caused due to wide network traffic. It utilizes a connected vehicle infrastructure and efficiently manages the time delay in emergency vehicle arrival. Further, the pre-emption model considers the worst-case non-emergency vehicle's waiting time issues. The work in [15] utilizes an emergency vehicle signal coordination method to offer a green wave to the emergency vehicles. The signal coordination method effectively clears the queue traffic on the road and creates a green phase for quick navigation of emergency vehicles. The work in [16] considers daily emergency vehicle routing issues in a specified network with high spatial resolution and offers effective decision support for emergency vehicular systems. The spatial resolution introduces two advanced technologies that are pre-hospital screening and lane pre-clearing.

The pre-hospital screening offers injury diagnosis of patients and lane pre-clearing assures that the ambulance is moved with desired speed in all lanes. Such a model exploits three various ambulances which can support first aids based on the pre-hospital screening. Moreover, it presents mixed-integer linear programming (MIP) strategy to allocate emergency vehicles to the patient location and navigate the vehicles promptly by planning the shortest traveling routes. Thus, it manages the ambulance fleet properly. A Virtual Traffic Light plus for Emergency Vehicle (VTL+EV) has been proposed in [17] to prioritize the emergency vehicles in an intersection. The VTL+EV is a decentralized and self-coordinated traffic control system in which the movement of emergency vehicles is expedited, and the normal vehicle waiting time is also minimized.

The work in [18] proposes a Global Positioning System (GPS) based traffic light preemption model to diminish the travel time delay of emergency vehicles. With the GPS data,

the emergency vehicle can become aware of its position and destination position. The GPS assists the preemption model by incorporating software programs with GPS technology and developing electronic maps to determine the shortest paths. Thus, the emergency vehicle selects the shortest paths based on GPS information and arrives on time. Also, the GPS-based preemption model clears the normal vehicles on the emergency vehicle path by effectively managing the traffic lights of intersections using transmitters. An innovative traffic signal control model in [19] diminishes the response time of emergency vehicles by utilizing connected vehicle infrastructure. Based on the beacons received from an emergency vehicle, such a model instructs the traffic signal to adjust the green phase earlier to reduce the arrival delay of emergency vehicles. The work in [20] proposes a priority signal control algorithm with transit signal priority to improve the emergency vehicle preemption. The transit signal priority model is a proven technique to offer an enhanced public transit operation quality in urban scenarios. The priority model tunes the traffic signal phases based on transit signal priority and serves quick preemptions to an emergency vehicle. Thus, it assists in alleviating the delay in emergency vehicle arrival and minimizing the impact of preemption on general road traffic. An emergency vehicle pre-clearing model in [21] prioritizes the emergency vehicle on the corresponding path by employing the cooperative driving of connected vehicles in a particular area. Such a model converts the connected vehicle cooperative driving issue as a mixed-integer nonlinear programming (MINP) to guarantee the emergency vehicle desired speed and to minimize the impact of pre-emption on connected vehicles. The MINP achieves the objectives by formulating a bi-level optimization model. Initially, the connected vehicles proceeding of the emergency vehicle are divided into various blocks. Further, an emergency vehicle sorting algorithm is applied in each block to sort vehicle trajectories. Thus, the MNP is solved based on the sorting trajectories, and the emergency vehicles are allowed with desired speed on the corresponding path. A novel traffic light-assisted emergency vehicle preemption method at an intersection has been introduced in [22]. Such a model employs wireless vehicles to infrastructure communication among the emergency vehicle and the traffic lights controller for preemption. It estimates the vehicle density at the intersections based on the messages and builds a dynamic mathematical model to discharge the vehicles in the queue.

The work in [23] utilizes a multi-objective programming model for emergency vehicle pre-emption at intersections. The main intention of such a model is to clear the emergency vehicles quickly at the intersection and increase the passing rate of normal vehicles by minimizing the emergency vehicle preemption impact. The work in [24] mainly focuses on constructing better routes for emergency vehicles by designing a realistic traffic-based optimization model. It obtains real-time traffic knowledge from the Google Maps Distance Matrix API. Finally, it finds the best shortest emergency vehicle path with less congestion. The real-time traffic flow-based dynamic and efficient traffic light scheduling algorithm in [25] adjusts the finest green phase time at the signalized road intersection based on realistic traffic information. It also considers the emergency vehicle presence in green phase time adjustment and assists for quick emergency vehicle passing. A multi-agent preemptive longest queue first system has been proposed in [26] to handle the emergency vehicle crossings at interrupted intersections. Further, an efficient preemption strategy is selected to diminish the negative impact of preemption on general traffic in [27]. It utilizes the VANET communication through the emergency vehicle path. Thus, it clears the entire route of an emergency vehicle in advance without disturbing the normal traffic flow. The work in [28] utilizes the internet of things technology to facilitate emergency vehicles crossing the intersections quickly. Such a model gathers the EV data along its route periodically and intermittently and provides high priority to the EVs, especially at intersections. A signal priority algorithm in [33] develops a queue length-based green signal activation model in which the signal green phase is extended to the specific road that experiences a high delay. The priority algorithm considers queue length to solve the arrival time issues of the emergency vehicle and reduce the impact on normal vehicles along the emergency vehicle route. The smart emergency vehicle plan model in [30] designs an efficient EV communication model by utilizing app monitoring and a centralized network. The vehicles in the traffic control system have a unique identity number to establish a connection with a centralized server around the traffic signal. The centralized network maintains the vehicular network data, and it plans effective routes to the emergency vehicles. A novel EV preemption method in [31] exploits the advantage of the vehicle to infrastructure communication and vehicle density queue information to manage the traffic light controller. However, it lacks to consider the negative impact of EV preemption on normal traffic.

C. Research Gap and Problem Statement

Numerous emergency vehicle preemption and route selection methods are designed in the existing literature to pass the emergency vehicles quickly to the destination. Most of the emergency vehicle route selection model considers the traffic congestion and route length in the emergency vehicle path discovering. However, an emergency vehicle may be delayed due to the signalized intersections along its selected shortest path in urban scenarios. A minute of emergency vehicle delay causes tremendous loss of lives, and hence, it is crucial to minimize the impact of intersection delay caused due to inefficient traffic light scheduling. With aiming to solve such issues, the later researches utilize efficient preemption methods in which the vehicle and infrastructure communication are used to clear the emergency vehicle path or lane in an advanced manner. Such models allow the emergency vehicles to take high priority at the intersections even the signal is in the red phase. In such situations, there is a chance of accidents due to inexperienced and careless driver behaviors. Therefore, it is crucial to activating the signal preemption at the right time using appropriate scheduling methods. The proposed work attempts to design an efficient emergency vehicle preemption model in which geographic routing and timely traffic light scheduling are used to quickly navigate the emergency vehicles and reduce the impact of normal traffic.

III. Design Overview of SG-TSE

Generally, a less congested route with a short travel time is suggested for emergency vehicles. However, the suggested best routes may be suffered by normal vehicle traffic flows and the hindrances of normal road topology. This work proposes an emergency vehicle pre-emption method with the assistance of vehicular geographic routing and traffic light scheduling to reduce the negative impacts of normal vehicle traffic on emergency vehicle speed. Fig. 1 shows the block diagram of the proposed methodology. Initially, the traffic light and RSU detect the EV on the selected less congested and shortest traveling route based on the routing messages disseminated by EV. The disseminated messages include information about EV presence and speed. Thus, the EV is detected based on the messages. The pre-emption distance measurement is applied to compute the distance between the EV and road intersections in which the normal vehicle traffic is high. Secondly, the multi-criteria-based preclearance is utilized to clear the normal vehicles quickly in the approaching route and minimizes the disturbances associated with emergency vehicle speed. Further, the normal vehicles are cleared rapidly from the approaching route based on green phase adjustment. Moreover, the proposed methodology minimizes the negative impacts on emergency vehicle speed and provides timely help to the public.

A. Network Model

The vehicular network is modeled as a communication graph G (N, E), where N refers to the number of nodes classified into emergency vehicles, non-emergency vehicles, RSUs, and traffic light controllers. It is assumed that the virtual traffic lights are installed at every intersection, referred to as RSUs. The term N refers to the communication link between any two entities. The vehicles in SG-TSE move with the desired speed S. The speed of EV is high than normal vehicles $S_{EV}>S_{NV}$. The EV does not change its speed along its path using geographic routing and traffic light scheduling-based preemption. Every vehicle in the SG-TSE is equipped with GPS, and it updates its location itself. The other vehicles knew the location of emergency vehicles based on dissemination messages. The vehicles are also equipped with On-Board Units (OBUs) for enabling wireless communication. The emergency vehicles disseminate the beacons to the others in the corresponding segment by using OBUs. Each vehicle has a road map for path selection. The shortest and less congested traveling route of EV is suffered by normal network traffic in urban scenarios. To reduce the negative impact of normal vehicles on the EV route, the SG-TSE divides the EV approaching lane into multiple segments $S=\{S_1, S_2,.....S_n\}$. In each segment, a Segment Head (SH) is selected for centric-based geographic routing. Further, the signal head green phase adjust time t_g is computed using the minimum detection distance metric.

B. Segment based Geographic Routing (SGR)

Initially, the emergency vehicle selects the best traveling route with minimum congestion and short travel time [32]. However, the shortest route includes some normal traffic that influences the desired speed of emergency vehicles. Therefore, it is essential to create an alert about the emergency vehicle arrival to reduce the negative impacts of normal vehicles. The TG-TSE utilizes the SGR to inform the emergency vehicle presence and speed to the traffic controllers along its selected path. The SGR exploits segments to accomplish information dissemination segment-based routing. The SGR routing decision depends on vehicle location information, direction, vehicle density, and link quality among the two communicating parties. Initially, the SGR separates the selection path into multiple segments and inaugurates the geographic routing through the head node elected in each segment. The SGR routing decision is based on the segment data with a look ahead of the next segment data. The message is forwarded through the segment heads until it reaches the traffic controller.

Segment Formation: The SGR is based on various routing parameters like location information, direction, link quality, and traffic density of road networks. The main intention of SGR is to quickly inform about its presence to the traffic controller for efficient preemption. The EV routes the messages in two ways. Firstly, the EV straightly informs the traffic controller when the EV and traffic controller is in the same segment. Otherwise, the EV divides the corresponding path into multiple segments based on the location information and road map data for multi-hop forwarding. In Fig. 2, the segment formation of the proposed SGR is depicted. The SGR forms the segments based on the road trajectory and the number of intersections of the EV vehicle path [29].

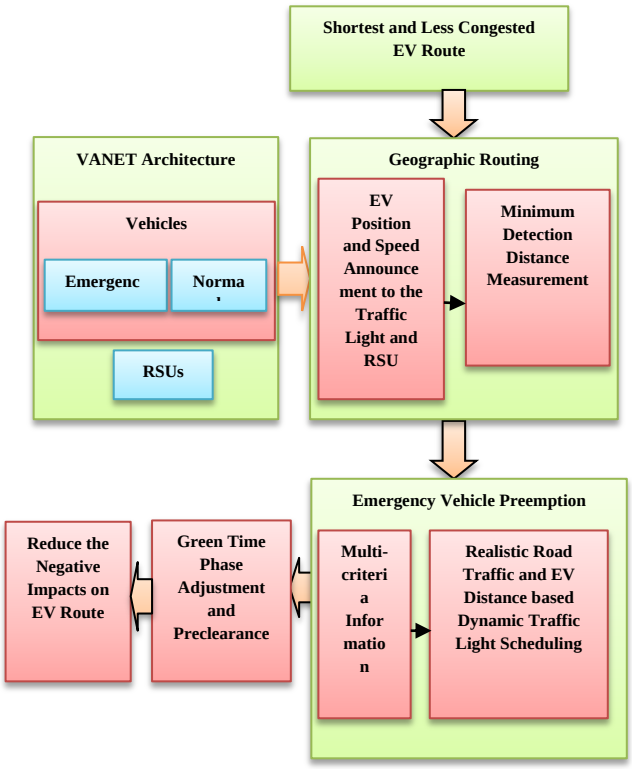


Fig. 1. Block Diagram of Proposed Methodology.

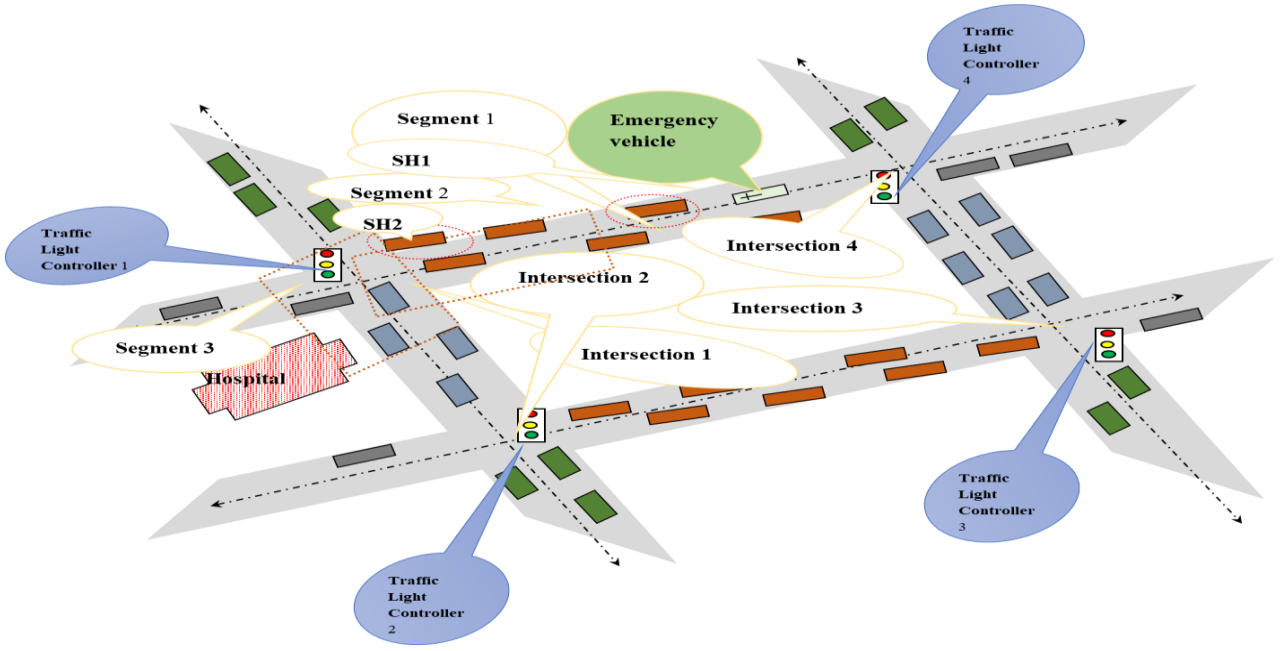


Fig. 2. SGR Segment Formation and Routing.

Segment Head Selection: The segment head is a crucial forwarding node in SGR, as the emergency vehicle arrival is informed through the head to the traffic light controller for preemption. The segment head (SH) selection is initiated after segment formation. Every vehicle announces its position, speed, direction, and link quality through its beacon messages in VANETs. The segment centric distance of a node D(c) is measured using the location information of the node n at t and t+1 time intervals. It is estimated as follows.

$$D(c) = L(t+1) - L(t) \quad (1)$$

The terms $L(t+1)$ and $L(t)$ are location coordinates of (x2, y2) and (x1, y1), at t+1 and t time, respectively. Further, the SGR utilizes the Pythagoras theorem to estimate the location of the SH node, as depicted in equation (2).

$$D_{c\to n} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

More than one nodes are presented near the centric location. Therefore, the SGR considers the link quality and direction parameters in SH selection to speed up the message delivery rate without compromising the reliability. Consequently, the SGR computes a score value for the nodes that are suitable for SH. The first metric is location information, in which the node should have to present near or exactly in the central position of the segments. Further, the progressive distance towards the destination is an essential routing metric in which the vehicles are moving in the same direction of EV is selected as SH. The direction difference score is estimated using the following equation.

$$S_{DD} = \begin{cases} 1; & \text{Same Direction} \\ 0; & \text{Opposite Direction} \end{cases} \quad (3)$$

Finally, the SGR computes the overall score value of the SH nodes using the following equation (4).

$$S_{SH} = \alpha * D_{c\to n} + \beta * S_{DD} + \gamma * L_Q + \varphi * S_v \quad (4)$$

In equation (4), the terms $D_{c\to n}, S_{DD}, L_Q$ snd S_v are the distance, direction, link quality, and speed values of the vehicle suitable for SH. The terms $\alpha, \beta, \gamma,$ and φ are weighting factors. The summation of weighting factors is equal to 1. Finally, the SGR selects the node that has high S_{SH} value as SH. This process is performed in all the segments of the EV traveling path. Further, the EV informs the traffic controller about the arrival through the selected SH vehicles.

C. Dynamic Traffic Light Scheduling and EV Preemption (DTSE)

After receiving the EV arrival information, the traffic light controller in SG-TSE initiates the traffic light scheduling process DTSE. The traffic light controller has to identify the EV in advance with a minimum detection time to pre-clear the approaching lane before the EV arrival and neglect the EV waiting time at the intersection. Hence, efficient scheduling is essential to minimize the impact of the normal vehicle on EV preemption. For traffic light scheduling, the DTSE uses the minimum detection distance metric, which is a type of distance measurement applied between the emergency vehicle location and the traffic light intersection. The controller receives the location and speed information of EV from the SH node, and it starts to calculate the minimum detection distance between the EV from the intersection using equation (5).

$$t \geq t_{switchover} + t_{min} + STI \quad (5)$$

Where the term $t_{switchover}$ is the switchover time of the signal head and the term t_{min} is the discharge time of the signal. The term STI is the safety time interval to pre-clear the vehicles in the approaching lane of EV. The switchover time is the interval of switching the signal state. The discharge time t_{min} is estimated from the average queue length and the queue discharge speed of the EV approach using historical information. The STI value is kept at the constant of 2 s. The t

value and speed information are used to calculate the minimum detection time. Thus, the TSE preemption method inaugurates the preemption phase at the correct time using minimum detection distance. Thus, the TSE avoids the hindrances of EV traveling path and minimizes the impact of the preemption on normal vehicular traffic. Further, the TSE starts the green phase adjustment for quick EV navigation through the corresponding intersection.

1) Green phase adjustment: Based on the distance measurement and multiple criteria like congestions and road conditions, the EV preemption is timely activated in the SG-TSE. A significant parameter is the signal head green time requirement (t_g) on the EV preemption, which is computed as follows.

$$t_g = t - (t_{switchover} + t_{min} + STI) \quad (6)$$

By using the t_g value, the SG-TSE effectively pre-clears the vehicles in the EV approaching lane and assists the EVs to maintain their desired speed in the corresponding intersection. In the coordinated route intersections case, the TSE assumes that the distance between two successive intersections is smaller than the detection time of EV detection distance. The EV detection point suffers the first intersection. To rectify such an issue, the SG-TSE considers the notification period of the discharge time of both consecutive intersections. In other words, the SG-TSE treats the two consecutive intersections as a single intersection. Otherwise, the distance between the two consecutive intersections is high than the minimum detection distance, and the SG-TSE re-estimates the detection distance for the second intersection by using the TSE model for preemption. Moreover, the green phase of the traffic light is scheduled at the correct time according to the emergency vehicle distance from the intersection. The SG-TSE protocol process is explained in algorithm 1.

Algorithm. 1. SG-TSE Protocol Process

//SG-TSE Protocol Process//

Input: Selected less congested and shortest EV travelling route
Methods: SGR and DTSE
Output: EV preemption
SG-TSE Do {
Inputs the EV travelling route and initializes the network;
Starts the SGR and DTSE;

SGR Do {
Initiates the segment formation for geographic routing;
Divides the approaching route into multiple segments;
Elects an SH in each segment based on a multi-criteria value;
Passes the EV arrival messages to the traffic light controller through SHs;
}

DTSE Do {
Measures the minimum detection distance using equation (5);
Initiates green phase adjustment;
Calculates signal head green time requirement (t_g);
Switchovers the green phase at right time;
Pre-clears the EV route;
EV preemption;
}};

IV. Performance Evaluation

The SG-TSE performance is analyzed using NS-2. The performance of the proposed work is compared with existing MDRP [11], PGRP [8], MM-GPSR [9], and Geo-LU [10] for performance evaluation. The simulation parameters are demonstrated in Table I.

TABLE I. Simulation Parameters

Parameter	Value
Simulation Tool	NS-2
Network Area	1000x1000
Number of Nodes	10 to 30
Number of EVs	1-3
Vehicle Communication Range	50m
RSU Communication Range	250m
Routing Protocol	SG-TSE
Traffic Simulator	SUMO
Transport Protocol	UDP, CTP
Speed of Normal vehicles	45 Km/hr
Speed of EVs	60 Km/hr
Propagation Model	Two Ray Ground
Simulation Time	2 Seconds
Application Type	CBR
Packet Size	128 Bytes
Data Rate	3 Mbps

A. Performance Metrics

The efficacy of SG-TSE is analyzed in terms of packet delivery ratio, overhead, throughput, delay, and EV preemption speed.

- Packet Delivery Ratio (PDR): It is the percentage of successfully delivered packets to the total number of generated packets.
- Overhead: It is the number of extra packets used to perform network operations.
- Throughput: It is the rate of data delivery.
- Delay: It is the time taken to deliver a packet from a source to a destination.
- EV preemption Speed: It is the speed maintained by the EVs on the approaching lane.

B. Simulation Results

Fig. 3 portrays the comparative PDR results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU observed under different node density scenarios. From the results of Fig. 3, the SG-TSE increases the PDR from vehicle density 10 to 20, whereas the PDR is decreased after the point of 20 node density scenario. It is caused due to the adequate number of vehicles offer better connectivity to packet forwarding, and the high number of nodes competes to access similar links, resulting in some packet loss in the network. For example, the

SG-TSE accomplishes 99%, 99.6%, and 98% of PDR for 10, 20, and 30 node densities, respectively. However, the PDR of SG-TSE is higher than the other four geographic routing protocols from 10 to 25 node densities. For instance, the SG-TSE improves the PDR by 1.4%, 3%, 1.9%, and 1.5% than the existing MDRP, PGRP, MM-GPSR, and Geo-LU protocols, respectively, when 10 nodes are present in the network. The figure shows that the SG-TSE and Geo-LU accomplish 98% and 99.4% of PDR in 30 node scenarios. The link utility aware geographic router node selection in Geo-LU improves the PDR than the proposed SG-TSE. However, the SG-TSE attains better PDR values than the MEDRP, PGRP, and MM-GPSR protocols, when 30 numbers of vehicles are presented in the network.

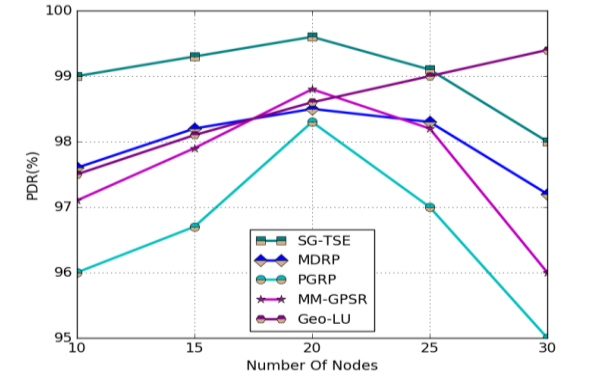


Fig. 3. Number of Nodes vs. PDR.

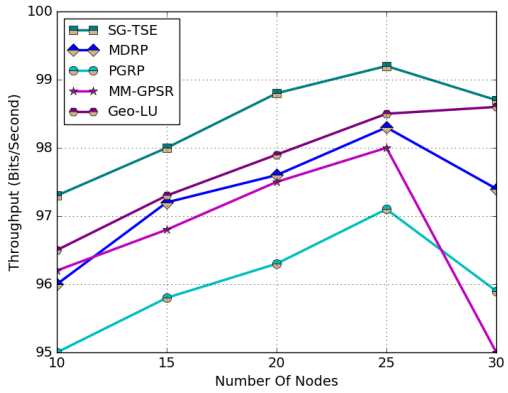


Fig. 4. Number of Nodes vs. Throughput.

Fig. 4 demonstrates the throughput comparison results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU geographic routing protocols. All protocols increase the throughput from the point of 10 node densities to 25 node densities. For example, the SG-TSE obtains 97.3% and 99.2% of throughput under 10 and 25 number of nodes scenario, respectively. It is high in the range of 1.9%. After point 25, the throughput value of SG-TSE is decreased, as the link may fail due to high competition nodes. The SG-TSE minimizes the throughput by 0.5% after the point 25 number of nodes. However, the SG-TSE attains better throughput performance than the other existing protocols. The main reason is that the SG-TSE selects the best SH nodes for data forwarding using multiple parameters that are position, speed, direction, and link quality. Thus, it maximizes the throughput even the network is highly congested. For instance, the SG-TSE improves the throughput by 1.3%, 2.3%, 1.1%, and 0.8% than the existing MDRP, PGRP, MM-GPSR, Geo-LU when 10 numbers of vehicles are present in the network.

Fig. 5 shows the delay results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU obtained with different numbers of nodes. The SG-TSE escalates the delay by adjusting the number of nodes from 10 to 30. The main reason is that the nodes have to retransmit the packets frequently due to high packet loss under a high vehicle density setting. For example, the delay of SG-TSE is 0.35 seconds and 1.7 seconds for 10 and 30 nodes scenarios. However, the delay performance of SG-TSE is better than the other four geographic routing methods. The segment-based geographic router selection assists the SG-TSE to diminish the delay even the highly congested network. Also, the SG-TSE selects the best SH node by taking into account the position, speed, direction, and link quality parameters. Thus, it minimizes the delay in packet delivery and motivates the traffic lights for timely green phase activation, resulting in minimum EV arrival delay. For instance, the SG-TSE reduces the delay by 30%, 58.8%, 78.1%, and 63.2% than the MDRP, PGRP, MM-GPSR, and Geo-LU under 10 nodes scenario. It is varied by 69.6%, 51.4%, 63%, and 57.5% for 30 number of nodes scenario.

Fig. 6 illustrates the overhead comparison results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU by adjusting the number of nodes from 10 to 30. All protocols increase the overhead by varying the node density from low to high. This is caused due to the utilization of a high number of control packets under the high-density scenario that maximizes the overhead in the network. For instance, the SG-TSE accomplishes 110 and 310 packets of overhead for 10 and 30 numbers of nodes, respectively. However, the SG-TSE diminishes the overhead than the MDRP, PGRP, MM-GPSR, and Geo-LU, as demonstrated in Fig. 6. The main reason is that the segment-based geographic routing along the EV route limits the control packets within the segment, resulting in minimum overhead. For example, when 30 nodes are present in the network, the SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU attain 310, 350, 375, 410, and 375 packets of overhead, respectively.

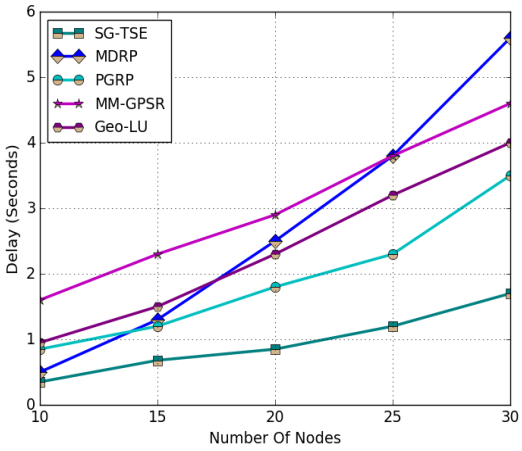


Fig. 5. Number of Nodes vs. Delay.

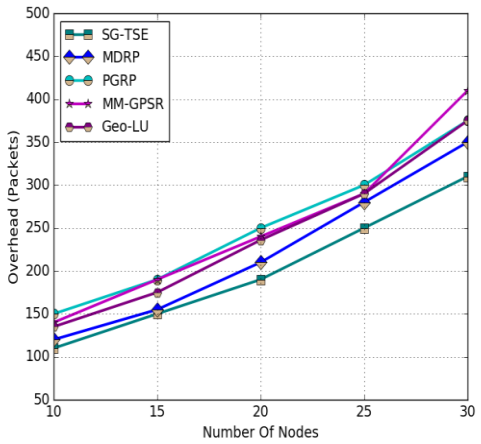


Fig. 6. Number of Nodes vs. Overhead.

The EV preemption speed results of SG-TSE obtained by varying the Number of Intersections (NoI) are depicted in Fig. 7. The EV preemption speed is diminished when varying the vehicle density from 10 to 30. The main reason is that the traffic controller requires a long time to pre-clear the vehicles under a high vehicle density scenario, impacting EV preemption speed. For instance, the speed of EV is 60 and 55 Km/Hr for 10 and 30 node density scenarios under one intersection. The SG-TSE design considers the NoI into account, and there is a need to recalculate the minimum detection distance when more than one intersection is presented along the EV route. Thus, it creates some impact on EV speed. For instance, the speed of EV is diminished by 8.3% for 20 nodes with two NoI scenarios. However, the timely green phase activation with minimum detection distance measurement in SG-TSE speeds up the EV at intersections rapidly and minimizes the arrival delay even the network is congested.

Fig. 8 obtains the PDR comparative results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU by varying the vehicle speed from 20 to 60Km/hr. The results show that the SG-TSE decreases the PDR by adjusting the vehicle speed from 20 to 60 Km/Hr. It is caused due to the frequent link disconnections of high-speed vehicles in the network. For example, the SG-TSE accomplishes 99% and 97.1% of PDR when the nodes move with speed 20 Km/Hr and 60 Km/Hr, respectively. However, the SG-TSE obtain better PDR results by selecting the SH nodes with multi-criteria information like position, speed, direction, and link quality than the existing protocols. Thus, it effectively handles the frequent link disconnections and boosts the PDR even when vehicles move at high speed. For instance, the SG-TSE increases the PDR by 0.3%, 1.1%, 4.1%, and 2.6% than the existing MDRP, PGRP, MM-GPSR, and Geo-LU protocols under a high vehicle speed scenario of 60 Km/hr.

Fig. 9 shows the delay results of SG-TSE, MDRP, PGRP, MM-GPSR, and Geo-LU protocols. The results are accomplished by adjusting the speed values of the vehicles from 20 to 60 Km/hr. All protocols increase the delay by varying the vehicle speed from low to high. The main reason is that the nodes have to retransmit the packets frequently due to link disconnections. For example, the delay of SG-TSE is 0.5 seconds and 2.3 seconds for 20 and 60 Km/Hr of vehicle speeds, respectively. However, the delay performance of SG-TSE is better than the other existing MDRP, PGRP, MM-GPSR, and Geo-LU protocols. For instance, the SG-TSE reduces the delay by 54.5%, 37.5%, 58.3%, and 66.7% than the MDRP, PGRP, MM-GPSR, and Geo-LU when the vehicles move with 20 Km/Hr speed. The reason is that the multi-criteria-based SH router selection assists the SG-TSE to diminish the delay even the highly congested network. Thus, it maximizes the green time activation accuracy. Moreover, the SG-TSE minimizes the negative impact on normal vehicles owing to EV preemption by activating the green phase at the right time.

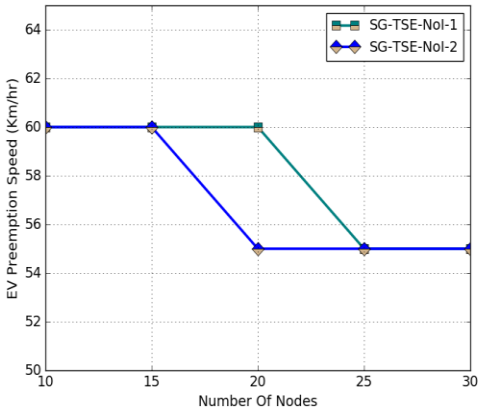


Fig. 7. Number of Nodes vs. EV Preemption Speed.

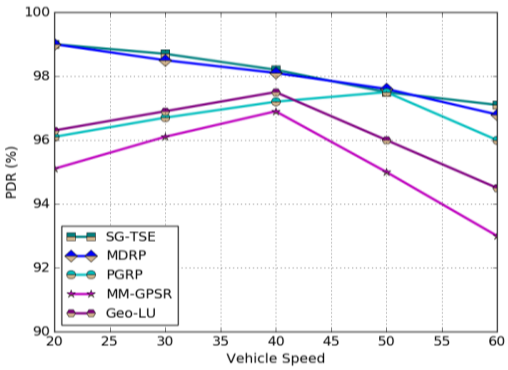


Fig. 8. Vehicle Speed vs. PDR

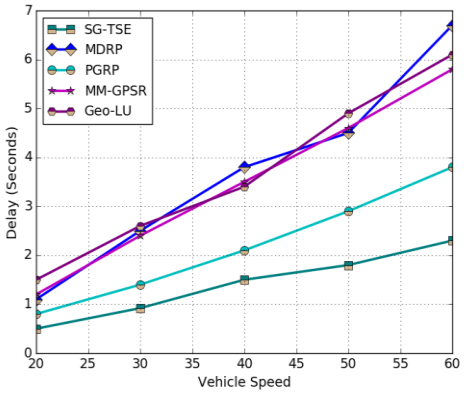


Fig. 9. Vehicle Speed vs. Delay.

V. Conclusion

In this paper, a novel EV preemption method SG-TSE has been proposed to reduce the negative impacts of normal vehicles on EV speed. To achieve the objective, the SG-TSE includes two mechanisms that are SGR and DTSE. By dividing the approaching route into multiple segments and performing segment-based geographic routing, the SGR instructs the traffic light controller about EV arrival. The minimum detection distance measurement based green phase adjustment in DTSE reduces the negative impacts of normal traffic on EV speed and neglects the EV preemption negative effects on normal traffic. Moreover, the NS-2 based simulation depicts the effectiveness of the proposed SG-TSE with different performance metrics like PDR, overhead, delay, throughput, and EV preemption speed. From the results, the EV maintains its speed in the approaching lane without disturbing the normal traffic conditions.

References

[1] Zeadally, S., Hunt, R., Chen, YS. et al, "Vehicular ad hoc networks (VANETS): status, results, and challenges", Telecommun Syst, pp. 217–241, 2012, https://doi.org/10.1007/s11235-010-9400-5.

[2] Michael Lee, Travis Atkison, "VANET applications: Past, present, and future, Vehicular Communications", Vol. 28, 2021.

[3] V. Paruchuri, "Adaptive Preemption of Traffic for Emergency Vehicles", UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim), pp. 45-49, 2017, doi: 10.1109/UKSim.2017.34.

[4] Almuraykhi, K. M., &Akhlaq, M, "STLS: Smart Traffic Lights System for Emergency Response Vehicles", International Conference on Computer and Information Sciences (ICCIS), 2019.

[5] W. Min, L. Yu, P. Chen, M. Zhang, Y. Liu and J. Wang, "On-Demand Greenwave for Emergency Vehicles in a Time-Varying Road Network With Uncertainties", in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 7, pp. 3056-3068, July 2020, doi: 10.1109/TITS.2019.2923802.

[6] R. Anil, M. Satyakumar and A. Salim, "Emergency Vehicle Signal Pre-emption System for Heterogeneous Traffic Condition : A Case Study in Trivandrum City", 2019 4th International Conference on Intelligent Transportation Engineering (ICITE), 2019, pp. 306-310, doi: 10.1109/ICITE.2019.8880151.

[7] SouaadBoussoufa-Lahlah, FouziSemchedine, and LouizaBouallouche-Medjkoune, "Geographic routing protocols for Vehicular Ad hoc NETworks (VANETs): A survey", Vehicular Communications, Volume 11, 2018.

[8] Karimi, R., &Shokrollahi, S, "PGRP: Predictive geographic routing protocol for VANETs", Computer Networks, 2018.

[9] Yang, X., Li, M., Qian, Z., & Di, T, "Improvement of GPSR Protocol in Vehicular Ad Hoc Network", IEEE Access, pp. 39515–39524, 2018.

[10] Alzamzami, O., &Mahgoub, I, "Link Utility Aware Geographic Routing for Urban VANETs using Two-Hop Neighbor Information", Ad Hoc Networks, 2020.

[11] MeenaakshiSundhari, R. P., Murali, L., Baskar, S., & Shakeel, P. M, "MDRP: Message dissemination with re-route planning method for emergency vehicle information exchange", Peer-to-Peer Networking and Applications, 2020.

[12] P. Devi and S. Anila, "Intelligent Ambulance with Automatic Traffic Control", 2020 International Conference on Computing and Information Technology (ICCIT-1441), pp. 1-4, 2020, doi: 10.1109/ICCIT-144147971.2020.9213796.

[13] Subash Humagain, Roopak Sinha, Edmund Lai & Prakash Ranjitkar, "A systematic review of route optimisation and pre-emption methods for emergency vehicles", Transport Reviews, Vol. 40, No. 1, pp. 35-53, 2020, DOI: 10.1080/01441647.2019.1649319.

[14] Oza, P., &Chantem, T, "Timely and Non-Disruptive Response of Emergency Vehicles: A Real-Time Approach", 29th International Conference on Real-Time Networks and Systems, 2021.

[15] Wenwen Kang, Gang Xiong, YishengLv, Xisong Dong, Fenghua Zhu, &Qingjie Kong, "Traffic signal coordination for emergency vehicles", 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2014, doi:10.1109/itsc.2014.6957683.

[16] Ziling Zeng; Wen Yi; Shuaian Wang; and Xiaobo Qu,"Emergency Vehicle Routing in Urban Road Networks with Multistakeholder Cooperation", Journal of Transportation Engineering, Part A: Systems, Vol. 147, No. 10, 2021.

[17] S. Humagain and R. Sinha, "Dynamic Prioritization of Emergency Vehicles For Self-Organizing Traffic using VTL+EV", IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society, pp. 789-794, 2020, doi: 10.1109/IECON43393.2020.9254313.

[18] A. S. Eltayeb, H. O. Almubarak and T. A. Attia, "A GPS based traffic light pre-emption control system for emergency vehicles", International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), pp. 724-729, 2013, doi: 10.1109/ICCEEE.2013.6634030.

[19] Hamed Noori, Liping Fu, and SajadShiravi, "A Connected Vehicle Based Traffic Signal Control Strategy for Emergency Vehicle Preemption", 2016.

[20] Asaduzzaman, M., &Vidyasankar, K, "A Priority Algorithm to Control the Traffic Signal for Emergency Vehicles", IEEE 86th Vehicular Technology Conference (VTC-Fall), 2017, doi:10.1109/vtcfall.2017.8288364.

[21] Wu, J., Kulcsár, B., Ahn, S., & Qu, X, "Emergency vehicle lane pre-clearing: From microscopic cooperation to routing decision making", Transportation Research Part B: Methodological, pp. 223–239, 2020, doi:10.1016/j.trb.2020.09.011.

[22] VítObrusník, Ivo Herman, ZdeněkHurák, "Queue discharge-based emergency vehicle traffic signal preemption", The research was funded by Technology Agency of the Czech Republic within the program Epsilon, the project TH03010155.,IFAC-PapersOnLine, Vol. 53, No. 2, 2020.

[23] Mu, H., Song, Y., & Liu, L, "Route-Based Signal Preemption Control of Emergency Vehicle", Journal of Control Science and Engineering, 2018, 1–11. doi:10.1155/2018/1024382.

[24] Rathore, N., Jain, P. K., &Parida, M, "A Routing Model for Emergency Vehicles using the Real Time Traffic Data". IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2018.

[25] Younes, M.B., Boukerche, A, "An efficient dynamic traffic light scheduling algorithm considering emergency vehicles for intelligent transportation systems", Wireless Networks, pp. 2451–2463, 2018.

[26] Louati, A., Elkosantini, S., Darmoul, S., &Louati, H, "Multi-agent preemptive longest queue first system to manage the crossing of emergency vehicles at interrupted intersections", European Transport Research Review, Vol. 10, No. 2, 2018.

[27] Shaaban, K., Khan, M. A., Hamila, R., & Ghanim, M, "A Strategy for Emergency Vehicle Preemption and Route Selection", Arabian Journal for Science and Engineering, 2019.

[28] Talebi, M., & Sabaei, M, "Smartly, reduce the latency of high-priority vehicles using loT technology" In IEEE 29th Iranian Conference on Electrical Engineering (ICEE), pp. pp. 514-520, 2021.

[29] Kamble, Shridevi J., and Manjunath R. Kounte. "On Road Intelligent Vehicle Path Predication and Clustering using Machine Learning Approach." 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, 2019.

[30] M. V, A. V and S. Allirani, “Smart Emergency Vehicle Plan with App Monitoring and Centralized Network”, 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1-5, 2021.

[31] Obrusník, V., Herman, I., & Hurák, Z, “Queue discharge-based emergency vehicle traffic signal preemption”, IFAC-PapersOnLine, Vol. 53, No. 2, pp. 14997-15002, 2020.

[32] Kamble, Shridevi Jeevan, and Manjunath R. Kounte. "Machine learning approach on traffic congestion monitoring system in internet of vehicles." Procedia Computer Science 171 (2020): 2235-2241.

[33] SEO, J., KIM, D., KANG, S., LEE, J., & LEE, S, “Development of Signal Priority Algorithm considering a Recovery Time for Emergency Vehicle”, 2021.

Detection of Data Leaks through Large Scale Distributed Query Processing using Machine Learning

Kiranmai MVSV[1]
Assistant Professor (C) and Research Scholar
Department of CSE, UCEK, JNTUK
Kakinada, India

D Haritha[2]
Professor
Department of CSE, UCEK, JNTUK
Kakinada, India

***Abstract*—With the growth in the distributed data processing and data being the fuel for each of the processes, the query processes of the data are expected to be significantly lower. Hence, the distribution of the data is highly expected and during the distributing of the data, the chances for data leakage increases to a significant extend. The data leakage problems are not generally caused by intentional errors, rather this is caused by the higher visibility of the data over multiple clusters. Henceforth, the detection process is also very critical. Many of the parallel research attempts have demonstrated various methods for the detection and as well as the prevention methods. The works in the direction of the detection of the data leaks are highly dependent either on the historical information of the leaks or depends on the contextual importance of the data. In both the cases, the outcomes of the detection process accuracy cannot be ensured. In the other hand, the preventive measures can also turn into a reactive process for detection by reversing the principles proposed in these research outcomes, but the computational complexities are significantly higher. Thus, this work proposes a novel strategy for detection of the data leakages after the data distribution during the query processing events. This work proposes an initial Occurrence Based Rule Set Extraction method using Adaptive Threshold for generating the rulesets, further for reducing the time complexity and reducing the loss of dataset attribute information, this work introduces yet another algorithm for Dynamic Inference-based Rule Set Reduction. After the inferences are generated, finally this work deploys the Attribute Subset Equivalence-based Leak Detection mechanism for final detection of the clusters with data leaks. This work demonstrates nearly 89% accuracy for the detection process.**

Keywords*—Distributed query processing; distributed data leak; data leak detection; attribute subset equivalence; dynamic inference; adaptive threshold model introduction

I. Introduction

Leakage of the data during centralized or distributed query processing environment is one of the primary concerns in the recent cyber security domain. Number of professionals and researchers have aimed to define the problem of data leakage in the past decade and aimed to provide solutions considering the higher number of cases are reported for data leakage, which leads to data breaches. As per the reports by Breach Level index [1], a gigantic volume of data, as close to 4.5 Billion, are compromised only in 2018.

The empirical study by C. Missaoui et al. [2] have critically analysed the data leakage situations and formulated the correct sequence of events, which might lead to the data breach. The common reason for data leakage, as per this study, showcased that the use of unused data stored in the centralized or distributed storage solutions leads to data leakage. The example also demonstrated that the data, which is left unattended or not in motion, must be revoked or should be distributed with a specified life time and expiry to be restricted for reuse.

Yet another case study presented by IBM showcased that the cost of recovery of stolen or leaked data can be as high as 3.86 Billion dollars [3].

Thus, the demand for data protection and data leakage detection is one of the primary demands from the current research trends and must be addressed. Also, the parallel research outcomes have demonstrated that the security concerns are higher in case of the distributed data sources for the obvious reasons. Thus, this work focuses on the distributed query processing environments to detect the data leaks.

The rest of the work is furnished such as in Section – II, the fundamentals of the distributed query processing is discussed and understood, in Section – III, the outcomes from the parallel research attempts are analysed, in Section – IV, the problem is formulated using mathematical model, in Section – V, the proposed solution is again formulated using mathematical modelling method, in Section – VI, the proposed algorithms are furnished and elaborated, in Section – VII, the obtained results are analysed, in Section – VIII, the comparative analysis is presented and in Ssection – IX, the final conclusion of the research is presented.

II. Distributed Query Processing Fundamentals

In this section of the work, the fundamental of distributed query processing method is discussed in order to realize the recent outcomes from the parallel research attempts.

Assuming that, every query Q is the collection of relations and predicates as R and P respectively. This relation can be formulated as,

$$Q \rightarrow \sum R . \sum P \tag{1}$$

Or for example, the above relation can be re-written as,

$$Q \rightarrow [R_1 \cap R_2] \cup R_3 \tag{2}$$

Also, assuming that the relation R, is distributed over the cluster set C[], where each and every cluster can be identified as C_X for "n" number of clusters. Thus, can be represented as,

$$C[] \leftarrow \sum_{x=1}^{n} C_x \tag{3}$$

And,

$$R[] \rightarrow C[] \tag{4}$$

Further, assuming the data distribution is as follows,

$$\{R_1, R_2, R_3\} :: \{\langle C_1, C_2, C_3 \rangle, \langle C_2, C_3 \rangle, C_3\} \tag{5}$$

During the query processing for distributed systems, the primary objective is to find the allocation and minimum distribution of the relations. As,

$$\{R_1, R_2, R_3\} :: \{C_3\} \tag{6}$$

Further, the optimization possibilities for predicates can also be identified and performed. Once the optimization task is completed, the result of the query can be return to the queue buffer.

Henceforth, this fundamental understanding of the distributed query processing shall help in realizing the parallel research outcomes, which are discussed in the next section of the work.

III. Parallel Research Outcomes: Survey

After the fundamental understanding of the distributed query processing and chances for the data leakages, in this section of the work, the parallel research outcomes are discussed.

In the recent years, multiple organizations have aimed to provide the complete solution for the detection and up to some extend prevention of the data leakages problems. The primary working principles of these systems are to manual perform exhaustive search with the previously leaked data on the existing shared data for finding the leakage. The point to be mentioned here is that, the data leakage may not be an intentional issue every time. Many of the times, it is been observed that, the wrong distribution of the data attributes leads to the data leakages. Thus, the search option of the previous information may demonstrate machine learning characteristics, but eventually leads to failure in case of insufficient previous or historical data; thus, the other approaches getting popularity in the practice.

One of the major benchmarks in the domain of data leakage detection was the research outcome by P. Papadimitriou et al. [4]. This work demonstrates the data leakage detection by introducing watermarking methods to identify the source of leakage. However, this method was highly criticised by many other researchers due to the higher complexity of the computational models. Also, the size of the actual data, which is distributed, increases to a significant extend because of the replication for each watermarking information and the second issue was that the digital watermarking process is fragile as because of the pre-processing methods used by many algorithms can lead to the loss of watermarks and making the complete process again vulnerable.

The other parallel research outcome by L. Cheng et al. [5] has demonstrated another method to data leakage problem solution. This method demonstrates an approach to classify the content based on the sensitivity of the information and manage the distribution of the higher sensitive data with maximum care. This work is also criticised by the parallel researchers due to the facts that, firstly, the data sensitivity also depends on the context of the data, which is highly variable in all the instances of the query, and secondly, the higher time complexity and chances of low data visibility is also a challenge.

Further, in the complete other direction from these solutions, the work by S. Liu et al. [6] have demonstrated and listed the principles of data leakage preventions. This direction, in contract to the detection of the data leakage, promotes the prevention methods. Elaborating this fact, the work by B. Hauer et al. [7] have also showcased the functional rules for making the data leakage preventions. The set of mentioned rules can also be reversed to identify the leakage sources. Nevertheless, it is natural to realize that, the computational time complexity can be very high for this detection method and also for the distributed environments, this method is prune to errors. The next method is one of the extensions, proposed by T. Malderle et al. [8]. This work elaborates the mechanisms for collecting evidences of the data leakage and further validates the evidences against the prevention principles.

The data leakages are not only limited into the scope of centralized data, rather also extended in higher scale for distributed data. The work by J. Schütte et al. [9] has elaborated on the challenges of data leakage for distributed mobility devices.

The work by S. Trabelsi et al. [10,11,12] provides the summery of the challenges and failure points of each of the above-mentioned mechanisms.

Henceforth, in order to provide the solution to data leakage detection, in the next section of the work, the mathematical model of the actual problem is furnished for better identification of the solution possibilities.

IV. Problem Formulation

After the fundamental understanding of the query processing and the review of the parallel research outcomes, in this section of the work the research problem is formulated.

Assuming that the complete data schema is denoted as DSC[] and every attribute is denoted as AR_X. Thus, for n number of attributes, the total relationship can be formulated as,

$$DSC[] \leftarrow \sum_{i=0}^{n} AR_i \tag{7}$$

Also, one of the attributes in the total attribute set must be identified as the class variable and can be denoted as AR_C. This can be formulated as,

$$\prod_{i=0}^{n} AR_i \cup \prod_{j=1}^{n-1} AR_j \rightarrow AR_C \tag{8}$$

It is natural to realize that the class variable or the class attribute can also be retrieved using other attributes as well and can be denoted as,

$$\prod_{x=0}^{n} AR_x \cup \prod_{y=1}^{n-1} AR_y \rightarrow AR_C \tag{9}$$

In case of a distributed query processing environment, the data is expected to be distributed over multiple clusters, denoted as C[] and each and every cluster can be identified as, C_i. Thus, this relation for k number of clusters can be identified as,

$$C[] \leftarrow \sum_{i=1}^{k} C_i \tag{10}$$

The data is expected to be distributed over the clusters from the initial schema sets and can be realized as,

$$\prod_{i=1}^{n} DSC[i] \Rightarrow \prod_{j=1}^{k} C[j] \tag{11}$$

This can be re-written as,

$$\prod_{i=1}^{n} AR_i \Rightarrow \prod_{j=1}^{k} C[j] \tag{12}$$

It is often to be realized that during the query processing, the similar information can be generated from different attributes and the data leaks can happen.

For example, as the attribute sets AR_i, AR_x and AR_j, AR_y can contain similar information, thus these sets if become part of same clusters during data distribution, then the data leakage is obvious and cannot be prevented.

$$\langle AR_i, AR_x \rangle \neq C_\alpha \tag{13}$$

And

$$\langle AR_j, AR_y \rangle \neq C_\beta \tag{14}$$

Or,

$$C_\alpha \neq C_\beta \tag{15}$$

Henceforth, detection of the data leakage from the distributed schema is the identified problem to be solved. In the next section of this work, the proposed mathematical model for the data leak detection is proposed.

V. Parallel Research Outcomes: Survey

After the fundamental understanding of the parallel research outcomes and the formulation of the problem, in this section of the work, the mathematical model for solution is elaborated.

Assuming that the complete data schema is denoted as DSC[] and every attribute is denoted as AR_X. Thus, for n number of attributes, the total relationship can be formulated as,

$$DSC[] \leftarrow \sum_{i=0}^{n} AR_i \tag{16}$$

Also, one of the attributes in the total attribute set must be identified as the class variable and can be denoted as AR_C. This can be formulated as,

$$\prod_{i=0}^{n} AR_i \cup \prod_{j=1}^{n-1} AR_j \rightarrow AR_C \tag{17}$$

Further analysing the item set frequency for each data item sets, the ruleset, R[] can be generated for "d" number rules and each rule in the ruleset can be considered as R_X. Thus, this relation can be formulated as,

$$R[] \leftarrow \sum_{i=1}^{d} R_i \tag{18}$$

Also, assuming that two different rules, R_X and R_Y, implies the same class variable data instance, AR_C, as,

$$R_X \rightarrow AR_C \tag{19}$$

And,

$$R_Y \rightarrow AR_C \tag{20}$$

However, the rules can contain different attribute sets, as,

$$R_X = \langle AR_i, AR_{i+1}, AR_{i+2}, ...AR_n \rangle \tag{21}$$

Or,

$$R_X = AR'[] \tag{22}$$

And,

$$R_Y = \langle AR_j, AR_{j+1}, AR_{j+2}, ...AR_m \rangle \tag{23}$$

Or,

$$R_Y = AR''[] \tag{24}$$

Finally, in order to detect the data leak based on the large-scale distributed query processing, both the attribute sets, must not coexist on the same cluster, as formulated as,

$$(AR'[], AR''[]) \rightarrow C_{\alpha} \tag{25}$$

If the above situation is detected, then the data leakage can occur, and the security challenges can be increased.

Further, in the next section of this work, based on the problem formulation, the proposed algorithms are furnished and elaborated.

VI. Proposed Algorithms

Furthermore, in this section of the work, based on the proposed mathematical model of the solution in the previous section, the proposed algorithms are furnished.

The first algorithm is designed to generate the rulesets from the given schema and the dataset items. The algorithm is furnished here:

Algorithm - 1: Occurrence Based Rule Set Extraction using Adaptive Threshold Model (**OBRSE-ATM**)

Input: {Read the schema definition for attribute sets, A[] and Class variable, ARC}
Output: {Rulesets, FR}
Process:
Step - 1. Accept the list of attributes as A[]
Step - 2. Accept the class variable as ARC
Step - 3. For each item sets
  a. For each attribute value as A[i]
    i. If A[i] and A[i+1] generates ARC[i]
    ii. Then, count the number of occurrence as O[i] and Rule[i] = A[i] and A[i+1]
  b. End
Step - 4. For each occurrence as O[i]
  a. Calculate the mean as OM = {Sum(O[])}/{Count(O[])}
  b. Calculate the position of the OM as O[k]
  c. Calculate the adaptive threshold as AT = K/{Count(O[])}
Step - 5. For each occurrence as O[j]
  a. If O[j] > OM * AT
  b. Then, Accept R[j] as final rule and add to FR[i]
Step - 6. Report FR[]

To instigate the best principles dependent on a given perception, RULES family start by choosing (isolating) a seed guide to assemble a standard, condition by condition. The standard that covers the best models and the least negative models are picked as the best principle of the present seed model. It permits the best guideline to cover some negative guides to deal with the expansion adaptability and lessen the over fitting issue and uproarious information in the standard enlistment.

The second algorithm is designed to reduce the redundant rule sets and build the final reduced rulesets.

This articulation expresses that at whatever point over the span of some legitimate inference the given premises have been gotten, the predetermined end can be underestimated too. The specific conventional language that is utilized to portray the two premises and ends relies upon the real setting of the determinations.

Algorithm - 2: Dynamic Inference-based Rule Set Reduction (**DI-RSR**)

Input: {Rulesets, FR}
Output: {Reduced Rulesets, FRR}
Process:
Step - 1. Accept the rule sets as FR
Step - 2. For each FR[i]
  a. Generate the attribute set as AR[j]
  b. If AR[j] is subset of AR[j]
  c. Then, remove the AR[j] and FR[i]
  d. Else, Keep FR[i] into FRR[k]
Step - 3. Build the final rule set as FRR[]

The final algorithm is built for detecting the leaks in the distributed query situations and as furnished here.

Algorithm - 3: Attribute Subset Equivalence-based Leak Detection (**ASLD**)

Input: {Final Rulesets, FRR and Cluster Sets, CS}
Output: {Leaked Clusters, LC}
Process:
Step - 1. Accept the final rule sets as FRR[]
Step - 2. Accept the cluster distributions CS[]
Step - 3. For each FRR[i]
  a. Identify the attribute sets as AR[j]
  b. For each AR[j]
    i. If AR[j] infers to ARC[k] and AR[j+1] infers to ARC[k]
    ii. Then, Check AR[j] and AR[j+1] cluster allocation
      1. If AR[j] -> CS[r] and AR[j+1] -> CS[r]
      2. Then, detect data leakage at CS[r] and CS[r]->LC[]
      3. Else, Continue
Step - 4. Report the final leaked clusters as LC[]

Quite a bit of science is grounded in the investigation of equivalences, and request relations. Cross section hypothesis catches the scientific structure of request relations. Despite the fact that equality relations are as omnipresent in arithmetic as request relations, the mathematical structure of equivalences isn't too known as that of requests.

The working flow of the proposed algorithms as a framework is furnished here [Fig. 1].

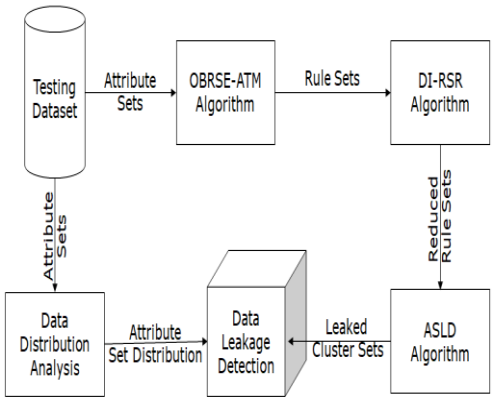


Fig. 1. Working Model of the Proposed Algorithms.

The obtained results from the proposed algorithms are highly satisfactory and discussed in the next section of the work.

VII. Result and Discussion

After the detailed discussed on the mathematical model and the proposed algorithms, in this section of the work, the obtained output from the algorithms are discussed here.

Firstly, the rule extraction results are discussed [Table I].

TABLE I. Rule Extraction Results

Test Number	Number of Rules Extracted	Time Complexity (Sec)
Test Run - 1	329	8.065
Test Run - 2	321	8.301
Test Run - 3	303	8.199
Test Run - 4	326	8.802

The number of extracted rules is the indications of the detailed understanding and deep consideration of all the attribute sets from the dataset. The greater number of rules in this phase of the result defines that most of the attributes are considered and further detection of the leaks can be performed more efficiently.

The results are analysed graphically here [Fig. 2, Fig. 3].

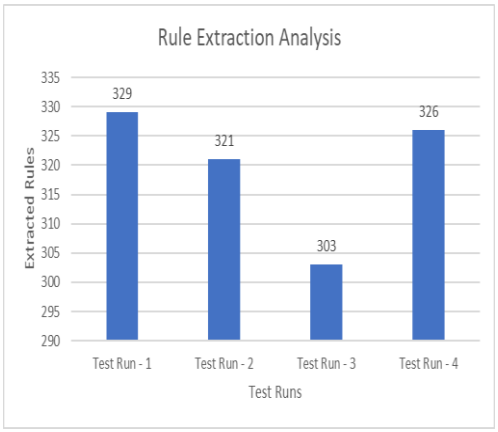


Fig. 2. Rule Extraction Analysis.

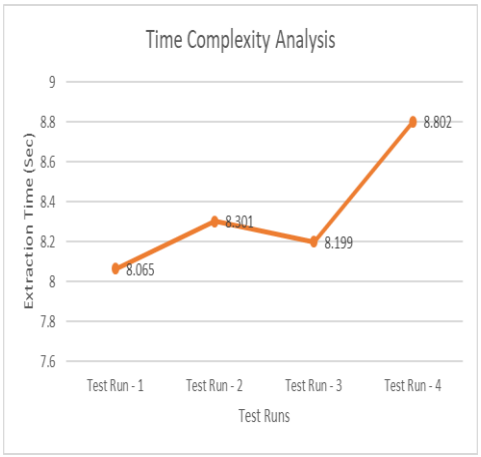


Fig. 3. Rule Extraction Time Complexity Analysis.

However, a greater number of rules can lead to higher time complexity for detection of data leakages. Thus, the reduction of the rulesets is highly expected here. Also, during the rule set reduction process, the point of caution is to maintain the attribute inference properties and relations, which will be helpful in deep detection of the data leakages. Henceforth, the rule set reduction results are discussed here [Table II].

TABLE II. Rule Reduction Results

Test Number	Number of Rules Extracted	Number of Rules after Reduction	Percentage of Reduction (%)	Time Complexity (Sec)
Test Run - 1	329	136	58.66	0.118
Test Run - 2	321	133	58.57	0.120
Test Run - 3	303	126	58.42	0.121
Test Run - 4	326	135	58.59	0.164

The results are analysed graphically here [Fig. 4, Fig. 5, Fig. 6].

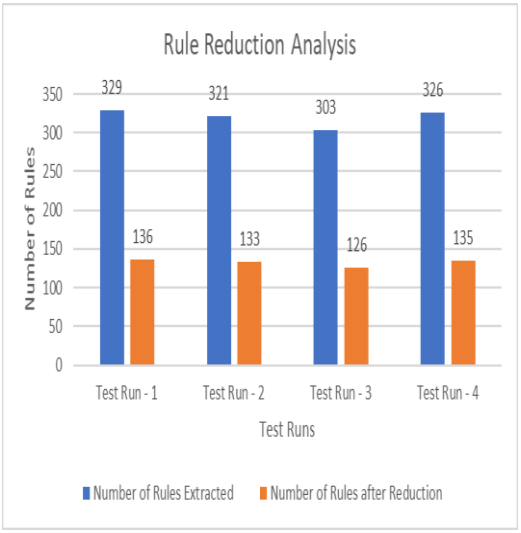


Fig. 4. Rule Reduction Analysis.

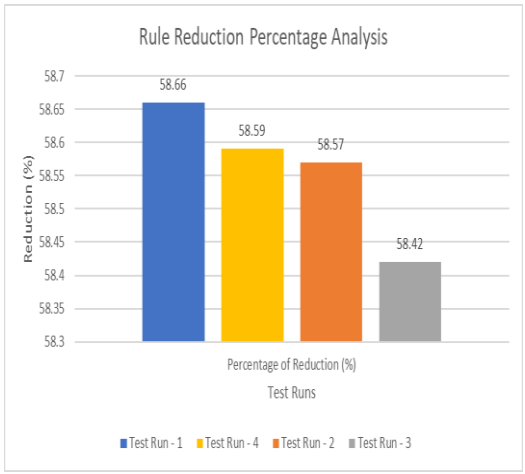


Fig. 5. Rule Reduction Percentage Analysis.

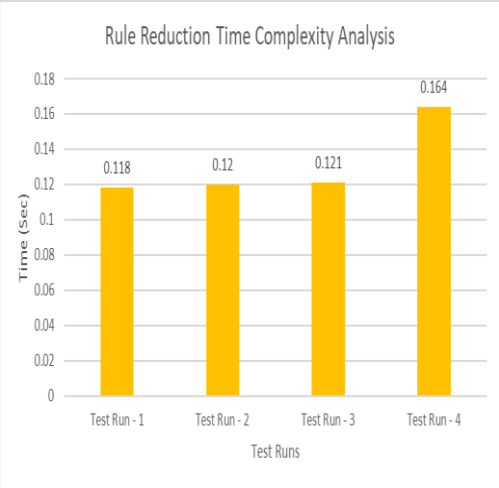


Fig. 6. Rule Reduction Time Complexity Analysis.

Once the rulesets are reduced and the inference properties are extracted, the actual distribution of the data must be analysed as the number of attributes distributed over the test clusters are the key points for detection of the leakages. Hence, the data distribution is analysed here [Table III].

TABLE III. DATA DISTRIBUTION ANALYSIS RESULTS

Test Number	Number of Clusters Detected	Number of Attributes (Mean) stored	Time Complexity (Sec)
Test Run - 1	5	4	0.023
Test Run - 2	4	7	0.022
Test Run - 3	4	7	0.022
Test Run - 4	5	6	0.023

Further after the analysis of the data sets, which are distributed over multiple clusters, finally the data leakages are detected, and the result is furnished here [Table IV].

TABLE IV. DATA LEAKAGE DETECTION ANALYSIS RESULTS

Test Number	Number of Clusters Detected	Number of Clusters with Leakage	Time Complexity (Sec)	Detection Accuracy (%)
Test Run - 1	5	2	0.010	89.55
Test Run - 2	4	3	0.012	89.55
Test Run - 3	4	3	0.009	89.55
Test Run - 4	5	4	0.018	89.32

It is natural to realize that, the higher accuracy of the detection process leads to higher security of the data distribution during the distributed query processing.

The results are analysed graphically here [Fig. 7, Fig. 8].

Henceforth, with the detailed analysis of the obtained results, in the next section of this work, the results are compared with the parallel research outcomes.

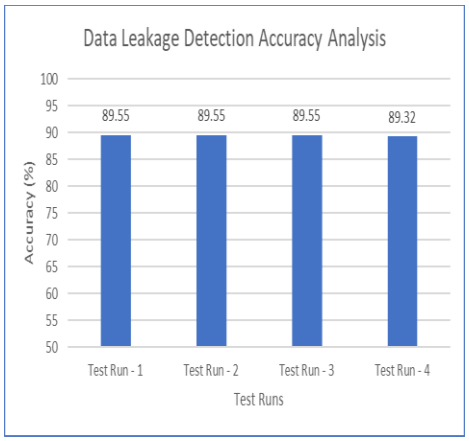


Fig. 7. Data Leakage Accuracy Analysis.

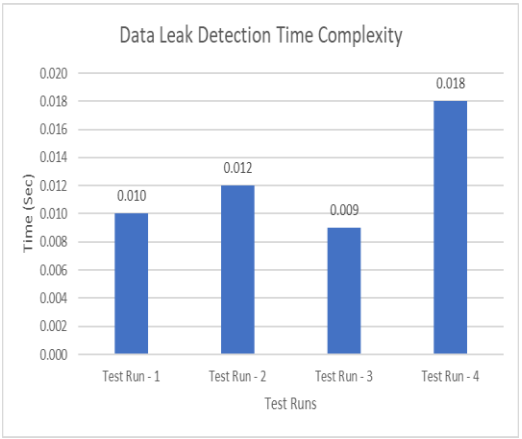


Fig. 8. Data Leakage Time Complexity Analysis.

VIII. COMPARATIVE ANALYSIS

In order to establish the believe that the proposed model is better performing than the other parallel research outcomes, the comparative analysis is most important. Thus, in this section of the work, the proposed algorithms are compared with the parallel research outcomes [Table V].

Hence it is natural to realize that, the proposed algorithms have outperformed the other parallel research attempts. The detailed reason for this achieved benefits are discussed in the previous sections of this work.

Further, with the analysis of the results and comparative analysis, in the next section of the work, the research conclusion is presented.

TABLE V. COMPARATIVE ANALYSIS

Proposed Methods	Year	Fundamental Mechanism	Detection Accuracy (%) [Mean]	Time Complexity (Sec) [Mean]
Appcaulk by J. Schütte et al. [9]	2014	Distributed Data & Taint Tracking	80	0.37
Proactive Warning by T. Malderle et al. [8]	2018	Centralized Data & Historical Data Analysis	88	0.40
Monitoring Methods by S. Trabelsi et al. [10]	2019	Centralized Data & Cost-Based Analysis	85	0.24
OBRSE-ATM, DI-RSR & ASLD Method [12]	2020	Distributed Data, Adaptive Threshold Model, Dynamic Inference & Subset Equivalence Analysis	89.49	0.01

IX. CONCLUSION

The distributed query processing is an essential part of today computing and the challenges of the distributed query processing is the leakage of the data. The data leakage can lead to critical security issues. Thus, this work identifies the solutions to detect the data leaks, which can further be used to ensure data distribution carefully. As mentioned in the previous sections of this work, the detection of the data leakages is highly difficult and demands a deep machine learning based approach. Thus, in order to solve the data leakage detection this work demonstrates a step by step process as initially the data relation between the data set attributes are extracted in form of rulesets. During the further processing, it is been observed that, due to higher number of rulesets in the system, the computational complexity is increasing to a greater extend, thus, this work again deploys a novel mechanism for reduction of rulesets. The deployed algorithm for rule reduction is designed carefully not to lose any inference properties. The rule reduction algorithm demonstrates a nearly 50% reduction in rulesets. Further, the data distribution is analysed, and the number of clusters are detected. Finally, this work deploys yet another novel machine learning based algorithm for detection of the data leakages based on data information equivalence and demonstrates a nearly 89% accuracy. The algorithms designed in this case are highly generic and can be applied for any data distribution scenarios and can be considered as a benchmark in this field of research for making the distributed query processing domain safer and faster.

REFERENCES

[1] Data Breach Index, https://breachlevelindex.com/.

[2] C. Missaoui, S. Bachouch, I. Abdelkader, S. Trabelsi, "Who Is Reusing Stolen Passwords? An Empirical Study on Stolen Passwords and Countermeasures", International Symposium on Cyberspace Safety and Security, pp. 3-17, 2018, October.

[3] Cost of a Data Breach Study: Global Overview, July 2018.

[4] P. Papadimitriou, H. Garcia-Molina, "Data leakage detection", IEEE Transactions on knowledge and data engineering, vol. 23, no. 1, pp. 51-63, 2011.

[5] L. Cheng, F. Liu, D. D. Yao, "Enterprise data breach: causes challenges prevention and future directions", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 5, 2017.

[6] S. Liu, R. Kuhn, "Data loss prevention", IT professional, vol. 12, no. 2, 2013.

[7] B. Hauer, "Data and information leakage prevention within the scope of infor-mation security", IEEE Access, vol. 3, pp. 2554-2565, 2015.

[8] T. Malderle, M. Wübbeling, S. Knauer, A. Sykosch, M. Meier, "Gathering and analyzing identity leaks for a proactive warning of affected users", Proceedings of the 15th ACM International Conference on Computing Frontiers, pp. 208-211, 2018.

[9] J. Schütte, D. Titze, J. M. De Fuentes, "Appcaulk: Data leak prevention by injecting targeted taint tracking into android apps", 2014 IEEE 13th International Conference on Trust Security and Privacy in Computing and Communications, pp. 370-379, 2014.

[10] S. Trabelsi, "Monitoring Leaked Confidential Data," 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), CANARY ISLANDS, Spain, 2019, pp. 1-5.

[11] A.Sindhura, J.Rajeshwar, M.V.Narayana , M.Ram Babu, "An Effective Semantic Web Knowledge Processing Mechanism by Using an Adaptive Swarm Intelligence Technique for Ontology (ASITO)", International Journal of Engineering Trends and Technology, Volume 69 Issue 3, 195-200, March 2021, ISSN: 2231 – 5381.

[12] Niladri Shekar Dey, Purnachand Kollapudi, M V Narayana, I Govardhana Rao, "An Automated Framework for Detecting Change in the Source Code and Test Case Change Recommendation", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 8, 2020, pp.270-280, ISSN : 2156-5570 (Online), 2158-107X (Print).

Knowledge Graph-based Framework for Domain Expertise Elicitation and Reuse in e-Learning

Jawad Berri
Information Systems Department
King Saud University, Saudi Arabia

***Abstract*—Reusing knowledge expertise of different domains in e-learning is an ideal approach to sustain knowledge and disseminate it throughout the different organizations' processes. This approach generates a valuable source for instruction which can enrich significantly the quality of teaching and training as it uses effortlessly expertise from its original sources. It is also very useful for teaching activities since it connects learners with real-life scenarios involving field experts and reliefs instructors from the tedious task of authoring teaching material. In this paper we propose a framework that allows gathering automatically expertise from domain experts while doing their activities and then represents it in a form that can be shared and reused in e-learning by different types of learners. The framework relies on knowledge graphs that are knowledge representation structures which facilitate mapping expertise to e-learning objects. A case study is presented showing how inspector reports are handled to generate on-demand e-courses specifically adapted to learners' needs.**

Keywords*—Knowledge graph; domain expertise; e-learning; knowledge elicitation; learning web

I. Introduction

Expertise is a valuable asset in today's competitive world. Organizations are striving to safeguard and make the most of their expertise to sustain development and maintain an advantage with competitors. Expertise is retained by human experts; it is the set of know-how and skills developed through practicing and experiencing their knowledge in a specific domain [1]. Identifying, representing and sharing expertise requires the development of knowledge management systems that are able to sustain knowledge and disseminate it through the different organizations' processes. These two requirements represent a twofold challenge facing the development of knowledge management systems. First, acquisition of expertise is not a straightforward task; it requires specific settings and work context to obtain it from experts. Most of the methodologies that have been proposed to acquire expertise relied mainly on eliciting expertise manually or by using elicitation systems that rely on the availability of experts where meetings are organized to acquire their expertise [2]. This ends up generally with additional load for experts which is not always in line with their daily commitments and duties resulting in less involvement and motivation. Second, transferring and reusing expertise requires that the representations of knowledge have the articulacy and flexibility to smoothly transfer it from an initial domain to another domain and to adapt it to different contexts and users. The ability to reuse expertise remains the ultimate goal of organizations. e-Learning systems can contribute a great deal to reuse expertise by mapping it to e-courses for on demand learning and training [3]. In order to provide such e-learning environment for organizations, these systems should be able to handle expertise in a mechanistic way so that to not add a burden on experts while transferring their expertise and also on instructors while preparing e-courses for learners.

In this research we propose a framework that is designed to transfer domain expertise through e-learning systems which can generate and adapt the learning material to different target learners. The framework maps conceptual representations of expertise acquired during normal experts' activities while on duty into learning material that can be used by different types of learners. Expertise is represented as knowledge graphs that are knowledge representation structures which facilitate mapping expertise to e-learning courses. Knowledge graphs capture the expertise concepts and their relationships offering a semantic organization of concepts which allow a fluent mapping into an e-learning course where concepts are prerequisite of other concepts. Also, accessing concepts over the web is facilitated by available technologies that allow querying remotely knowledge graphs which opens the possibility to share and reuse all knowledge graphs on the web. The transition from knowledge graph representation to a e-course is made possible by the learning web constructor algorithm, proposed in this research, which is used to structure e-learning material as a tree of concepts that are adapted to fit the learner's context and profile. A case study is presented to illustrate a real life application of the proposed framework. It shows a company that employs experts who write inspection reports that are automatically represented as knowledge graphs during the report elaboration phase. Domain knowledge and reports are stored into a knowledge graph base that is used to generate on-demand e-courses specifically adapted to learners' needs. The knowledge graph base is then queried to retrieve the necessary concepts which are used to generate a e-course for two types of learners: inspection trainees and new recruits.

This paper is organized as follows: Section 2 provides a background on knowledge engineering and knowledge graphs. Section 3 presents research works in relation to the present research. The following section presents the knowledge management framework and explains the different stages to handle expertise. Section 5 details the inspector report case study that illustrates how expertise is acquired and reused. Finally, we conclude this research and provide future potential research paths to investigate.

II. Background

A. Knowledge Engineering

Knowledge engineering is a critical task and a bottleneck facing any knowledge management system aiming to handle human expertise. During this task expertise is acquired from domain experts by specialized knowledge engineers. Many methodologies have been suggested and used to perform this task and the majority relies on the availability and willingness of experts to transmit their tacit knowledge to the knowledge engineer who encodes it into the knowledge base of a knowledge management system [4], [5]. Unfortunately, experts are not always available and experience has shown that classical knowledge engineering could not be applied at large scale. Indeed domain dependence of the developed systems was not easy to bypass towards flexible and domain independent systems. In order to circumvent this difficulty and make this phase at large scale, knowledge engineering should be transparent to experts relieving them from the burden of eliciting knowledge for every domain from scratch [6], [7]. Also knowledge engineers should focus more on the development of knowledge management systems which serve the daily tasks of experts and at the same time can detect, acquire and represent knowledge into knowledge bases that can be reused and shared effortlessly. Such systems can be developed with the widespread of mobile technologies and context-aware mobile systems which are able to adapt to complex context situations and recommend sophisticated solutions to users. This has been the target of many research such as in the medical, engineering and manufacturing fields [8]. In these domains experts have mobile applications that facilitate their work and at the same time perform a great management work on the background such as efficient storage and retrieval of information, efficient context management to adapt information to users and recommendation systems to guide them through the best and optimal solution [9].

B. Knowledge Graphs

Knowledge graphs are graphs of data intended to accumulate and convey knowledge of the real world [10]. Knowledge Graphs involve interlinked descriptions of entities – objects, events or concepts. These are semantic graphs which can capture subtle meanings that can enhance inference in knowledge management systems. The implementation of knowledge graphs and their use in different applications can also foster the development of intelligent systems able to reason and recommend knowledge. Knowledge graphs can use ontologies to provide an abstract representation of a domain where graph concepts and relationships are concisely defined. Besides, available technologies allow accessing concepts over the web which facilitates sharing and reusing remotely knowledge graphs. Also, query languages, such as SPARQL, have been defined to query knowledge graphs which opens the possibility to search and retrieve knowledge graphs easily [11].

A knowledge graph is defined as a graph G = {E, R, F}, where E is a set of entities, R a set of relations and F a set of facts that have the form of a triple (e, r, e') [12]. For instance, (Tom, FriendOf, Jerry) is a triple denoting a relationship FriendOf between two objects Tom and Jerry. This simple fact can be represented in predicate calculus as: FriendOf(Tom, Jerry). Predicate calculus is a language which handles knowledge graph representations in the form of facts, rules for inferring new knowledge and allows expressing queries for retrieving knowledge. For instance, the query FriendOf(Tom, X) retrieves all friends of Tom by instantiating X with all objects that match similar predicates and hence knowledge graphs.

III. Related Work

Sharing and reusing expertise is a broad research field which involves many disciplines such as knowledge management to elicit knowledge of experts, artificial intelligence to represent knowledge and use it in reasoning, computer information systems to develop applications able to diffuse knowledge to the right user in the right context. e-Learning is an ideal application domain that can contribute to reuse expertise as it transfers tacit knowledge to explicit knowledge through learning, training, coaching, or mentoring [2]. e-Learning systems developed for education or training implement instructional material into user-friendly sophisticated systems that have the ability to adapt to learners in need of domain expertise in a specific context [9]. Knowledge representations are naturally used in teaching and learning in the form of concept maps and taxonomies to categorize concepts and to illustrate them through concept visualization. In this section we present research works that have used concept graphs in education from different perspectives.

Shi et al. in [13] propose a learning path recommendation model which uses specific semantic relationships and knowledge graphs to propose learning objects to learners. The objective of the proposed framework is to increase learning efficiency and recommend personalized learning paths. This framework has been applied to learn machine learning algorithms. Learning objects are categorized into three classes namely: basic knowledge, algorithm, and task. A recommendation algorithm is then used to recommend the optimal learning path for learners based on scores of learning algorithms features such as publication time, citation count, search frequency and impacts of the publisher and author. While this approach seems to work well in the specific domain of machine learning algorithms, it has two major limitations: i) it needs a manual elicitation of the domain knowledge to identify and categorize learning objects for every domain, and ii) some of the semantic relationships defined to link learning objects such as: Ori-algorithm (current LO was improved from an original algorithm), ApplyToAlgorithm and ApplyToTask, are local to this particular domain and can hardly applied to another domain. In [14] a system is developed based on knowledge graphs to provide personalized learning content to learners that are categorized by their learning abilities skill set. Knowledge graphs are constructed based on the concepts extracted from learning objects and relationships are set between concepts. Based on the core concepts and the relationships semantic of the knowledge graph, graphs are generated automatically to the three categories of learners. Accordingly, slow learners are given only the core concepts, then additional graph relationships are considered to generate learning content for moderate learners, and finally highly skilled learners are offered more learning objects based on an

extended knowledge graph. In [15] authors developed a learning path generator based on knowledge graphs to provide guidance for learners. The method uses a topological ranking algorithm to generate a topological structure of the learning path and then learning objects are serialized using ant colony optimization. Evaluation of the method shows that the generated learning path is comparable to expert learning path in terms of learning outcomes. Authors in [16] propose KnowEdu, a system that construct educational knowledge graphs that can be used for online learning in school. The system uses pedagogical data and learning assessment data to extract instructional concepts of courses and then identifies significant relations holding between these concepts. Concepts and relations are extracted using respectively neural sequence labeling algorithm on pedagogical data such as textbooks and course tutorials, and association rule mining on learning assessment data to identify the relations such as prerequisite and inclusion. The authors present a case study where the system was used to build a knowledge graph for mathematics course.

IV. Knowledge Management Framework

The framework depicted in Fig. 1 shows the different stages that allow handling knowledge from its source till its utilization in e-learning systems. The framework exhibits four main phases namely: Elicitation, Management, Reuse and Sharing, for converting knowledge expertise as it is acquired from experts till it is shared by learners in various learning systems.

A. Knowledge Elicitation

Knowledge is sought through daily expert activities such as coaching, auditing, brainstorming, training, consultation and mentoring. Knowledge undergoes a four step process: Identify, Extract, Validate, and Represent. Knowledge identification is done from resources such as videos, audios, manuals, reports, regulations, procedures, etc. produced by the expert and made available for performing the activities. These resources are either available internally within the organization and are queried from databases or are retrieved using web services for activities that are posted on the web or online social networks. Knowledge is then extracted automatically from these resources and then validated to make sure that it fits with the objectives set for reusing and sharing learning content. In the last process step knowledge is represented into the knowledge base for further use. This process is supported by the Elicitation Model which specifies what knowledge is sought, from where to get it and how to extract it and validate it. It is also supported by the knowledge representation language which allows representing knowledge into the knowledge graph base.

B. Knowledge Management

Knowledge Management step handles mainly the storage of knowledge in the knowledge graph base and provides a reasoning engine to infer new knowledge or to adapt knowledge according to the user's context. For this purpose, context is constantly gathered and updated to allow adaptation of learning content. Knowledge is also scored which is necessary in order to retrieve quality knowledge in response to requests of learners.

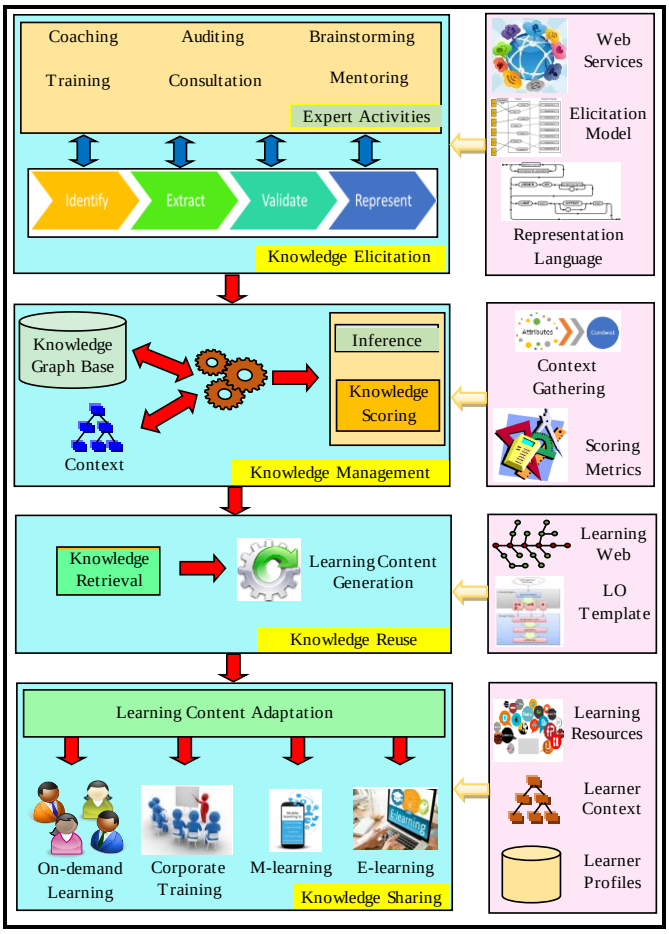


Fig. 1. Knowledge Management Architecture.

C. Knowledge Reuse

This step provides the necessary modules for retrieving knowledge and generating the learning material. Reusing knowledge for learning relies on two components: the Learning Web (LW) and the Learning Object Template. LW is the main learning structure which organizes the learning material. LW is generated by the Learning Content Generation module. It is similar to a tree of nodes where the root is the first learning object and the other nodes represent LOs available for the learner to visit to complete the learning requirements. LOs are learning units including a set of resources that are organized into specific templates. Algorithm ConstructLW in Fig. 2 takes three parameters Le, KG and LW. Le is an input list of ordered concepts in the knowledge graph base selected by an instructor to deliver a particular e-course. This list should be ordered according to the pre-requisite teaching precedence. KG is the input knowledge graph base as described in Section 2; it is defined as {E, R, F}. LW is the learning web which is initially set to an empty list and which will include a list of ordered concepts in relation with the concepts in Le. ConstructLW retrieves all direct concepts e' in relation with concepts e in Le and adds them to LW. It is noted that the union set operation (Line 9) prevents the duplication of concepts already in LW. The complexity of ConstructLW is in $O(n^2)$ where n is the maximum number of concepts e' that can be in relation with a concept e in LW.

```
1. ConstructLW (Le, KG, LW)
2. input: Le                          is a list of concepts in the
                                      knowledge graph KG
3. input: KG = {E, R, F}              is a knowledge graph;
4. output: LW = ∅                     is the Learning Web
5.    for each e ∈ Le do
6.    {
7.     LW = LW ∪ {e}                  add e to LW
8.     for each e' such that (e, r, e') ∈ KG do
                                      e' is a concept in relation with e
9.          LW = LW ∪ {e'}            add e' to LW
10.   Le = Le – {e}    remove e from Le
11.   }
12. return LW
```

Fig. 2. Learning Web Generator Algorithm.

D. Knowledge Sharing

In this phase learning generated is adapted to fit different learners according to their context and profiles. Sharing knowledge is vital for organizations. It should be a daily integral part of a learning organization which must develop clear strategies for diffusing knowledge and define the appropriate use of the transmitted knowledge. In order to be able to share and adapt knowledge this phase uses three components namely learning resources, learners' profile and context which allow content adaptation to different types of learners.

V. Case Study

The Inspection Report case study presented in this section illustrates the knowledge management architecture (Fig. 1) and shows how expertise is elicited and reused. A petroleum and gas company has many natural gas processing plants which includes satellite stations that treat Liquefied Natural Gas (LNG). These stations have specialized maintenance engineers who monitor the station's devices (such as piping systems, storage tanks, vaporizers, control devices, pressure regulating valves, etc.) in order to conduct risk assessment, predict deficiencies and propose corrective maintenance actions. The objective is to improve operational efficiency, guaranty safety and protect the environment. Engineers perform regular on-site visits and write reports about the station devices. In case there is a deficiency they propose the repairing to be done and the devices to replace. Reports have a specific template which include an identification section, a description of the problem (if any), reference to previous reports about the same device (or problem), corrective actions proposed and possible maintenance (replacement, reparation, …) to be done on the device along with the execution schedule. While on-site engineers can access all the reports stored in the database to check the maintenance history and evaluate the progress of previous defects. They can also communicate with each other seeking advice or corroboration of their diagnostics. Reports are recorded using a mobile tablet through a template allowing the engineer to write text, take pictures and record audios and videos. Also communication with peers is recorded and added to the report.

A. Eliciting Knowledge Reports

Reports represent a valuable source of expertise that is consulted by managers, experts, maintenance engineers. Also trainees and new recruits can learn and practice their knowledge from inspector reports. When inspectors do their inspection task a knowledge graph is created based on the report attributes such as the inspector ID, the task name, and the component inspected. Fig. 3 shows part of the knowledge graph including two types of concepts: domain knowledge concepts about Refrigeration Piping (green colored circles), and task related knowledge (blue colored circles) which captures the inspection tasks done by inspectors to check pressure of a specific Refrigeration Piping. The figure shows a superposition of concepts for *Inspection Report* and *Inspector* exhibiting the fact that multiple reports have been done by many inspectors for the same task that is to check pressure of component Piping GNL 225-R3674. Knowledge elicitation is done in a transparent manner without intervention of inspectors. While performing their tasks, the task related knowledge is acquired whenever an inspector creates an inspection report to inspect a component. The report is then added automatically as a subgraph in the knowledge graph base as a task related knowledge.

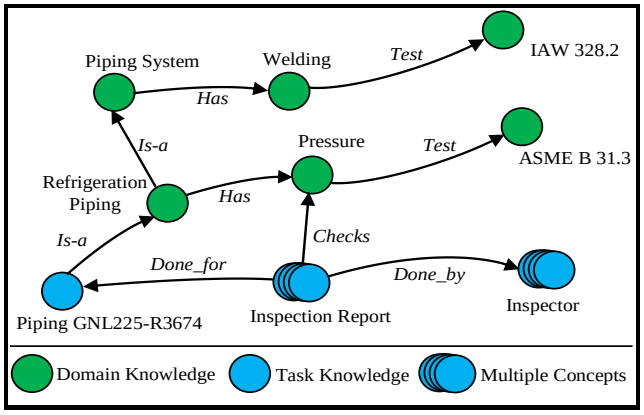


Fig. 3. Piping System's Knowledge Graph.

B. e-Learning Content Generation

Inspection reports expertise represented as knowledge graphs is used by experts, inspectors, and engineers to plan on-site inspections and to follow up on maintenance. It is also exploited by the training and development unit to provide learning content for trainees and new recruits. Training material is generated, as e-learning courses, by querying the knowledge graph base for both inspection trainees and new recruits for the plant. These two types of learners have different backgrounds and different learning objectives and hence require different learning material. Inspection trainees are company employees who need to be trained on inspecting the site, they are knowledgeable about the plant processes. They need to be trained on testing the refrigeration piping system by using the test specification ASME B 31.3. New recruits are engineers who need to acquire knowledge about the plant processes and get some practice on the unit piping system. In order to cover the learning needs of these two types of learners, two different learning webs are generated as shown in Fig. 4.

1) Learning web generation: The learning web is a learning structure organized as a tree of units including the building blocks of learning called Learning Objects (LO) [3]. LW represents the e-course that is traditionally designed by instructors. In our case LW is automatically generated from the

knowledge graph base which is queried to retrieve the knowledge subgraph containing the necessary knowledge concepts to fulfill the requirement of learning for a specific learner [17], [18]. These concepts are then ordered to form a tree of concepts. Each concept is materialized by a LO (or set of LOs) that includes the necessary learning resources to be exposed to the learner [9].

Both LWs in Fig. 4 are generated by *ConstructLW* algorithm (Fig. 2). They include mandatory and elective LOs. Mandatory LOs represent the required knowledge that is essential in any learning course. They correspond to domain knowledge concepts in the knowledge graph (Fig. 2). Elective LOs are supporting learning units which are not essential to achieve the learning outcomes but they can support learners to understand, practice or have examples about the core concepts. Elective LOs correspond to task knowledge concepts in the knowledge graph.

Trainees' LW (TLW) is focused on a specific task that is to check and maintain the refrigeration piping system. They need to learn about this system and also be able to elaborate a report using the test specification ASME B 31.3. The input list of concepts Le for generating TLW is Le = {Refrigeration Piping, Pressure}. Algorithm ConstructLW generates LW by including the following: i) mandatory LOs: Refrigeration Piping, Pressure, and ASM B 31.1 representing the necessary domain knowledge and ii) elective LOs: Piping GNL225-R3674 and all reports that have been done for this particular component of the refrigeration piping. The new recruits' LW (RLW) is generated to acquire knowledge about the piping system and get some practice about the regular system maintenance. The input list of concepts Le is Le = {Piping System, Welding, Refrigeration Piping, Pressure}. In this case, ConstructLW generates RLW including the following: i) mandatory LOs: all domain knowledge concepts to allow engineers to learn about the piping system and ii) elective LOs: Piping GNL225-R3674 and the inspection and maintenance reports to allow engineers to have practical sessions on the piping system.

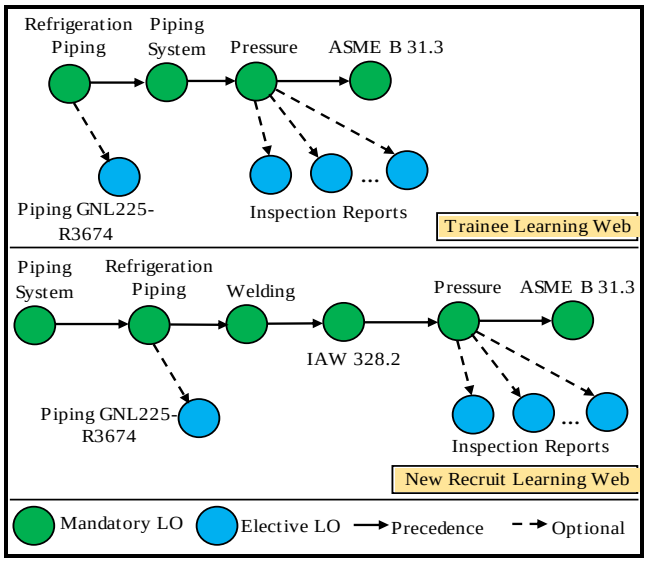


Fig. 4. Generated Learning Webs for Inspector Trainee and New Recruit.

2) Learning objects generation: Once the LW is built learning object are generated for each concept. During the elicitation phase resources are gathered and stored. These resources are then retrieved and packaged as LOs using specific templates. An example of a domain concept, the *Piping System LO*, is illustrated in Fig. 5. The generated LO includes a text description about the concept *Piping System*. It includes also a video, an image and a list of topics organized the same way as the LW, i.e. a sequence of domain concepts and task concepts supporting domain concepts.

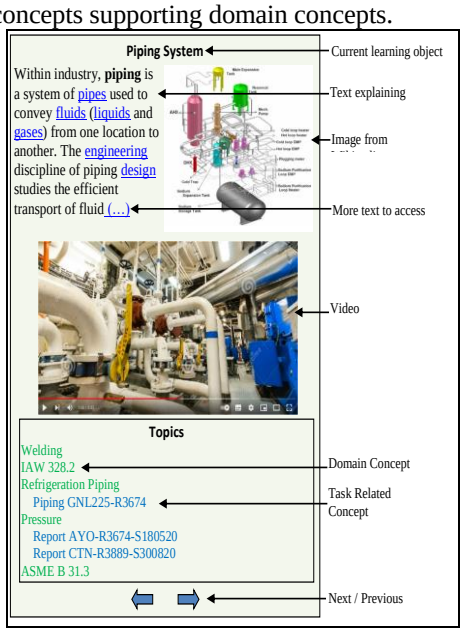


Fig. 5. Piping System Learning Object.

Navigation in the LW can be done in two ways. The first possible navigation is to use *Next* and *Previous* arrows at the bottom of the LO. This allows a sequential navigation in the LW moving forward (or backward) to the next (or previous) domain concept triggering the generation of the corresponding LO. The second possibility is to navigate freely in the LW by clicking on any LO listed in the *Topics* area. This gives more freedom to the user to focus on the concepts he is more interested in and do not waste his time with concepts already known.

VI. Discussion

The framework presented in this paper is designed to safeguard expertise and promotes a smooth transition from expertise represented as knowledge graphs to e-learning courses adapted to different learner types. Usually experts (such as doctors, consultants and auditors) in organizations perform their daily duties using computer programs or applications to facilitate their tasks. While these activities are generally recorded in databases (such as the inspection reports) rarely this expertise is shared and reused within or outside the organization. The framework contributes to not only gather expertise during experts' activities but explicitly turns expertise recorded into e-learning material for sharing by learners in the organization. The representation of expertise using knowledge graphs offers a clear flexibility to the web learning generation algorithm which allows to extract a subgraph on-demand. The

proposed algorithm extracts the target concepts and their related concepts to generate the learning web used for e-learning. Although the algorithm in its actual version is restricted to the first level related concepts, it can be easily updated to consider additional levels offering more learning depth and more adaptation to learners.

VII. Conclusion

Reusing expertise is e-learning opens many opportunities for organizations to exploit and share efficiently their know-how and skill set. The framework presented in this paper supports such fluent transition from expertise elicitation till its reuse for learning and training. Expertise is represented as knowledge graphs which provides two advantages: i) concepts are semantically organized which allows an inherent mapping in e-learning where concepts are prerequisite of other concepts, ii) accessing concepts over the web is facilitated by available technologies that allow querying remotely knowledge graphs which opens the possibility to reuse all knowledge graphs on the web. The transition from knowledge graph representation to a learning web is made possible by the learning web constructor algorithm which maps concept graphs into a tree-like structure of learning objects composing a e-course. This gives the possibility to adapt knowledge to different learner types based on their needs and context. The case study presented illustrates the framework's feature while the same knowledge graph has been queried for Inspection Trainee and New Recruit to generate learning material to both learners taking into account their context and learning needs. Future research plans are to develop the scoring function which scores learning resources according to diverse criteria such as author profile, resource creation time, resource popularity (such as number of likes, number of views). This will promote the use of the most interesting resources to fulfill effectively the learning requirements.

References

[1] N. Stehr, R. Grundmann, *Experts. The Knowledge and Power of Expertise*, 1st Ed., Routledge, 2011.

[2] A., E. M., and H. Ghaziri, *Knowledge Management*, 2nd Ed., International Technology Group, LTD., North Garden, VA, 2010.

[3] J. Berri, R., Benlamri, Y., Atif, H., Khallouki, "Web Hypermedia Resources Reuse and Integration for On-Demand M-Learning", *International Journal of Computer Science and Network Security*, vol. 21, no. 1, pp. 125-136, Jan., 2021.

[4] B. J. Wielinga, A. Th. Schreiber, J. A. Breuker, "KADS: a modelling approach to knowledge engineering", *Knowledge Acquisition*, vol. 4, no. 1, pp. 5-53, Mar, 1992.

[5] R. Studer, V. R. Benjamins, D. Fensel, "Knowledge engineering: Principles and methods", *Data & Knowledge Engineering*, vol. 25, no. 1–2, pp. 161-197, Mar., 1998.

[6] A. T. Bimbaa, N. Idris, A. Al-Hunaiyyan, R. B. Mahmud, A. Abdelaziz, S. Khana, V. Chang. "Towards knowledge modeling and manipulation technologies: A survey", *International Journal of Information Management*, vol. 36, no. 6, Part A, pp. 857-871 Dec., 2016.

[7] G. Leu, H. Abbass, "A multi-disciplinary review of knowledge acquisition methods: From human to autonomous eliciting agents", *Knowledge-Based Systems*, vol. 105, no. 1, pp. 1-22, Aug., 2016.

[8] T. Ali, J. Hussain, M. B. Amin, M. Hussain, U. Akhtar, W. A. Khan, U. U. Rehman, S. Lee, B. H. Kang, M. H., M. Afzal, H.-S. Han, J. Y. Choi, H. W. Yu, A. Jamshed, "The Intelligent Medical Platform: A Novel Dialogue-Based Platform for Health-Care Services", *Computer*, vol. 53, no. 2, Feb., pp. 35-45, 2020.

[9] S. Al-Marshad, J. Berri, "Learning from Web Searching: Enhancing Users' Experiences with NaviWeb Mobile System", *Journal of Digital Information Management*, vol. 19, no. 4, pp. 113-124, Dec., 2021.

[10] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. Ngonga Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, "Knowledge Graphs", *ACM Computing Surveys*, vol. 54, no 4: 71:1-71:37, 2021.

[11] R. Angles, M. Arenas, P. Barceló, A. Hogan, J. L. Reutter, D. Vrgoc, Foundations of modern query languages for graph databases, *ACM Computing Surveys*, vol. 50, no. 5, 68:1– 68:40, 2017.

[12] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications", *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-27, Apr., 2021.

[13] D. Shi, T. Wanga, H. Xing, H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning", *Knowledge-Based Systems*, 195, pp. 1-11, May 2020, 105618.

[14] A. J. Martin, M. M. Dominic, "Personalization of Learning Objects according to the Skill Set of the Learner using Knowledge Graph", *Turkish Journal of Computer and Mathematics Education*, vol.12 no.6, pp. 3974-3987, Apr., 2021.

[15] J. Gao, Q. Liu, W.-B. Huang, "Learning Path Generator Based on Knowledge Graph", *International Conference on E-Education, E-Business, E-Management, and E-Learning (IC4E)*, Tokyo, Japan, Jan. 10–13, pp. 25-33, 2021.

[16] P. Chen, Y. Lu, V. W. Zheng, X. Chen, B. Yang, "KnowEdu: A System to Construct Knowledge Graph for Education", *IEEE Access*, vol. 6, pp. 31553-31563, May, 2018.

[17] W. L. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, J. Leskovec, "Embedding Logical Queries on Knowledge Graphs", *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, pp. 1-12, 2018.

[18] S. Liang, K. Stockinger, T. M. de Farias, M. Anisimova, M. Gil, "Querying knowledge graphs in natural language", *Journal of Big Data*, vol. 8, no. 3, pp. 1-23, 2021.

Leveraging Artificial Intelligence–enabled Workflow Framework for Legacy Transformation

Abdullah Al-Barakati
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—The rapid advancement of web technologies coupled with evolving business needs make legacy transformation a necessity for enterprises around the world. However, the risks in such a transformation must be mitigated with an approach that is flexible enough to allow for a gradual and low risk transformation process. This paper presents a Service Oriented Architecture (SOA) workflow-based legacy transformation approach that allows for phased transformation in which a legacy system is first transformed into self-contained modular services accessible via a dedicated service layer. These modular services are managed through an AI-enabled workflow management layer that interacts with improved UI frontend for the system's end users. This paper presents a hypothetical prototype in which an Oracle 5 legacy system is transformed using the proposed architecture. ASP .NET Core MVC as well as Pega business process management platform are utilized to practically assess the feasibility of the proposed approach.**

***Keywords*—Legacy systems; service oriented architecture; workflow management; legacy transformation; digital transformation; artificial intelligence (AI)**

I. Introduction

Legacy systems can be defined as applications that were built with old technology but are still in use in many business environments [1]. Despite the apparent advantages of transformation from legacy systems to modern web-based solutions, government and private enterprises consider the process risky and challenging and, therefore, show reluctance in initiating the change. Such reluctance can be attributed in part to the heavy investments that were associated with developing legacy applications [2]. Furthermore, enterprises incur heavy costs to train their employees and tailor their processes to benefit from legacy systems [3]. On the other hand, the legacy transformation and modernization process itself can be lengthy and costly. Major risks include the inherent complex designs of legacy systems [4], tightly coupled components, system performance issues, and difficulties in mapping current systems to target architectures and platforms [5]. Additionally, the underlying knowledge about such systems is usually scanty due to limited documentation and the unavailability of the developers who originally built these systems [4]. As a result, most legacy modernization tenures tend to begin with lengthy reverse engineering periods to document current systems before paving the way for technology transformation.

It can be argued that legacy transformation is inevitable with the rapid advancements that technology is witnessing, especially the digital transformation phenomenon. Digital transformation places special emphasis on legacy transformation as one of the cornerstones of successful transformation strategies [6]. While legacy transformation and modernization processes can be lengthy and costly as mentioned earlier, they can offer long term cost savings, increased efficiency, better resource utilization and the ability to adapt to the dynamic business needs of any given enterprise [7]. Therefore, enterprises need an optimal transformation approach that will enable them to part with legacy systems and take advantage of the possibilities offered by modern web-based technologies [8].

In this paper, we propose a legacy modernization approach that aims for gradual technology upgrade from legacy systems to modern web-based solutions without disturbing business operations. It is based on a Service Oriented (SO) architectural approach that wraps exciting legacy applications with an AI-enabled workflow management layer. The workflow management layer acts as a service orchestrater that reduces the risks of inadequate service mapping when migrating from legacy systems to target modernized systems. Workflow management functionality is achieved via Business Process Management Solutions (BPMS) that can sit on top of the legacy system services. This approach emphasizes service orientation where business logic is captured and managed in a dedicated middle service layer that can potentially integrate with any future core systems that may replace current legacy systems. While this approach can be technology agnostic, we are showcasing a hypothetical case study where Pega BPMS is utilized to manage the workflows of an Oracle 5 form-based legacy system while having ASP.Net Core MVC as the main technology for a dedicated service layer.

The reminder of this paper is organized as follows. Section II sheds light on some of the research in legacy system transformation approaches. Section III outlines the overall architectural approach, its layers, and the integration points with legacy systems. Section IV introduces the suggested technology stack related to the proposed architecture. Section V showcases a practical implementation of the proposed approach as a hypothetical proof of concept. Section VI presents the findings of this paper and highlights possible areas of future work.

II. Legacy Transformation Approaches

Due to the importance of legacy system transformation, several studies have focused on finding the best way for a safe and fast transformation process. In this context, the work produced by [9] examines several options for legacy system

transformation in which replacement is considered the best option for old systems which are undocumented, outdated, or not extensible. However, authors in [9] note that the replacement of such systems is often a resource-intensive and risky process. On the other hand, [6] present a lightweight agile approach for effective low-risk legacy transformation as opposed to waterfall-based transformation approaches. Such an approach can potentially address the technical and procedural complexities associated with legacy transformation.

In the work presented by [10], the authors showcased the process of transforming a legacy social services information system to a modern digital platform. This platform capitalized on advanced technologies (Artificial Intelligence [AI] and Machine Learning) for analyzing and processing big data. From an architectural perspective, this transformation was enabled via the utilization of cloud computing, big data innovations, and the emerging microservices architectural principles. In a similar manner, [8] proposed a tiered architectural approach for legacy system transformation. In this approach, component configuration is specified in Extensible Markup Language (XML) files to facilitate legacy service wrapping and integration. The work presented by [11] is in conformity with the approach presented by [8] as the author argues that legacy systems can be transformed by exploiting modern, faster, and cheaper technologies such as Java and XML. He also indicates that such an approach can shift focus to functionality not technology, hence allowing for better response to the evolving business requirements of any given enterprise. Furthermore, [11] presented a legacy transformation software tool (RescueWare) that acts to decompose business knowledge into self-contained e-components tasked with performing certain business functionalities. These components are defined within standard Application Programming Interfaces (APIs) which can be accessed by other systems and interfaces that can potentially replace the legacy system in question.

The work done by [1] emphasized a component-based approach for legacy transformation. Their methodology includes a reengineering process to transform legacy systems into new components with upgraded software architecture design. They adopted a reverse engineering approach that is based on the extraction of architectural information from the existing codebases of legacy systems. Based on the extracted information, in conjunction with business domain knowledge, modular system components were to be developed to replace the existing legacy system. Similarly, [12] present an interactive tool to extract database and business logic components from legacy systems. The aim is to minimize the complexity of the migration process by introducing a decomposition step. This step can help slice the legacy system into encapsulated components that can be migrated into a client-server platform.

The work in [13] encompasses a legacy migration approach based on the conversion of legacy system architecture into Service Oriented Architecture (SOA) within a systematic predefined process. The process that they suggest is feature oriented as it focuses on a reengineering approach to transfer existing legacy services into web services facilitated by the Web Services Composition Language (WSCL). The author in [13] validated their approach with a case study which presented a prototype based on a layered architecture comprising three main layers: Interaction Layer, Translation Layer, and Repository Layer.

Much of the previous research focused on the concepts of modularity and service orientation for successful transformation. However, to our knowledge, integration of workflow management layers as a part of the legacy transformation process is an area that is not yet fully examined. For this reason, we propose an AI-enabled workflow-based approach for legacy system transformation.

III. Proposed Architecture

To counter the risks and complexities that accompany legacy transformation tenures, a layered SOA-based architectural approach is proposed. SOA can play the role of a transformation and integration enabler for legacy systems [7]. Exposing legacy services in a service-oriented manner will provide for modular services that can be exploited by a variety of interfaces. Such interfaces can be frontend systems or other core systems that benefit from the services of the legacy system. Furthermore, Service Orientation - as a concept and a transformation enabler - will allow for greater interoperability for legacy services. More importantly, business logic transformation to a service layer will lead to decoupling the services of the core system to further facilitate legacy replacement/enhancement.

In addition to the advantage of transformation to service-oriented components, this model is complemented with a dedicated workflow management layer which is tasked with orchestrating the different business services of the legacy system in transformation. Such a model aims to facilitate the process of managing the usually complex services and workflows of a typical legacy system. Additionally, by having a dedicated workflow management layer, it will be possible to gradually move the legacy services from the core system to a service layer. In such a scenario, the workflow management layer will orchestrate uninterrupted business operations by managing the right mix of legacy services, external integration on the one hand and user interactions on the other.

Two main architectural principles will govern and shape the proposed legacy modernization architecture. Firstly, a microservices approach will allow for rapid delivery of the system's business services [14]. Microservices will also be an enabler for a robust technology stack that can be enhanced or modified when and if needed. Enhancement can be achieved by plugging in more service layers to cover any evolving business needs. In conjunction with the utilization of microservices architecture, the proposed solution focuses on the development of domain-specific services. Hence, the legacy application's services will be divided into self-contained modules based on business areas. This approach adheres to the principles of Domain Driven Design (DDD). Thus, business context will be divided into individual areas that can be developed and managed separately. Greater modularity and manageability can be considered as important advantages of this approach. Based on the proposed approach, the following architectural layers will be required as illustrated in Fig. 1.

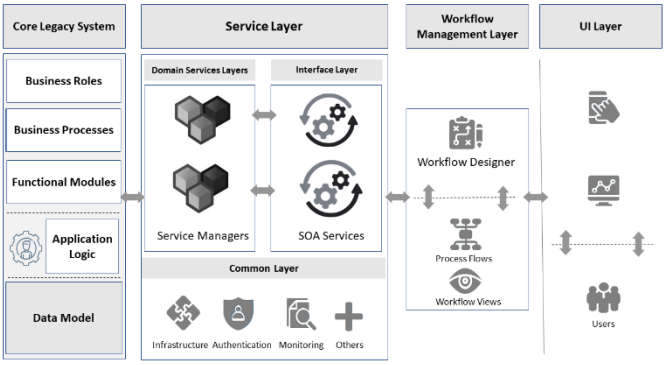


Fig. 1. Proposed Architecture.

A. Core Legacy System Layer

In the initial phases of transformation, the core legacy system will be considered as a backend system that the other system layers will be interacting with. As is the case with typical legacy systems, this layer will comprise application software and any associated databases.

B. Service Layer (Middle Layer)

To be able to adopt a service-oriented approach, most business logic should be separated from the core system into a dedicated API-based Service Layer (middle layer). This layer will contain the business logic such as all validations and custom roles related to the core legacy system. The service layer will be divided into the following sub-layers:

1) Interface Layer (API): This layer will be dedicated to exposing the services of the core legacy system into Representational State Transfer (RESTful) APIs that are consumable by the Workflow Management Layer. Furthermore, RESTful APIs will be exposed as web services that can be accessed by either internal or external systems.

2) Domain services layer: Following the principles of Design Driven Design (DDD), the Domain Services Layer will contain a number of business-oriented modules. These modules will correspond to the respective business areas in the core legacy system. Hence, each module will encapsulate several microservices that provide business-specific functions. In line with DDD implementation patterns, each module will have a service manager to manage its microservices. Moreover, the service manager will enable collaboration among the business modules as required. Service calling in this instance is achieved via the communication between the service managers of the different system modules.

3) Common layer: The Common Layer will provide generic services required by the application. For example, user management and authentication, integration with external parties, and data access. This layer will also comprise a dedicated Data Access Layer (DAL) that will provide all the data connectivity and access functionalities. Having a separate DAL is vital in the context of legacy transformation as it will enhance the adopted SOA approach by avoiding native access to databases.

C. Workflow Management Layer

One of the important elements of the proposed architecture is based on wrapping legacy solutions with Business Process Management (BPM) functionality. BPM functionality will be managed via a Workflow Management Layer that will act as an orchestration tool to manage the interactions between the legacy core system and its related end points. Following the proposed layered approach, a workflow management software component will sit on top of a dedicated service middle layer. Hence, the Workflow Management Layer will comprise three main components as follows:

1) Process workflows: The process workflows capture the workflows of the legacy system and manage the system services accordingly. It should be noted that domain driven workflows will not only capture the business-level models of the legacy system, but the approach adopted by [15] will be employed in which IT-level models will also be captured for effective workflow management. Hence, in addition to capturing the workflows of business processes, IT-level models will also be captured to address specific technical requirements such as infrastructure considerations and user access and authentication.

2) Workflow and data views: Workflow models constitute process models (views) that capture the actual sequence of activities/validations that a typical workflow process contains [16]. Workflow views will be used to create and manage the workflows that map the legacy system's functionality. Another layer of workflow modelling within the proposed architectural model is the data models (views). Data views contain the data objects required to define data fields, data field mapping, and connections to database. Hence, they bridge the connection between the backend database systems.

3) Artificial intelligence (AI) Layer: This layer will aim to streamline the system's workflows, reduce redundancies by intelligently handling large amounts of data, reduce user errors and increase the efficiency of routine tasks. It will offer the greater advantages of the legacy transformation process.

D. User Interface Layer

The User Interface (UI) Layer will provide users with the ability to interact with the backend legacy system to achieve the required business functionality. Frontends can be in the form of purpose-built desktop applications communicating with the Workflow Management Layer. They can also be in the form of web-based applications whether it be a website, web portal, tablet, or mobile applications.

IV. Technology Stack

There is a variety of technology solutions that can support a gradual transformation from legacy systems to modern web-based solutions. Based on the proposed workflow-managed SOA approach, the following software technologies can be utilized for a prototype implementation:

A. Pega Platform

Since one of the pillars of the proposed approach is the utilization of a Workflow Management Layer, the use of Pega

as a Business Process Management (BPM) platform is suggested. The low-code nature of Pega coupled with its App Studio that allows for business and IT cooperation in the design stage makes it a powerful tool for transformation and modernization projects [17].

Another reason to choose Pega is its wide range of data and integration capabilities that allows connecting Pega applications with distributed backend systems. The Pega platform also supports a wide spectrum of integration standards and protocols allowing for high connectivity levels with external systems [18]. Additionally, Pega offers a wide range of AI and machine learning tools that allow for optimized workflows and increased efficiency [19]. These capabilities are particularly important in relation to the proposed architecture which emphasizes communication with legacy systems via dedicated integration layers.

B. ASP.Net Core MVC

ASP.Net Core MVC is a modular and cross-platform development framework for developing web-based applications [20]. It provides a concrete framework for developing RESTful web services that can expose data operations [21]. This development framework was selected for the prototype implementation due to its ability to expose backend services as RESTful web services that can be consumed by other software layers (namely, the Workflow Management Layer in our proposed architecture).

Developing a middle layer using ASP.Net Core MVC can provide the required flexibility in terms of transforming a legacy system to a web-based system. Within this context, the main advantage of ASP.Net Core MVC is its ability to provide headless web services [21]. Headless API services do not have User Interface (UI) as they are meant to be consumed by other systems that may have their own UI elements. This approach provides the necessary flexibility to expose system services via different interfaces such as websites, web portals, and mobile applications.

C. Devart

Devart is a database connectivity tool that supports a variety of database platforms. To avoid direct (native) access to the legacy database, Devart can be a good tool for building Data Access Layers (DAL) that can provide the necessary interfaces to the service layer to access legacy databases. Furthermore, Devart's developer tools support reverse and forward engineering which makes it a suitable tool for legacy modernization implementation [22].

V. Solution Implementation

To identify the exact components of the proposed architectural approach, a hypothetical proof of concept is presented in this paper. We examine the effectiveness of the proposed approach through a prototype based on one of the common legacy transformation scenarios. This scenario is represented in the transforming of an Oracle forms-based system (Oracle 5) to a web-based application. In this proof of concept, the scenario of exposing a legacy HR system to the web is highlighted through the implementation of the proposed transformation approach.

A. Transformation Steps

A piece-by-piece transformation process is followed as opposed to a risky big bang approach where all system components are migrated at once to a new system/platform. Based on this gradual approach, two main transformation stages can be envisaged:

1) Phase I: Transformation to Service Orientation (SO): The main goal of this phase is to transform legacy services into modular services that can be accessed from a service-oriented middle layer (service layer). To achieve this goal, the legacy system will be analyzed and documented on an as-is basis. Then, business logic will be captured in the service layer that will directly interact with the legacy system and its database (acting as a backend system in this instance).

2) Phase II: Legacy system replacement: Since Phase I will separate business logic from the legacy core system, it will be relatively a lower risk process to replace the backend legacy system with a new system that will interact with the existing service layer. In such a scenario, business operations and end-user experience will not be interrupted as they will still be interacting with the same frontend systems. Such frontend systems can be either a workflow management interface as manifest by the proposed architecture, customized desktop applications or web-based applications.

B. Case Study

The prototype system includes four layers (Core Legacy System, Service Layer, Workflow Management Layer and UI Layer). In this context, it is assumed that the Legacy System's services were mapped into several clearly defined APIs that can loosely integrate with other systems. As illustrated in Fig. 2, the API services act as an entry point and perform the required services using the business modules' service managers. Furthermore, the API service layer will use the Common Layer for generic functions such as the management of user authentication, getting database context, etc. Additionally, API services will share the database context with all business services allowing the system activities to be handled in a few database transactions.

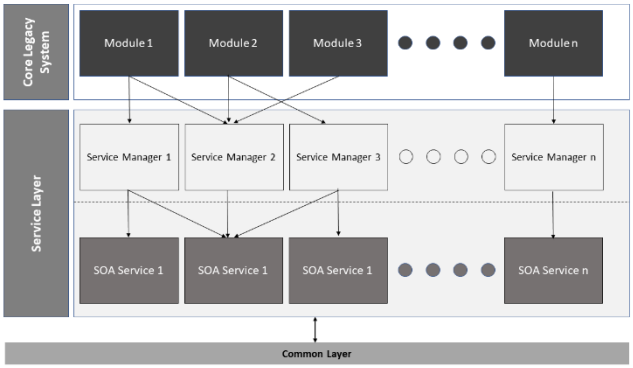


Fig. 2. Modules and Service Managers' Interactions.

C. System Services

Table I summarizes the main API Layer services that require the basic HR system functionality. If we take the example of a simple service to enter the details of a new employee, several RESTful API services can be used to populate the dropdown menus used in the data entry form within the system's UI. These services will be called via the designated service manager to pull the required data for use in the UI. Furthermore, business validation can be performed using services from several modules based on the action being performed by the end user.

D. Workflow Management

The legacy system functionality is decomposed into service-oriented workflows managed by the Workflow Management Layer. In our prototype, the Workflow Management Layer is represented in Pega BPMS, which is a low code workflow management platform that has the flexibility to integrate with a variety of backend systems. Fig. 3 illustrates the Pega-designed workflow for the employee addition process within the prototype HR system. This workflow contains three main steps: identification of employee details, employee addition, and closure.

TABLE I. API LAYER SERVICES

System Functions	**Service APIs (CommonDBContext = InfraSrctureManage.GetDatabaseContext)**
Search Function	API. Search Employee In this service, the user inputs the search attributes and then calls the following service to fetch the required results: HRSrvcMngr.GetEmployeeList
Add New Employee	API.NewEmployee: A data entry form allows the user to enter the attributes related to a new employee. The following services are used to fill the dropdown lists associated with employee attributes. For example: RefSrvcMngr.GetEmployeeTypes(CommonDBContext) RefSrvcMngr.GetContractTypes(CommonDBContext) RefSrvcMngr.GetContractTypes(CommonDBContext) RefSrvcMngr.GetDepartmentsList(CommonDBContext)
Update Employee Details	There are two steps in this process: 1. Show employee details, attributes are initiated by fetching the current employee data by using the service: HRSrvcMngr.GetEmployeeByID 2. Users can update the employee's details and then either save the record or cancel the process.
Delete Employee Details	API. DeleteEmployee The API.SearchEmployee can be used to fetch the employee details. Once an employee record is selected, it can be deleted by using the following service: SgnSrvcMngr.DeleteEmployee(CommonDBContext)
Save Details	API. SaveEmployee Saving an employee's details can be done through either of two processes: 1. Adding a new employee: HRSrvcMngr.CreateEmployee(CommonDBContext) 2. Updating employee details: HRSrvcMngr.UpdateEmployee(CommonDBContext)
Cancel Operation	This functionality will be achieved via the UI level by clearing the data entry form

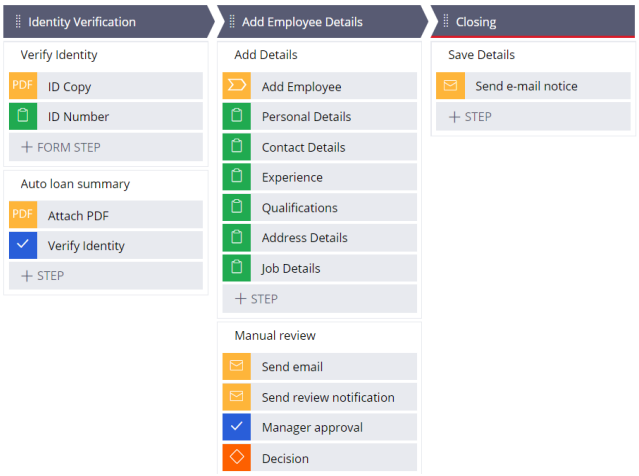


Fig. 3. Proposed Architecture.

The first step in the workflow involves the verification of the employee's identity before adding his details to the system. To verify the identity, it is assumed that Pega will integrate with a third-party provider to validate the employee's ID. In our prototype implementation, Pega capabilities are utilized to create the system's frontend in the form of a series of HyperText Markup Language (HTML) pages that correspond to the designed workflow.

Once the employee's identity is validated, the "Add Employee Details" step is invoked which will involve displaying the employee addition data entry form shown in Fig. 4. The different drop downs in the entry form will be populated with dynamic values pulled from the legacy system database. For example, the following services will be called to fetch Employee Types, Contract Types and List of Departments:

RefSrvcMngr.GetEmployeeTypes(CommonDBContext)

RefSrvcMngr.GetContractTypes(CommonDBContext)

RefSrvcMngr.GetDepartmentsList(CommonDBContext)

Once the employee details are added, they can be saved by invoking another service from the Service Layer:

HRSrvcMngr.CreateEmployee(CommonDBContext)

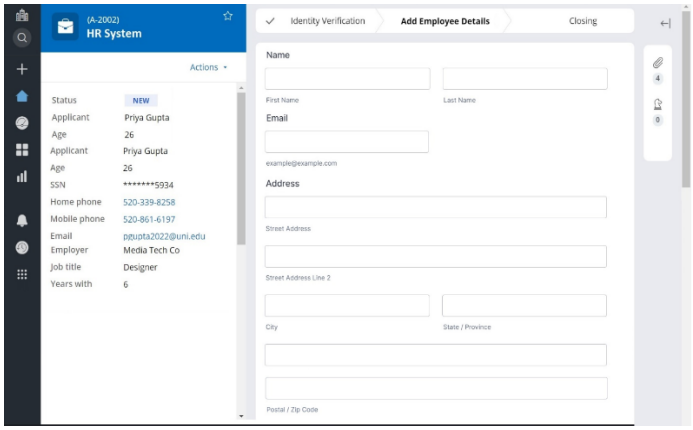


Fig. 4. Employee Addition Form.

E. Integration

Pega integration capabilities with RESTful APIs were utilized to integrate with the services provided by the Service Layer. In this scenario, Pega acted as a client application that uses HTTP protocols to access GET or POST methods to achieve the required functionality. An example of RESTful service consumption is the process by which the list of contract types is fetched to populate the relevant dropdown list in the employee addition form. HTTP GET requests are passed through service HTTP query strings that contain the required operations. In this example, GetContractTypes service is used:

https://www.legacyhr.com/GetContractTypes.php?operation=fetchtypes

The fetched data is formatted into JavaScript Object Notation (JSON) string that can be easily used in the system's frontend as illustrated in Fig. 5. Similarly, when there is a need to write data to the legacy system's database, HTTP POST operations can be used to pass the required data (for example, new employee's details) to the core system.

```
{
  "status"   : "SUCCESS",
  "count"    : "3",
  "contract_types"  : [
            {
              "id" : "1",
              "name" : "Fulltime",
              "category" : "contracts",
              "description" : "Contract for permanent staff"
            }, {
              "id" : "2",
              "name" : "Parttime",
              "category" : "contracts",
              "description" : "Contract for temporary staff"
            }, {
              "id" : "3",
              "name" : "Projects",
              "category" : "contracts",
              "description" : "temporary contracts for projects "
            }
             ]
}
```

Fig. 5. Contract Types JSON Sample.

VI. Conclusion and Future Work

A. Conclusion

The rapid advancement of web technologies coupled with evolving business needs make legacy transformation inevitable for enterprises around the world. However, the risks of such a transformation should be mitigated with an approach that is flexible enough to allow for a gradual and low risk transformation.

The proposed SOA workflow-based transformation approach offers several benefits in terms of legacy system transformation into web-based applications. The key advantage here is the adoption of a microservices architecture where the legacy system's functionality is decomposed into self-contained functional units. On top of that, an AI-enabled Workflow Management Layer orchestrates the system's functionality by calling the required legacy services from a dedicated Service Layer (middle layer). In our prototype implementation, we utilized ASP.Net Core MVC for the Service Layer implementation and Pega BPMS for the Workflow Management Layer.

B. Future Work

Future work will involve progressing further with the transformation approach by examining the process of replacing the legacy backend system with a new core system. The aim here will be to validate the success of full transformation by utilizing the suggested architectural approach and transformation steps.

References

[1] H. Kim and Y.-K. Chung, "Transforming a Legacy System into Components," In: Gavrilova M. et al. (eds) Computational Science and Its Applications - ICCSA 2006. ICCSA 2006. Lecture Notes in Computer Science, vol. 3982, pp. 198-205, 2006.

[2] H. M. Sneed, "Planning the reengineering of legacy systems," IEEE Software, vol. 12, no. 1, pp. 24-34, 1995.

[3] A. D. Ionita, M. Litoiu and G. Lewis, Migrating Legacy Applications: Challenges in Service Oriented Architecture and Cloud Computing Environments, 1 ed., IGI Global, 2012.

[4] R. Khadka, B. V. Batlajery, A. M. Saeidi, S. Jansen and J. Hage, "How do professionals perceive legacy systems and software modernization?," in ICSE 2014: Proceedings of the 36th International Conference on Software Engineering, Hyderabad, 2014.

[5] P. Gordon, R. Seacord, D. Plakosh, G. Lewis and J. Fuller, Modernizing Legacy Systems: Software Technologies, Engineering Processes, and Business Practices, 1 ed., Addison-Wesley Professional, 2003.

[6] D. Goerziga and T. Bauernhansla, "Enterprise Architectures for the Digital Transformation in Small and Medium-sized Enterprises," in Procedia CIRP, Naples, 2018.

[7] H. M. Hess, "Aligning technology and business: Applying patterns for legacy transformation," IBM Systems Journal, vol. 44, no. 1, pp. 25-45, 2005.

[8] Y. Zou and K. Kontogiannis, "Migrating and Specifying Services for Web Integration," Lecture Notes in Computer Science, vol. 1999, pp. 253-270, 1999.

[9] S. Comella-Dorda, K. Wallnau, R. C. Seacord and J. Robert, "A Survey of Legacy System Modernization Approaches," Defense Technical Information Center, 2000.

[10] S. B. Popov and P. V. Khripunov, "Digital Transformation Legacy Social Service Information System," in Journal of Physics: Conference Series, Britsol, 2019.

[11] L. Erlikh, "Leveraging legacy system dollars for e-business," IT Professional, vol. 2, no. 3, pp. 17-23, 2000.

[12] G. Canfora, A. Cimitile, A. De Lucia and D. L. Giuseppe, "Decomposing legacy programs: a first step towards migrating to client-server platforms," The Journal of Systems and Software, vol. 54, no. 2000, pp. 99-110, 2000.

[13] S.-H. Li, S.-M. Huang and D. C. C. C.-C. Yen, "Migrating legacy Information," Journal of Database Management, vol. 18, no. 4, pp. 1-25, 2007.

[14] P. Krivic, P. Skocir, M. Kusek and G. Jezic, "Microservices as Agents in IoT Systems," in 11th KES International Conference, KES-AMSTA 2017, Algarve, 2017.

[15] M. C. Branco, J. Troya, K. Czarnecki, J. K¨uster and H. V¨olzer, "Matching Business Process Workflows," in Model Driven Engineering Languages and Systems, Innsbruck, 2012.

[16] W. Yang and F. Li, "Workflow modeling: a structured approach," in 8th International Conference on Computer Supported Cooperative Work in Design, Xiamen, 2004.

[17] Gartner, "Gartner Magic Quadrant for Enterprise Low-Code Application Platforms," 2019. [Online]. Available: https://www.gartner.com/en/documents/3991199/magic-quadrant-for-enterprise-low-code-application-platf. [Accessed 30 10 2021].

[18] S. Mangu, "Business Process Management: Robotic Process Automation Approach," International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 11, no. 11, pp. 831-840, 2020.

[19] R. Walker, "Artificial Intelligence in Business: Balancing Risk and Reward," Pegasystems, 2018.

[20] J. Ciliberti, ASP.NET Core Recipes: A Problem-Solution Approach, 2 ed., Apress, 2017.

[21] A. Troelsen and P. Japikse, Pro C# 7: With .NET and .NET Core, 8 ed., Apress, 2017.

[22] H. Schwichtenberg, Modern Data Access with Entity Framework Core, 1 ed., Apress, 2018.

Developing the Mathematical Model of the Bipedal Walking Robot Executive Mechanism

Zhanibek Issabekov, Nakhypbek Aldiyarov
Satbayev University, Almaty
Kazakhstan

***Abstract*—The paper considers the accuracy of footstep control in the vicinity of the application object. The methodology of forming a simulation of the executive electro-hydraulic servomechanism is developed. The paper presents control algorithms in the dynamic walking mode. The issues of stabilization of the sensors installed in the soles are investigated. The description of the laboratory model and simulation of the main links of the exoskeleton, approximated to human parameters, allowing to insert the studied algorithms of motion of the executive mechanism into the program of automation of calculations of the links of motion are given. The authors for the first time simulated the bipedal walking robot using modern digital technologies, including the joint use of pneumatic electric drive. This paper proposes an automated control scheme for manipulators controlling immobilized human limbs. Considering the functions of the leg and the phases of movement, the structure scheme is chosen so that the same actuator performs several functions. This construction partially reduces the load on the person, because the drives of the various links due to their gravity can overturn a person. Using the kinematic structure of the model and the method of adaptive control of the manipulator, as well as replacing some movement parts with plastic material, the authors were successful in reducing the total weight by three times compared with foreign analogues, which is important for a sick person.**

Keywords*—Exoskeleton; manipulator; model; kinematics; dynamics

I. Introduction

In the 21th century, there arise more and more situations that demand from people with disabilities to perform a wide variety of works related to daily life. To help people, mobile robotics tools are being created which can be controlled via radio or cable. Relating to a person, being near the power source, he can control his functions by cable. However, as life shows, in order to expand human capabilities associated with his movement, control can be carried out by radio channel. Many scientists work for creation of the motion control algorithms of bipedal walking robots that move in various dynamic modes of movement: walking, running, jumping, etc., in other words, human motion simulation, which formed the basis for the study of using the theory for creation of mechanisms for energetically optimal regulation of human walking [1-2].

For these reasons, scientists develop exoskeletons that represent a technical device designed for physical relief of a person who performing different works thanks to load accommodation by the exoskeleton, provided that it repeats human biomechanics [1-3]. As one reason for the growing popularity of these devices, it is necessary to mention the areas of their possible application:

1) Military sphere.
2) Use by people with disabilities.
3) Elimination of consequences of various emergencies.
4) Use of heavy equipment in conditions of inapplicability.
5) Use in operations where it is possible to replace heavy equipment with human labor.

Exoskeletons are classified according to the following [3-5].

1) According to the power source and the drive operating principle: passive exoskeletons, active exoskeletons;
2) According to the drive: electric, pneumatic, hydraulic.

Exoskeletons of today are becoming a very powerful tool to assist soldiers and medical staff of specialized clinics with rehabilitation of patients who have suffered from limb diseases. It was found that the most of designed exoskeletons cannot be used for rehabilitation of patients with limited functions of the upper and lower limbs due to the large mass of the structure, dependence on external power sources and their significant cost. For designing the exoskeleton, it is necessary to solve a lot of technical problems, among which are the following: the problem of walking control; designing the executive mechanism using original design solutions; designing the system of interconnected drives; designing the power source with a high specific power; designing the system of sensitivity, orientation and navigation; designing the control system and designing its algorithms. As of today, there is a lot of versions of exoskeletons built using various drives (electric drive, hydraulic drive, pneumatic drive), but their practical use is very limited due to difficulties associated with an on-board energy source that can provide the exoskeleton with autonomy. However, this fact only serves as the progress accelerator and leads to the constant appearance of more and more new versions of exoskeletons [4-6]. The purpose of the research is the use of exoskeletons in medicine for the rehabilitation of patients with musculoskeletal direction. This article describes the mechanisms for controlling the feet and the center of mass of a human robot.

II. Relevance

The works of many scientists considered the problems of foot elasticity when modeling human movements [5-6], [5-6],

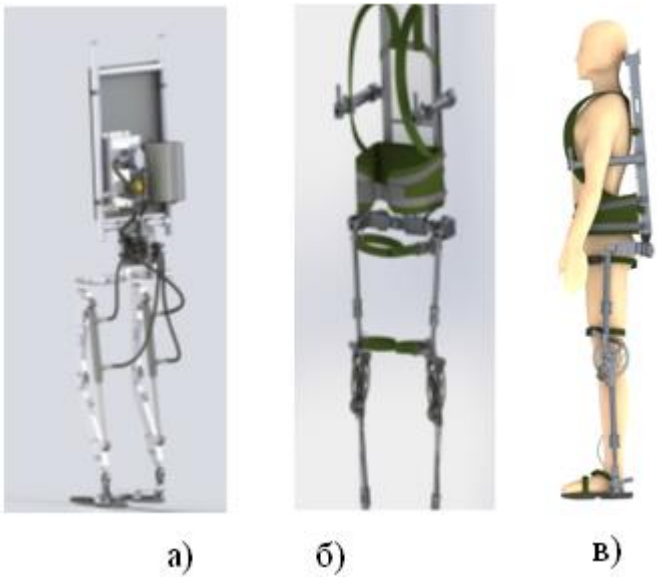
dynamic motion of the manipulator based on the closed kinematic chain with coordinate-parametric control, two-engined electric drives of coordinated rotation based on an asynchronous-valve cascade, dual-power machines [7-8], etc., developments of which facilitated the development of investigations in the field of the manipulator use for replacing the human legs. There are many examples in the world of using the developments of scientists to improve the situation of persons disabled from childhood and work, to return them to normal life, their participation in the Paralympic Games, etc. Fig. 1 shows the development of scientists of Moscow universities, active and passive exoskeletons, as well as the exoskeleton appearance on a man simulator.

It is of interest the ReWalk exoskeleton developed in Israel by ARGO Medical Technologies which allows people with paralysis of the lower half of the body (lower paraparesis) to stand on their feet and walk using sticks (Fig. 2).

The design of ReWalk exoskeleton is based on sensors that detect the body tilt ahead and transmit a signal to devices supporting the legs [4-7]. The power source is provided by a battery placed in a special backpack. It has 2 degrees of mobility. The structure operation is possible only for persons with preserved functions of the upper limbs.

A special feature of this layout diagram is installing 4 electric motors in the knee and hip joints of the exoskeleton, the ankle position is regulated by the springs that allow the foot to naturally stand on the ground (Fig. 3b).

A significant disadvantage of ReWalk exoskeleton is the complexity of ensuring the mechanism balance and hereto related safety problems of operator`s motion [9]. In most cases, these structural versions of exoskeletons are used as simulators for people undergoing rehabilitation.



(a) active exoskeleton; (b) passive exoskeleton; (c) exoskeleton on a man simulator.

Fig. 1. D-model of Active and Passive Exoskeleton.

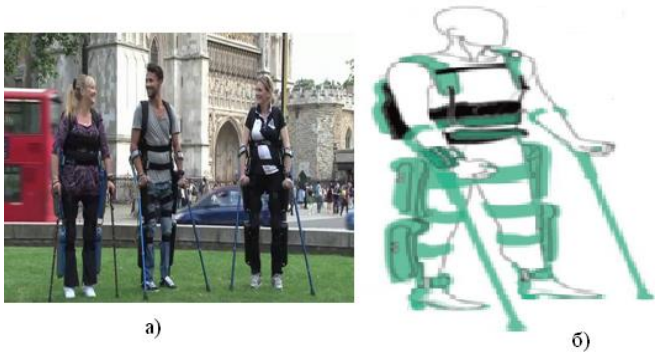
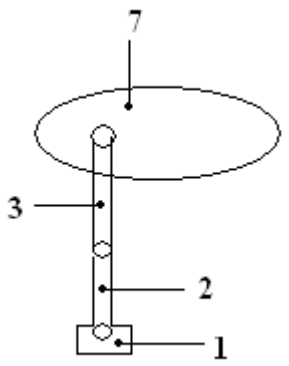


Fig. 2. ReWalk Exoskeleton (a) and its Layout Diagram (b).

As far back as 1948, the Russian professor N. A. Bernstein drew a man with prosthetics copying the leg skeleton but with electric motors, which was the development of the Prosthetics Research Institute. It is interesting to note that just after the war ended, it was a very vital invention which, unfortunately, had no practical continuation in future. In the 60`s, General Electric developed this idea but in a version of a full-fledged skeleton with hydraulic control. The same attempt was made by the Russian side in Russia (Leningrad, 1970) [5-6]. In Kazakhstan, such works are under development, so the study presented in this article is very relevant.

III. Methodology

A large number of investigations were devoted to the study of kinematics and dynamics of manipulation robots [7-9]. The manipulator executive mechanism has the treelike kinematical structure, a large number of mobility steps, does not properly secure to the fixed base, and during movement, it is connected with the supporting surface and ambient objects [10-11]. The authors for the first time develop a similar model with the use of modern digital technologies including the joint use of pneumatic-electric drive. This paper offers the automated control scheme for manipulators that control the immobilized human limbs. It is known that human manner of walking just as a robot determines by trajectories of the pelvis (main but dependent movement) and feet (auxilliary but forced movements). The main movement is carried out as a result of moving the legs [12-13]. The schematic structure of human musculoskeletal system is similar in many ways to the movement of bipedal walking robot which contains a system of solids and is in the form of closed kinematic chains with rotating and translating kinematic pairs connected to the body. Fig. 3 shows the model of human foot control in which relative movements of segments are carried out by drives. The leg mechanism has several degrees of mobility. Taking into account the leg functions and movement phases, schematic structure is chosen so that one and the same drive performs several functions. This construction partially reduces the load on the person, because the drives of the various links due to their gravity can overturn a person.



1- foot; 2-lower leg; 3-upper leg; 7- pelvic bone

Fig. 3. Human Leg Model.

Methods for describing the kinematic structure (KS) were developed for formation of mathematical models of the executive mechanism [1, 11].

To set the kinematic structure and write the kinematic expressions, the designations and indices given below were used. Kinematic connections of each segment are characterized by:

- number of one of the previous segment;
- numbers of one or more of subsequent segments;
- its sequential number for the previous segment.

Each segment is associated with as many coordinate systems as there are subsequent segments but not less than one. One of them is taken as the main one, which is assigned the number 1, the rest are auxiliary. All of them are assigned in accordance with the traditional Denavit-Hartenberg rules for robotics [1, 9]. Vector values can be set in different coordinate systems (CS). Its upper left index indicates in which CS the given vector is specified. If this index is missing or null, then the vector is set in the base CS. If a nonzero number is specified, then the vector is specified in the main coordinate system of the segment, number of which is the specified number. If two numbers separated by commas are - ${}^{i,j}\overline{X}$, then the vector is specified in j^{-th} the auxiliary coordinate system i. Let us denote by $L = \{1,2,3 \dots n\}$ an unordered set, the elements of which are the numbers of the executive mechanism segments.

The following index functions are introduced: [11-13].

1) f(i) - segment number that is the parent segment for segment i.

2) s(i,k) –segment number that is k-th son segment for segment i.

3) dg+ (i) – semidegree of segment i that defines the number of son segments of i segment.

4) D(i) – cortege of numbers-links that are son segments for i segment$\left(\Gamma(i) = \left\{\left(s(i,1), s(i,2), \dots, s(i,k), \dots, s\left(i,i,dg^{+}(i)\right)\right)\right\}\right)$;

5) ns(i) –defines what number as son segment has i segment for its father segment (sequence number of i segment in the cortege. $\Gamma\left(f(i)\right)$.

6) nan(i) — defines the number of ancestors of i segment.

7) an(i,j) — determines the number of the j-th from the root of the ancestor of the link i.

If i segment is more than 0, it is possible to write:

$$ns\left(s(i,k)\right) = k\ ; i = f\left(s(i,k)\right);\ k \in \{1,2,3 \dots dg^{+}(i)\} \qquad (1)$$

As it is for each segment$i(i \in L)$, we define the vector connecting the beginning of the father segment with the beginning of the son segment; the vector of the centre of mass (CM) position; the coefficient defining the type of articulation of main segment i with father segment; the type of ancestor of segments – the rotary joint 1; the type of articulation of segments - the telescopic joint 0.

This is how the exoskeleton robot segments are described and defined. However, if you transfer this system to a person, then you need to note the following: if the robot can turn and rotate according to the operator's instructions, then the person is limited in his motions. For example, the human foot (especially a sick person) cannot turn to 180°, as well as the knee and other parts of the lower limbs. Therefore, it is necessary to develop the motion algorithm limiting turns of feet from zero to 90°.

The method for describing kinematics of the bipedal walking robot executive mechanism was proposed by Denavit and Hartenberg with compatible coordinate systems all parameters of whose are presented in Fig. 4 [11-13].

The calculation system [4] is based on the use of homogeneous transformations matrices (4×4), which give unambiguous and clear rules for constructing a mathematical model of the robot's executive mechanism. At the same time, the number of parameters included into the matrix relative to the position of the successive segments of the executive mechanism is minimal. The matrix form Ai is identical both for rotational and translational joints.

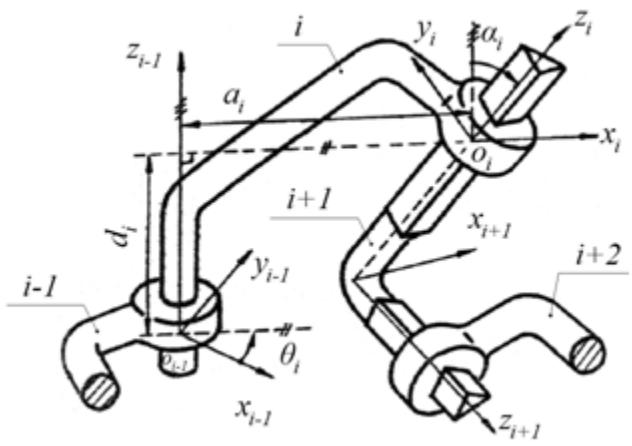


Fig. 4. Denavit-Hartenberg Compatible Coordinate Systems.

The advantage of this method of constructing connected CS is that you can specify only four parameters that determine the relative position of two consecutive CS i-1 and i, and, consequently, the conversion matrix Ai. Two consecutive coordinate systems of segments, for example, $i-1$ and i, can be always coincided using a rotation [14], two transfers and another rotation carried out by the $i-1$ coordinate system in the following order:

- Rotate by an angle θ_i around the axis z_{i-1} in the positive direction until the axes x_{i-1} and x_i become parallel and equally directed. If the joint is rotational, then the angle θ_i coincides with the generalized coordinate.
- Relocation over the distance d_i along the axis z_{i-1} until the axes x_{i-1} and x_i will coincide. If the joint is translational, then the d_i coincides with the generalized coordinate.
- Relocation over the distance a_i along the axis x_i until the coordinate origin will coincide. Parameter a_i is the constructive constant of the mechanism (depends on geometry of the structure).
- Rotation by the angle α_i αi about the axis x_i until all axes coincide.

Note that angles are positive if they are counted counterclockwise around the specified axes, and linear displacements are positive if they coincide with the positive directions of the corresponding axes.

IV. Result and Discussion

As a result of these movements the coordinate system $O_{i-1}x_{i-1}y_{i-1}z_{i-1}$ sequentially take up the possitions $O'_{i-1}x'_{i-1}y'_{i-1}z'_{i-1}$, $O''_{i-1}x''_{i-1}y''_{i-1}z''_{i-1}$, $O'''_{i-1}x'''_{i-1}y'''_{i-1}z'''_{i-1}$, and finally reaches the position $O_ix_iy_iz_i$. Moreover, each subsequent coordinate system is characteristed in the previous coordinate system by 4x4-coordinate transformation matrices sequentially.

From the four parameters θ_i , d_i , a_i , α_i (Fig. 4), two parameters a_i and α_i are always constant and determined by the design of the robot executive mechanism. One of two other parameters (θ_i or d_i) is a variable parameter. For the rotational joint, the value θ_i characterizes the angle of relative rotation of i-1 and i segments, and the liner value d_i is constant. Reverse, for the telescopic connection, d_i is a variable. The variable of i-th joint (θ_i or d_i) is usually called the generalized coordinate of the robot executive mechanism [15-18].

Each limb has three degree of freedom and is derived by an engine with mechanism (transmission, gearbox, reduction unit). In the lower part of the limb there are three force sensors for measuring the leg force response.

Using the methods of describing the kinematic structure there were made the prerequisites for adaptive control of the executive mechanism for motion of any part of the human body. The authors considered the akinetic human limbs as a non-stationary object (NO). Some scientists propose to use the adaptive identification of NO with Markov parameters on the example of the bipedal walking robot. However, this NO is considered as a non-stationary dynamic multi-dimensional multi-loop controlled object with inputs p and outputs q (Fig. 5) [19-21].

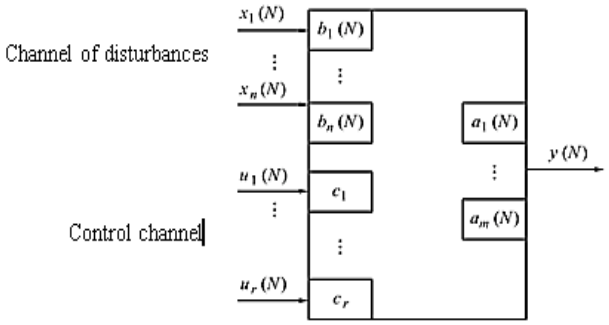


Fig. 5. Scheme of Non-stationary Dynamic Multi-dimensional Multi-loop Controlled Object.

There are two fundamentally different types of input channels at the object input. The perturbation channel is formed from n observed inputs:

$X^T(N) = (x_1(N), \ldots, x_n(N))$, multi-loop controlled object $X \in R_1^n$ where R_1^n is a set of alloweble inputs with unobserved unknown parameters bj (N), $j = \overline{1, n}$.

The control channel is also observed: UT(N) = (u1(N), ..., ur(N)), $U \in R_2^r$ where R_2^r − a set of acceptable controls, but unlike the perturbation channel, all its parameters are known: $a_1(N) \ldots a_m(N)$, then $Y \in R_3^q$, where R_3^q − a set of acceptable outputs [13,22].

Using Denavit-Hartenberg method and the adaptive method of controlling the executive mechanism of human movement, the authors developed two variants of variables for the human exoskeleton changing d or Q parameter.

These models were put into Program of calculation and received (Table I).

The table shows the results of eight from 21 steps.

TABLE I. Values of Simulated Parameters of Exoskeleton Segments

No. of step	Q, rad	d, м	a, м	α, rad	$f(i)$	$ns(i)$
1	$-\pi/2$	0	0	$-\pi/2$	0	1
2	$-\pi/2$	0	0	$-\pi/2$	1	1
3	$-\pi/2$	0	0	$-\pi/2$	2	1
4	$\pi/2$	0	0	$\pi/2$	3	1
5	$\pi/2$	0	0	$\pi/2$	4	1
6	0	−0,39	0	$-\pi/2$	5	1
7	$\pi/2$	−0,1	0,17	$\pi/2$	6	1
8	0	0	0	$-\pi/2$	7	1

It can be concluded from the table that angle Q changes from $-\frac{\pi}{2}, \frac{\pi}{2}, 0 \ldots d - 0{,}39; d = -0{,}1; d = 0$, etc.

TABLE II. RESULTS OF ARBITRARY POINTS OBTAINED ON MATLAB

$No.of\ step$	Q, rad	d, м	a, м	α, rad	$f(i)$	$ns(i)$
6	-π/2	0	0,237	- π/2	5	1
7	0	0,196	0,163	π/2	6	1
8	π/2	-0,109	0	- π/2	7	1
9	π/2	0	0,487	0	8	1
10	0	0	0,669	π/2	9	1
11	π	0,195	-0,097	π/2	6	1
12	π/2	-0109	0	- π/2	11	1
13	π/2	0	0,475	0	12	1

It can be concluded from the Table II that angle Q changes from $-\frac{\pi}{2}, \frac{\pi}{2}, 0 \ldots d + 0{,}196; d = -0{,}109; d = 0$.

The authors developed and applied the Program to the model in Fig. 3.

REFERENCES

[1] Sh. Aszhan, "Development of an automated control system for the executive mechanism of a robot manipulator," Scientific and Technical Conference: Innovative Technology in Engineering. Almaty: ETU, pp. 176–179, April 2021.

[2] R.W. Jackson, C.L. Dembia, S.L. Delp and S.H. Collins, "Muscletendon mechanics explain unexpected effects of exoskeleton assistance on metabolic rate during walking", J. Exp. Biol., vol. 220, no. 11, pp. 2082–2095, 2017.

[3] N.T. Zhetenbaev, G.K. Balbaev, Zh. N. Isabekov and E.S. Nurgizat, "The future of robots with exoskeletonsin health care," Global science and innovations 2020: Central Asia, vol. 4, no. 3, pp. 26–32, 2020.

[4] Zh.N. Isabekov A.K. Kovalchuk and N.T. Zhetenbaev Exoskeletons of lower limitations: a brief explanation. Bulletin of KazATK, vol. 1, no. 108, pp. 78–84, 2019.

[5] J. Wu, J. Gao, R. Song, R. Li, Y. Li, L. Jiang, "The design and control of a 3D of lower limb rehabilitation robot," Mechatronics. vol. 33, pp.13–22, 2016.

[6] A.K. Kovalchuk, "Designing drives of a medical robot actuator," Life Science Journal, pp. 337–340, 2014.

[7] D. Huamanchahua, Y. Tasa-Aquino, J. Figueroa-Bados, J. Alanja-Villanueva, A. Vargas-Martinez, R.A. Ramirez-Mendoza, "Mechatronic Exoskeletons for Lower-Limb Rehabilitation: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronies Conference (IEMTRONICS), 2021. pp.1–8. doi: 10.1109/IEMTRONICS 52119.2021.9422513.

[8] G.V. Prado, R. Yli-Peltola, M.B.C. Sanchez, "Design and analysis of a lower limb exoskeleton for rehabilitation," Interdisciplinary Applications of kinematics, 2019, pp. 103–114.

[9] Zh.N. Isabekov, "Mathematical model of kinematics and dynamics of arbuscular executive machinery of proactive exoskeleton," Polytechnic youth magazine, Bauman MSTU, vol. 4, no. 4, pp. 5–10, 2016.

[10] A.K. Kovalchuk, "Modified Denavit-Hartenberg Coordinate System for Robot Actuating Mechanismus with Tree-like Kinamatic Structure," Science and Education of the Bauman MSTU, vol. 11, pp. 12–30, 2015.

[11] Zh.N. Issabekov, K.A. Moroz, M.F. Kerimzhanova, "Study of the dynamics of the exoskeleton actuating unit," Bulletin of BSTU named after V.G. Shukhov, no. 11. pp. 99–105, 2021. doi: 10.34031/2071-7318-2021-6-11-00-00.

[12] D.A. Elias, D. Cerna, C. Chicoma and R. Mio, "Characteristics of a lower limb exoskeleton for gait and stair climbing therapies," Interdisciplinary Applications of Kinematics, Springer, pp. 81–92, 2019.

[13] Z.N. Issabekov, I.K. Tsybrii, K.A. Moroz, "Organization of walking of the lower-extremity exoskeleton using the control of the supporting foot. Advanced Engineering Research," Series Machine Building and Machine science. 2021;21(3):247-252.

[14] Sh. Aszhan, "Requirements for the accuracy of control of the lower extremities of the exoskeleton near the object," International Scientific and Practical Conference: Modern Kazakhstan: reforms of education and science. Almaty: ETU, October 2021, pp. 248–251.

[15] A.K. Kovalchuk, "Method of mathematical description of the kinematics and dynamics of tree-like actuators of walking robots," Natural and Technical Science, vol. 5, no. 73, pp.87–90, 2014.

[16] A.K. Tanyildizi, O. Yakut, B. Tasar, "Mathematical modeling and control of lower extremity exoskeleton," Biomedical Research, vol. 29, no. 9, pp. 1947–1952, 2018.

[17] E.K. Lavrovsky, E.V. Pis'mennaya, P.A. Komarov, "On the problem of organizing the walking of the exoskeleton of the lower extremities using control deficit," Russian Journal of the Biomekhanics, no. 2, pp. 158–176, 2015.

[18] T.M. Mukhidinov and K.S. Sholanov, "The movement of the biped walking robots," European research, vol. 1, no. 2, pp. 9–14, 2015.

[19] A.Zh. Toygozhinova, Research and development of automated deployment of air ozonisation, PhD thesis, Almaty, pp. 98–109, 2017.

[20] Song S., Collins S.H. Optimizing Exoskeleton Assistance for Faster Self-Selected Walking. IEEE Transactions on neural Systems and Rehabilitation Engineering, vol. 29, pp. 786–795, 2021. doi: 10.1109/TNSRE.2021.3074154.

[21] N.T. Isembergenov, K.N. Taissariyeva, U.O. Seidalieva, V.V. Danilchenko, "Microprocessor control system for solar power station," News of National Academy of sciences of the Republic of Kazakhstan. Series of geology and technical sciences, vol.1, no. 433, pp. 107–111, 2019. https://doi.org/1032014/2019.2518-170X.13.

[22] E.K. Lavrovsky, E.V. Pis'mennaya, "On the lower limb exoskeleton regular locomotion under input control deficit," Russian Journal of the Biomekhanics, vol. 18, no. 2, pp. 208–225, 2014.

Data Backup Approach using Software-defined Wide Area Network

Ahmed Attia, Nour Eldeen Khalifa[0000-0001-8614-9057], Amira Kotb
Information Technology Department
Faculty of Computers and Artificial Intelligence
Cairo University, Giza, Egypt

***Abstract*—Over the past several years, the traditional approaches of managing and utilizing hybrid Wide Area Network (WAN) connections, between sites across geographical regions, have posed many challenges to enterprises. Software-Defined Wide Area Network (SD-WAN) has emerged as a new paradigm that can overcome the traditional WAN challenges like the lack of visibility for WAN bandwidth utilization and the inefficient usage of expensive WAN resources. The flexibility and agility brought to WAN by applying the SD-WAN paradigm helped to improve the efficiency of bandwidth utilization and to address the surge of bandwidth demands. The SD-WAN capabilities become essential for meeting the heavy inter-data center's traffic exchange required for business continuity and disaster recovery operations. In this paper, a data backup approach is introduced using SD-WAN that makes the network centrally programmable. This will leverage the ability to make fine-grained traffic engineering for different data flows over WAN to optimize the bandwidth utilization of the expensive WAN resources by balancing the traffic load across network links between data centers and to minimize the time required to transfer backup data to disaster recovery sites. The proposed approach proved its efficiency according to the bandwidth utilization if it is compared to the other related works.**

Keywords*—Wide area networks; software defined network; software defined wide area network

I. Introduction

The Wide Area Network (WAN) is essential nowadays for large enterprises and businesses. It is a critical component to connect between data centers and different sites across geographies [1]. The first public WAN was deployed in early 1980 [2] and designed to connect between different sites using leased lines that had limited speeds and high cost.

In the previous decades, different WAN technologies have been developed to provide better service quality in terms of cost and speed such as VPN (Virtual Private Network), ATM (Asynchronous Transfer Mode), and MPLS (Multi-Protocol Label Switching). Despite the improvements in the bandwidth and speed of WAN networks and internet services, they are still congested with high traffic loads that cause data loss and jitters [3] which impact the performance and quality of the provided services.

During the COVID-19 pandemic, network traffic spikes have increased obviously and start to impact the services provided by different internet and cloud companies such as Amazon, YouTube, and Netflix [4]. To address these challenges and the surge in bandwidth demand, enterprises start to use technological innovations such as Software Defined Networks.

Software-Defined Networks (SDN) has emerged over the past years as a new promising model that simplifies network management. SDN aims to provide centralized management for multiple network devices like routers, switches, and firewalls by separating the control plane from the data plane [5]. The SDN framework consists of three layers. The lowest layer is the data plane layer that contains the network elements and devices responsible for packet forwarding. One layer above, the control plane is located where network intelligence is logically centralized to maintain a global view of the network.

The SDN controllers communicate with network devices to guide them in handling data packets and to execute network policies and rules. The application layer contains the network applications that introduce new network features and applications such as load balancing, network statistics monitoring, and security [6]. The application layer makes use of the holistic view of the network provided by controllers to get information about different data flows and give the appropriate guidance to the control layer for enhancing the network performance. Currently, there are several SDN controllers provided such as OpenDaylight [7], POX [8], NOX [9], and Beacon [10] that can simplify the management of network devices.

For the southbound interface, the OpenFlow protocol is the most used open protocol for the interaction between the SDN controllers in the control plane and SDN OpenFlow compliant switches in the data plane [6]. The OpenFlow SDN controller will insert flow entries on the OpenFlow compliant switches to instruct them for forwarding the incoming data packets. The OpenFlow switches will follow the basic packet forwarding mechanism to handle the incoming packets. They will check the packets header and match it to the entries inserted by the SDN controller in the flow table. In case no flow entry matches the incoming packet header, the OpenFlow compliant switch will notify the controller to identify the correct action for the packet and install the appropriate flow entries for it. The SDN controller can also interact with OpenFlow switches [11] to retrieve statistical information about data flows.

On the other side, the northbound interface is provided by the SDN controller to the application layer for network programmability. Applications neither have direct interaction

with data plane devices nor are aware of network topology [12]. They interact with controllers to make fine-grained traffic engineering, based on the data flow statistics controller can retrieve, and guide them with the path layout of the data flows [13].

The SDN paradigm is usually used in data centers local area networks (LAN) but recently it is used in wide area networks (WAN) to mitigate the challenges mentioned before. The Software-Defined Wide Area Networks (SD-WAN) is built on the concepts of SDN. The WAN networks will obtain many technical advantages by using the SDN concepts [14] that will help to leverage its performance, utilization, and efficiency to handle the massive data transmissions.

Multinational technology companies like Microsoft and Google used SDN concepts to deploy their SD-WAN solution for inter-data center connections. They have multiple data centers distributed across the planet that are exchanging data traffic extensively for communication between subsystems and data backup activities [5]. Gartner anticipates that by 2024 60% of enterprises will implement SD-WAN [15].

SD-WAN can play a big role in the enhancement of business continuity strategies. Data backup and disaster recovery mechanisms focus on what needs to be done to keep the business running when disasters hit. The long-distance WAN networks play a key role in disaster recovery operations that require a large amount of data to be exchanged over the long distance between primary and secondary sites. They are restricted by limited bandwidth and the high latency of WAN networks. SD-WAN can handle these issues [16] and cope with the increasing network demands to enhance the data backup and disaster recovery operations.

In this paper, an approach for data backup and replication is proposed using the concepts of SD-WAN. The main objective of the proposed approach is to use SD-WAN capabilities to increase the efficiency of WAN network utilization and the adaptation to network demands. This should help to minimize the time needed for data backup between sites by shrinking the recovery point objective down to meet business continuity demands for enterprises.

In the proposed approach, SD-WAN centralized management and programmability are used to run applications that can check the network statistics of data flows like throughput across various links in real-time, and based on these measurements, decisions are taken to meet the demands of other data flows and guarantee the fair utilization of the available WAN links. The OpenNetMon open tool [17] is used in this research to get data flow statistics and to make fine-grained traffic engineering. This will be discussed in the methodology and the results sections.

The remainder of this paper is structured as follows: In Section II, related works to this research are discussed. In Section III, the methodology and algorithm used for balancing the traffic load across all available WAN links are presented. Sections IV and V summarizes the results after applying the proposed approach on the test topology. Finally, Section VI concludes this paper and presents the future works.

II. Related Work

SD-WAN shows great potential in improving the performance and utilization of WAN networks. Currently, many SD-WAN vendors in the market provide different mature products for traffic management and real-time analytics to improve network performance and availability such as Cisco SD-WAN (Viptela), VMware SD-WAN, and Silver Peak. In the past several years, many pieces of research were published showing the benefits of deploying SD-WAN and how it can improve the performance of businesses' workflows and services. Some approaches that aim to improve bandwidth utilization will be reviewed in this section.

Tech giants like Google and Microsoft who have multiple data centers across the planet deployed their SD-WAN solutions for interconnection between data centers. B4 is Google's software-defined inter-data center WAN solution. It uses OpenFlow protocol to manage individual network switches with centralized controllers that run traffic engineering applications. In [18], the authors discuss their experience with deploying B4 in production and achieving better bandwidth utilization for WAN links to perform large-scale data copies between sites.

The second deployment is SWAN from Microsoft that aims to achieve highly efficient and flexible WAN interconnection between data centers. It also improves the fairness to meet different traffic demands when WAN links are utilized with background traffic [19]. It is obvious that both B4 and SWAN, use the capabilities of SD-WAN like the controller's global vision of the network and the easiness to retrieve network statistics for traffic engineering, to maintain better utilization for the WAN expensive resources.

In [20], the authors introduced a software-defined traffic load balancing (SD-TLB) module that is embedded into the ingress nodes at the edge of carrier networks. It is built on the CPU of the physical ASIC-based (Application Specific Integrated Circuit) SDN switches. The SD-TLB performs per-flow scheduling in a round-robin manner over the optical paths besides detecting any path outage or failure using a global network controller based on per-port statistics.

This research showed by experiment that the SD-TLB module is better than the link aggregation group (LAG) for optimizing the bandwidth utilization for optical paths in carrier networks. However, it doesn't abstract completely the control plane and forwarding plane as SD-TLB is still running on the CPU of the edge network switches.

In [21], the authors proposed the implementation of a wide area network defined by software, which guarantees a predefined quality of service (QoS) and traffic prioritization between Software-Defined Datacenters.

The proposal deploys an experiment test scenario based on the concept of the living lab which measures the performance of the network when Voice Over IP (VOIP) services are provided and verified that an adequate level of QoS of Bandwidth can be guaranteed by providing traffic priority in an SD-WAN network between Software-defined data centers (SDDCs).

However, this approach focuses only on maintaining QOS by prioritizing traffic based on the source and destination IPs by a policy/rules introduced by the SD-WAN OpenFlow controller Floodlight (by using OpenFlow flow tables that help OpenFlow switches to steer traffic) and it doesn't introduce any approach that can balance traffic or steer for flows to another transport link in case of having poor bandwidth or high packet loss which can highly impact the quality of VOIP services.

In [22], the authors proposed a simple implementation for SD-WAN using open source tools like OpenDaylight SD-WAN controller and Open vSwitch as virtual switches. They developed applications to run on top of a centralized SD-WAN controller to manage and improve the quality of service (QoS) of the interconnecting WAN network between headquarter and branch offices. The two applications used in this implementation are for monitoring and path switching. The monitoring module will collect real-time statistics for network metrics like packet loss, packet delay, and jitters by injecting traffic into the network paths to evaluate the network links and confirm if they will meet the required QoS. After that, the path switching module will install the path that meets the enterprise QoS thresholds on network switches.

This implementation helped to guarantee the QoS requirements for enterprise applications and services while transferring data between sites. However, the usage of active monitoring and injecting traffic to measure network statistics in real-time may introduce additional network load that impacts the accuracy of measurements.

In this research, fine-grained traffic engineering is enabled in the proposed approach using the real-time measured network metrics for each data flow. The per-flow throughput is measured using an opensource tool OpenNetMon by querying per-flow counters from source and destination switches without injecting extra packets through the network between sites to avoid any overhead that may be introduced. The measured throughput metric will provide an advanced bandwidth utilization monitoring for all available links between sites which will enhance routing decisions for new data flows in a way that can improve the traffic load balancing and the efficiency of utilizing the available bandwidth.

III. Methodology

SDN and SD-WAN are based on the same methodology of separating the control plane from the data plane. Both are using controllers and OpenFlow switches, the main difference is that SDN controllers manage every network device including switches and core switches in LAN networks while SD-WAN controllers manage and interact only with edge devices [23].

In this research, topology in Fig. 1 is suggested to simulate the WAN connection between two sites using different paths (simulating WAN links with different speeds/bandwidth).

Edge devices act as gateways that allow data centers to communicate with other sites through WAN links that can be served by different WAN providers. SD-WAN controllers can't manage or interact with intermediate devices that are under the control of WAN service providers but still can be programmable to interact with sites' edge devices to prioritize and control WAN traffic between different locations.

OpenNetMon software is used in the proposed approach, an open-source software that runs on SD-WAN controller for network monitoring. The OpenNetMon monitoring software will interact with edge OpenFlow switches to monitor TCP flows in real-time and gather statistical info about them. The real-time monitored information will leverage the ability to make fine-grained traffic engineering and hence make more efficient routing decisions. This will enhance data transfer operations through WAN between different sites which is essential for WAN-based backup.

OpenNetMon software is composed of two modules, the forwarding module, and the monitoring module. The forwarding module will discover paths between source-destination pairs when an un-matched TCP flow connection or packet arrives and install path on switches instructing them to route packets for this TCP flow. The monitoring module will monitor flow metrics (throughput/delay/packet loss) between source-destination pairs.

In this proposed approach, the per-flow monitored throughput metric is used for traffic engineering to enhance the bandwidth utilization and balance the traffic load across WAN links between sites while transferring data for WAN-based data backup operations.

Pox controller eel branch is used on Ubuntu Linux 3.13.0-24-generic virtual machine with 8 processor cores and 4GB RAM to run the OpenNetMon software, and mininet version 2.2.1 is used to build testing topology.

The TCP flow throughput metric is periodically measured by querying flow statistics via OpenFlow protocol. The SD-WAN controller will interact with edge Openflow switches to get individual flow statistics on a timely basis to measure the TCP flow throughput in real-time. Equation (1) is used to measure the TCP flow throughput in a given time [24].

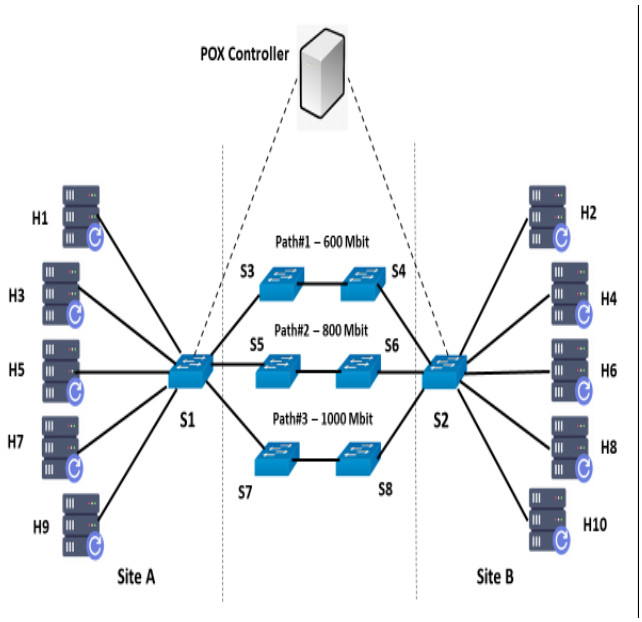


Fig. 1. Test Network Topology.

$$\text{Throughput (tp)} = \frac{\sum_{j=1}^{t} r_j}{t} \quad (1)$$

r_j Number of bytes processed for the TCP flow

t Time

The measured throughput metric of each TCP flow is used to find out the bandwidth utilization of each WAN link and based on this SD-WAN controller will select the least utilized link to be used by the newly initiated TCP flows for transferring data.

By calculating the average throughput of all TCP flows, it can be predicted if the WAN link used by each flow is highly loaded or has room for a new TCP stream. Equation (2) is used to measure the average throughput where tp donates to the measured throughput records for a specific TCP flow and n donates to the number of readings [24].

$$\text{Average throughput} = \frac{\sum_{i=1}^{n} tp_i}{n} \quad (2)$$

tp_i Represents measured TCP flow throughput at a specific interval.

n Represents the number of measured throughput records.

The Flow chart in Fig. 2 explains the proposed approach algorithm to select the least utilized path when new TCP flow starts:

In a nutshell, the following is a step-by-step implementation methodology for the proposed approach:

- Emulate the testing network topology in Fig. 1 on mininet for two sites with edge OpenFlow switches that have three WAN links each one has different Bandwidth (Path#1 600 Mbit – Path#2 800 Mbit – Path#3 1000 Mbit).
- Start running the OpenNetMon software on Pox Controller.
- Initiate three TCP flows at the same time using iperf tool between source-destination pairs H1-H2, H3-H4, and H5-H6 to transfer 10 Gig of Data.
- The three TCP flows will utilize the three WAN Links between the two sites.
- The forwarding module in OpenNetMon software was modified to discover all available paths between source-destination pairs.
- When the OpenNetMon software starts running on the POX controller, all paths between the two sites are in the ideal state. An index variable (i = 0) is used and will increment every time a new TCP flow starts.
- If 'i' is smaller than the number of available paths between source-destination pair, the controller will select one of the ideal paths to the new TCP flow.
- The aim of this is to generate traffic and flow congestion on all available paths.
- The monitoring module will keep a record of throughput for TCP flows on each path and calculate the average throughput for each link between 2 sites.
- After 30 seconds, a new TCP flow was initiated between source-destination pair H7-H8 to transfer 5Gig data using iperf.
- After 60 Seconds new TCP flow was initiated between source-destination pair H9-H10 to transfer 5 Gig of data using iperf.
- In the proposed approach, the forwarding module will select the path with the largest average throughput that will be the least utilized path and will have room to transfer more data.

In the next section, results for the proposed approach are highlighted regarding the throughput and the completion time needed for each TCP flow. Also, a comparison was made between the proposed approach and the Round-Robin approach [25].

Round-Robin: TCP flows are distributed among available paths in a sequential manner. In other words, the chosen path for the new TCP flows for source-destination pairs H7-H8 and H9-H10 is always the path of the next round in the available paths.

Proposed approach: The controller will direct TCP flows for source-destination pairs H7-H8 and H9-H10 to the least utilized path the moment a new flow is initiated based on real-time monitoring.

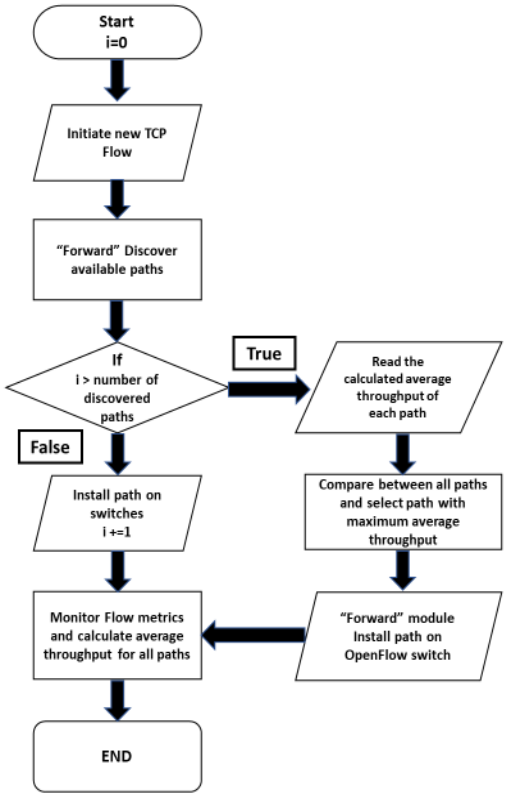


Fig. 2. Flow Chart for the Algorithm used in the Proposed Approach.

IV. Experimental Results

The two approaches were tested on the topology in Fig. 1 while transferring data (simulating data backup operations) between two sites.

The first approach used is Round-Robin. This approach will use all available paths with different Bandwidth (that simulates WAN links) in a sequential manner. Data was transferred between hosts/backup servers from site A to site B using iperf tool and Table I shows the path used by each source-destination pair in the Round-Robin approach.

Fig. 3 shows the throughput of all TCP flows on the 3 paths between source-destination host pairs. As it can be observed from the Round-Robin approach, the TCP flows from source-destination pairs H7-H8 and H9-H10 use path#1 and path#2 to transfer data, however, Path#3 has higher bandwidth, and this can be seen in the high throughput of H5-H6 TCP flow that utilizes this path in the graph (c). The usage of a single path by multiple TCP flows will impact the throughput share of each flow as shown in graphs (a) and (b).

The second approach tested is the proposed approach. Table II shows the path used by each source-destination pair while testing. In Fig. 4, the TCP flow for source-destination pair H7-H8 (that starts after 30 Seconds) uses path#3 based on the average throughput calculation done by the controller for the three paths that are utilized by TCP flows for source-destination pairs H1-H2, H3-H4, and H5-H6.

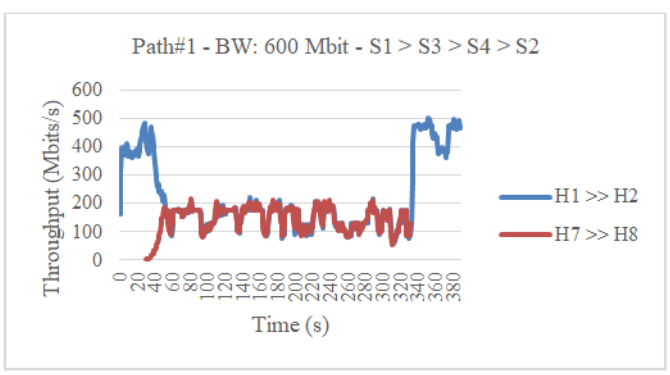
After 60 Seconds when a new TCP flow was initiated for H9-H10, path#2 was selected by the controller for it. The average throughput for path#2 shows better real-time results the moment this flow was started compared to path#3 that was already utilized by 2 TCP flows for H5-H6 and H7-H8 and path#1 that has the least bandwidth.

TABLE I. Paths used by each Flow in the Round-Robin Approach

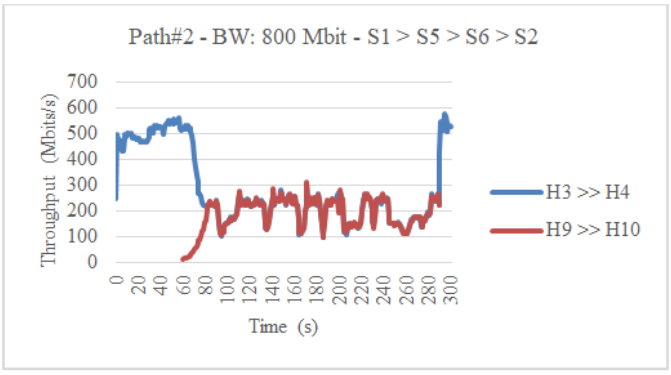
Src-Dst pair	Data Transferred	Path	Path Bandwidth
H1-H2	10 Gig	Path#1 - S1 > S3 > S4 > S2	600 Mbit
H3-H4	10 Gig	Path#2 - S1 > S5 > S6 > S2	800 Mbit
H5-H6	10 Gig	Path#3 - S1 > S7 > S8 > S2	1000 Mbit
H7-H8	5 Gig	Path#1 - S1 > S3 > S4 > S2	600 Mbit
H9-H10	5 Gig	Path#2 - S1 > S5 > S6 > S2	800 Mbit

TABLE II. Paths used by each Flow in the Proposed Approach

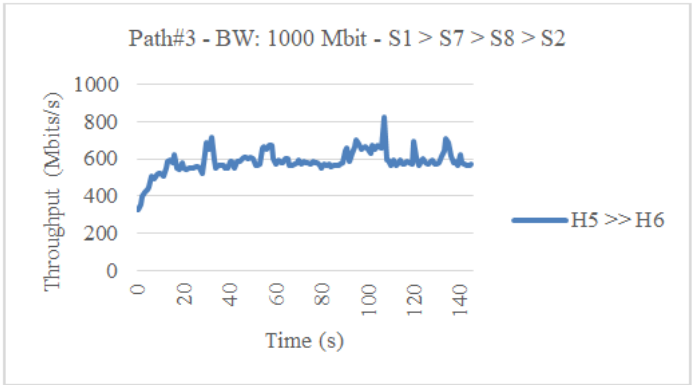
Src-Dst pair	Data Transferred	Path	Path Bandwidth
H1-H2	10 Gig	Path#1 - S1 > S3 > S4 > S2	600 Mbit
H3-H4	10 Gig	Path#2 - S1 > S5 > S6 > S2	800 Mbit
H5-H6	10 Gig	Path#3 - S1 > S7 > S8 > S2	1000 Mbit
H7-H8	5 Gig	Path#3 - S1 > S7 > S8 > S2	1000 Mbit
H9-H10	5 Gig	Path#2 - S1 > S5 > S6 > S2	800 Mbit



(a)



(b)



(c)

Fig. 3. (a) Network throughput for H1-H2 and H7-H8 TCP Flows on Path#1 using the Round-Robin Approach, (b) Network throughput for H3-H4 and H9-H10 TCP Flows on Path#2 using the Round-Robin Approach, and (c) Network throughput for H5-H6 TCP flow on Path#3 using the Round-Robin Approach.

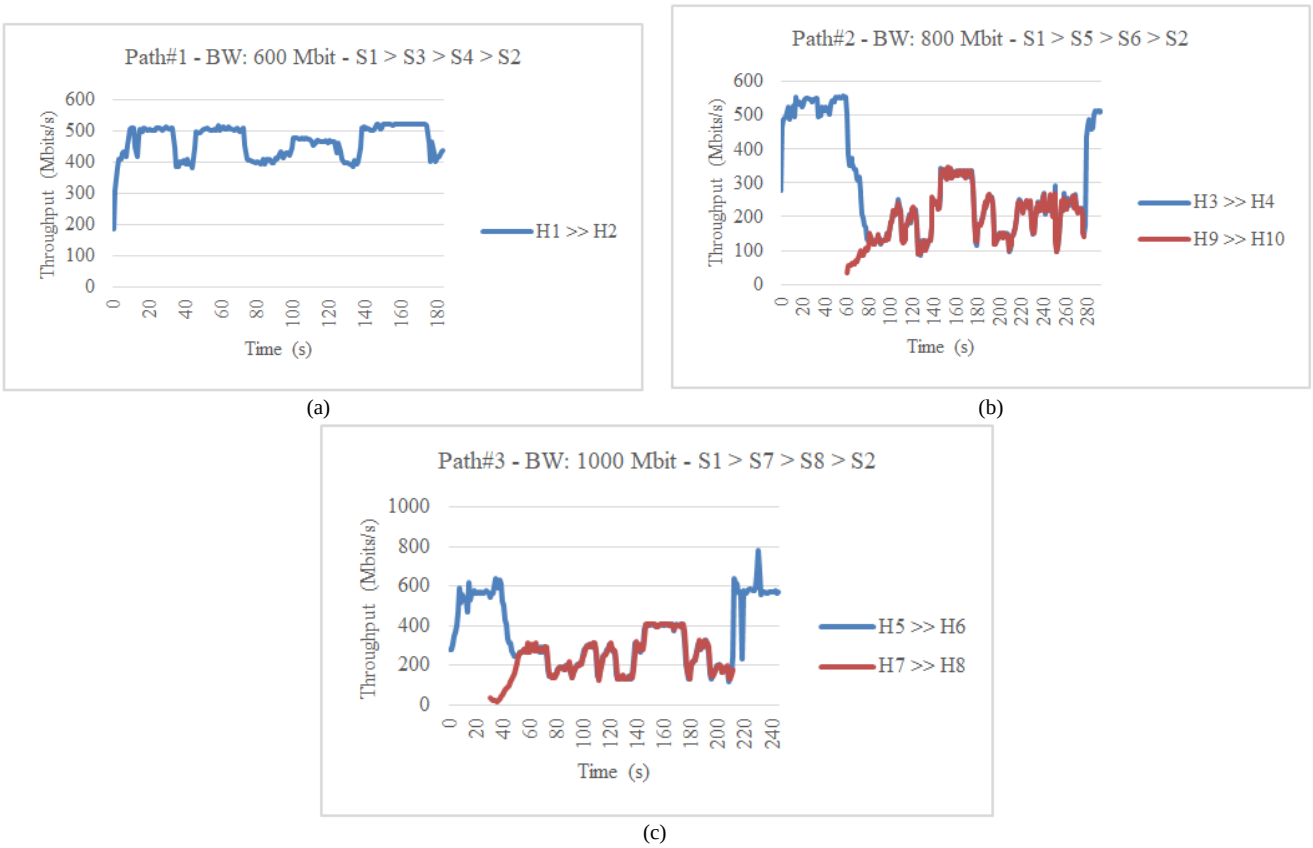


Fig. 4. (a) Network throughput for H1-H2 TCP Flow on Path#1 using the Proposed Approach, (b) Network throughput for H3-H4 and H9-H10 TCP Flows on Path#2 using the Proposed Approach, and (c) Network throughput for H5-H6 and H7-H8 TCP Flows on Path#3 using the Proposed Approach.

V. Results and Discussion

As it can be observed in the comparison between the 2 approaches in Fig. 5, for the average throughput and time taken by each TCP flow for completing data transfer, the proposed approach shows that a fair share and better utilization of the bandwidth resources can be achieved in challenging network conditions beside reducing the time needed to transfer backup data between sites.

In Fig. 5(a) it is clear that the total time required for the 5 data flows that are initiated between the 5 source-destination pairs to transfer 40Gig of data, has been reduced significantly using the proposed approach compared to Round-Robin. The total time required for transferring 40Gig using the Round-Robin approach is 389.3 seconds across the 3 available links used to connect between the two sites A and B. While in the proposed approach, using traffic engineering based on SD-WAN capabilities to balance the traffic load across network links, the time required to complete data transfer is reduced to 293 seconds.

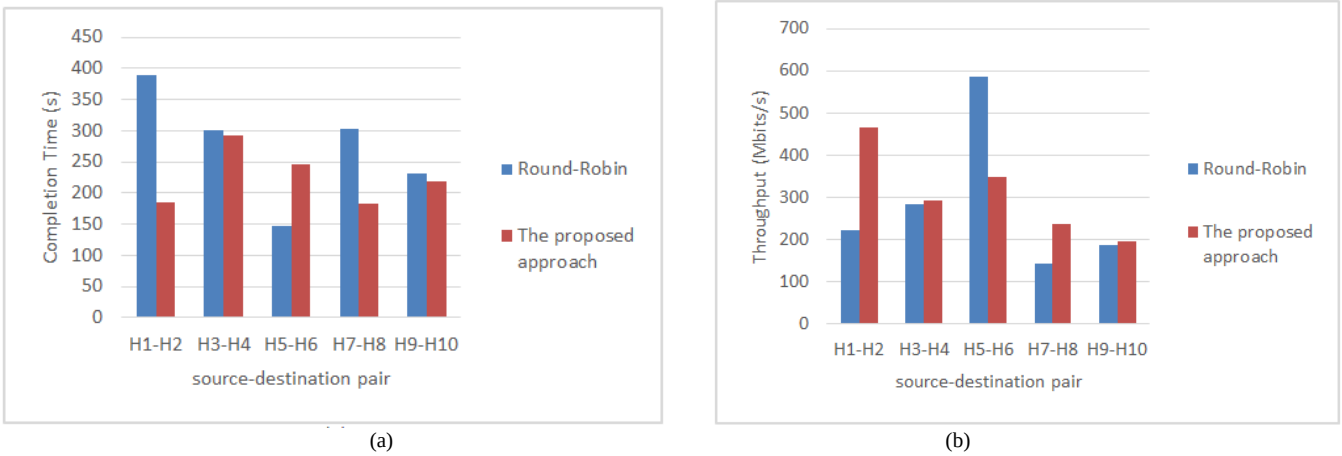


Fig. 5. Comparison between the Two Approaches for (a) Completion Time and (b) Average throughput of each TCP Flow.

These numbers reflect the bandwidth utilization enhancement between the data flows as shown in Fig. 5(b). The traffic load balancing mechanism introduced in this research, based on real-time measured network throughput, helps to improve the fairness in bandwidth utilization between the different data flows. The focus on background traffic initiated by the data flows (H1-H2, H3-H4, H5-H6) that initially utilized the three available network links, helps to make better routing decisions for newly initiated flows (H7-H8 and H9-H10). The average throughput for H7-H8 data flows while using the proposed approach is 236 Mbit/sec compared to round-robin 142 Mbit/sec. This enhancement occurs because of a better routing decision to use network Path#3 (bandwidth 1000 Mbit) and to share it with data flow (H5-H6) instead of sharing Path#1 (that has the least bandwidth 600 Mbit) with data flow (H1-H2). On the other hand, data flow (H1-H2) average throughput has improved to be 466 Mbit/sec, instead of 221 Mbit/sec in the round-robin case, after utilizing the limited bandwidth resources of Path#1 (bandwidth 600 Mbit) without having another competing data flow that tries to share resources on this network link.

It is obvious that the average throughput of H5-H6 data flow (transferring 10Gig of data) has been reduced from 587 Mbit/sec to 349 Mbit/sec after using the proposed approach for balancing data. This reduction is acceptable in the interest of improving fairness between all data flows and meeting the traffic bandwidth demands. This flow was using individually Path#3 (bandwidth 1000 Mbit) to transfer 10Gig of data in the round-robin approach, leaving the other four data flows to share the remaining two available network links Path#1 (bandwidth 600 Mbit) and Path#2 (bandwidth 800 Mbit) for transferring 30 Gig of data. After applying the proposed approach to balance traffic load, Path#3 is shared between two data flows (H5-H6 and H7-H8) to transfer 15 Gig of data.

The average throughput results for data flows H9-H10 and H3-H4 are nearly similar in the two approaches. The reason is that both data flows have shared the same network link Path#2 (bandwidth 800 Mbit) when applying the round-robin approach and the proposed approach as shown in Table I and Table II.

The results show that the aim of this research has been achieved by deploying fine-grained traffic engineering using SD-WAN to balance the traffic load across all available WAN links between data centers in a way that improves the bandwidth utilization and minimizes the time required for transferring data. The centralized traffic engineering used in the proposed approach helps to overcome the poor efficiency and sharing capabilities of WAN. This was achieved by considering the current state of the available WAN links before initiating new flows. The average throughput measurements are used to identify the reliability of each WAN link to handle new TCP flow without impacting the current running flows. From the test results, it is obvious that Round-Robin can cause an unplanned over subscription for some WAN links which degrades the performance of the competing TCP flows using them while other links are underutilized. On the other hand, the proposed approach helps to overcome the under and over subscription swinging state for available WAN links by enhancing the routing decisions for new data flows in a way that improves the fairness in sharing WAN resources and the utilization of bandwidth.

These enhancements are blessings for data backup and replication operations that are not a basic one-time function and they need to overcome many network limitations like latency for transferring data over long-distance networks. By minimizing the time required to transfer data through the network to the disaster recovery sites, the efficiency of data protection and business continuity will improve, and enterprises will be able to minimize their downtime.

VI. Conclusion and Future Work

In this work, the Software-Defined Network paradigm has been applied to achieve traffic load balancing across the hybrid WAN links available between data centers. The centralized programmable interface introduced by SD-WAN leverages the ability to make advanced network measurements for data flows that are propagating across network WAN links. An open-source network monitoring tool OpenNetMon is used to measure network throughput metrics for data flows based on SD-WAN capabilities. The measured network throughput metrics were used to balance the traffic load across available network links and to improve the efficiency of bandwidth utilization while transferring data between data centers.

The aim of this proposed approach is not to confirm or confute the performance improvement that can be achieved by other approaches that are used for traffic load balancing or to improve the bandwidth utilization, but the primary goal instead is to develop a new approach based on an open-source tool that was introduced in previous research to monitor per-flow metrics in real-time and provide the network statistical information essential for traffic engineering purposes.

By testing the proposed approach on a network topology simulated using mininet and compared it to the round-robin approach, the results showed that the approach developed in this research has shown an improvement in utilizing the bandwidth resources and helped to minimize the time required for transferring data between two sites. This will improve the efficiency of business continuity applications and helps the data backup operations to overcome network latency limitations that impact the time required to backup data between production and disaster recovery sites.

In the future work, it is planned to develop the used algorithm to add other network metrics like packet loss and latency to measure the reliability of available paths and to ensure meeting adequate quality of service policies before making routing decisions for new data flows. It is intended to extend the proposed approach by integrating algorithms to apply fault tolerance. The aim of fault tolerance is to failover data flows to use another WAN link in case of link failure or severe degradation in the link performance and throughput.

References

[1] B. Shin, *A Practical Introduction to Enterprise Network and Security Management*. Auerbach Publications, 2021.

[2] R. Graziani and B. Vachon, *Cisco Networking Academy: Connecting Networks Companion Guide*. Cisco Press, 2014.

[3] X. Tao, K. Ota, M. Dong, W. Borjigin, H. Qi, and K. Li, 'Congestion-aware Traffic Allocation for Geo-distributed Data Centers', *IEEE Trans. Cloud Comput.*, 2020.

[4] M. Said Elsayed, N.-A. Le-Khac, and A. D. Jurcut, 'Dealing With COVID-19 Network Traffic Spikes [Cybercrime and Forensics]', *IEEE Secur. Priv.*, vol. 19, no. 1, pp. 90–94, Jan. 2021, doi: 10.1109/MSEC.2020.3037448.

[5] O. Michel and E. Keller, 'SDN in Wide-Area Networks: A Survey', p. 6, 2017.

[6] G. Pujolle, *Software Networks: Virtualization, SDN, 5G, and Security*. John Wiley & Sons, 2020.

[7] A. Eftimie and E. Borcoci, 'SDN controller implementation using OpenDaylight: experiments', in *2020 13th International Conference on Communications (COMM)*, 2020, pp. 477–481.

[8] H. M. Noman and M. N. Jasim, 'POX controller and open flow performance evaluation in software defined networks (SDN) using mininet emulator', in *IOP conference series: materials science and engineering*, 2020, vol. 881, no. 1, p. 012102.

[9] V. S. Raju, 'SDN controllers comparison', 2018.

[10] M. Paliwal, D. Shrimankar, and O. Tembhurne, 'Controllers in SDN: A Review Report', *IEEE Access*, vol. 6, pp. 36256–36270, 2018, doi: 10.1109/ACCESS.2018.2846236.

[11] D. J. Hamad, K. G. Yalda, and I. T. Okumus, 'Getting traffic statistics from network devices in an SDN environment using OpenFlow', *Inf. Technol. Syst.*, vol. 2015, pp. 951–956, 2015.

[12] F. A. Lopes, M. Santos, R. Fidalgo, and S. Fernandes, 'A software engineering perspective on SDN programmability', *IEEE Commun. Surv. Tutor.*, vol. 18, no. 2, pp. 1255–1272, 2015.

[13] C. Trois, M. D. Del Fabro, L. C. de Bona, and M. Martinello, 'A survey on SDN programming languages: Toward a taxonomy', *IEEE Commun. Surv. Tutor.*, vol. 18, no. 4, pp. 2687–2712, 2016.

[14] J. Zhao, Z. Hu, B. Xiong, and M. Zheng, 'Performance Modelling and Optimization of Controller Cluster Deployments in Software-Defined WAN', in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Zhangjiajie, China, Aug. 2019, pp. 106–113. doi: 10.1109/HPCC/SmartCity/DSS.2019.00030.

[15] 'Gartner Dell - Transforming Networking for the Cloud Era with SD-WAN'. Dell Technologies, 2020. [Online]. Available: https://www.delltechnologies.com/asset/en-eg/products/networking/briefs-summaries/gartner-dell-newsletter-transforming-networking-for-the-cloud-era-with-sd-wan.pdf

[16] P. Kokkinos, D. Kalogeras, A. Levin, and E. Varvarigos, 'Survey: Live Migration and Disaster Recovery over Long-Distance Networks', *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–36, Nov. 2016, doi: 10.1145/2940295.

[17] N. L. M. van Adrichem, C. Doerr, and F. A. Kuipers, 'OpenNetMon: Network monitoring in OpenFlow Software-Defined Networks', in *2014 IEEE Network Operations and Management Symposium (NOMS)*, Krakow, Poland, May 2014, pp. 1–8. doi: 10.1109/NOMS.2014.6838228.

[18] S. Jain *et al.*, 'B4: Experience with a globally-deployed software defined WAN', *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 3–14, 2013.

[19] C.-Y. Hong *et al.*, 'Achieving high utilization with software-driven WAN', in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, 2013, pp. 15–26.

[20] Y.-J. Kim, J. E. Simsarian, and M. Thottan, 'Software-defined traffic load balancing for cost-effective data center interconnection service', in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, Lisbon, Portugal, May 2017, pp. 255–262. doi: 10.23919/INM.2017.7987287.

[21] R. E. Mora-Huiracocha, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, E. J. Cedillo-Elias, and V. M. Larios-Rosillo, 'Implementation of a SD-WAN for the interconnection of two software defined data centers', in *2019 IEEE Colombian Conference on Communications and Computing (COLCOM)*, Barranquilla, Colombia, Jun. 2019, pp. 1–6. doi: 10.1109/ColComCon.2019.8809153.

[22] S. Troia, L. M. M. Zorello, A. J. Maralit, and G. Maier, 'SD-WAN: An Open-Source Implementation for Enterprise Networking Services', in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, Bari, Italy, Jul. 2020, pp. 1–4. doi: 10.1109/ICTON51198.2020.9203058.

[23] S. Troia, L. M. M. Zorello, and G. Maier, 'SD-WAN: how the control of the network can be shifted from core to edge', in *2021 International Conference on Optical Network Design and Modeling (ONDM)*, 2021, pp. 1–3.

[24] V. Malagar, M. Kumar, S. M. V. Devi, and S. Gupta, 'Performance Evaluation of VANETs for Evaluating Node Stability in Dynamic Scenarios', *Int J Comput Appl Technol Res*, vol. 7, no. 9, pp. 376–385, 2018.

[25] A. Al-Najjar, S. Layeghy, and M. Portmann, 'Pushing SDN to the end-host, network load balancing using OpenFlow', in *2016 IEEE international conference on pervasive computing and communication workshops (percom workshops)*, 2016, pp. 1–6.

Critical Data Consolidation in MDM to Develop the Unified Version of Truth

Ms. Dupinder Kaur*, Dr. Dilbag Singh
Department of Computer Science and Engineering
Chaudhary Devi Lal University, Sirsa, India

***Abstract*—Organization seeking growth and competitive lead should use Master Data Management (MDM) as a foundation for efficient decision making. An MDM framework creates a trusted and reliable continuous record of customers, products, suppliers and other shared data sets. In master data, the critical data is consolidated to portray essential business entities into a Unified version of Truth. To create trusted view of master data challenges like quality, identity resolution, analytics and investment are faced. In proposed research, a technique has been designed to generate Master Data to assist the policy maker to address the said issues. In this paper, four steps have been taken for master data creation namely: Data Enrichment, Data Matching, Data Merging and Data Governance. To achieve legitimate data quality TALEND open studio has been used for data pre-processing and enrichment. An algorithm is designed to match and merge the master records. To validate the designed approach, results are evaluated using Pandas Data Frame on Python platform. This paper will assist the policy makers of the organizations in formulating the business strategies.**

***Keywords*—Master data management (MDM); master record; TALEND; data matching and merging**

I. INTRODUCTION

Data Management is concerned, with the entire lifecycle of a data resource from the creation to retirement, with advances and changes. Data Management comprises the procedures, practices, ideas and measures for proficient utilization of data resources. An enlargement of the data size, in many folds, makes the organizations to adopt the data management practices to maintain the quality. The transformation, integration and cleaning of data is needed for efficient handling of data [1]. The new challenge for maintaining the quality of most essential assets of business demands Master Data Management (MDM) program. MDM comprises of technologies, processes and disciplines to incorporate the cleaning, administration and controlling all shared data assets. An MDM arrangement formulates a single accurate set of data populated across different frameworks by ensuring the consistency and accuracy of organization's shared data assets. [2]. The core benefits of MDM includes informed decision making, reducing data duplication, better data compliance and handling change requests.

Master data deals with the critical data entities of a company such as clients, items and resources that are shared across value-based applications. It additionally enables the development of a 'single version of truth' [3]. The technical operations assisted by MDM activities include Data quality improvement, Master Data creation and Data engineering [4]. MDM is significant for a wide range of applications to make a complete beginning to end plan that drives progression and achieves better business results. A viable MDM execution enables better usage of basic data present in organization [5]. Relevant to Master data Management, various data taxonomies used in an enterprise are:

- Transactional Data: The inner or outer exchange or transaction that happen including sales, orders, purchase orders, card payment, etc.
- Reference Data: It addresses set of qualities that are referred by frameworks, applications, information store just as by transaction and master data, such as status codes, state contractions, segment fields and so forth [6].
- Reporting Data: The aggregated data compile for the purpose of analytic and reporting. For example: order status (Accept or Reject).
- Meta Data: The data that describe label or characterize other data. For example: properties of media file: its size, type, resolution and author, etc.
- Master Data: The persistent, non transactional data that defines primary business entities such as customer, product, employees and inventory. Master data is the unified version of common data that are often duplicated across the enterprise. The figure below depicts various categories of Master Data [7].

Master Data of an organization can include all entities involved in financial structure, transactions done by organization or locations such as address. If master data quality does not meet expectations, it can affect the efficiency of business operations [9]. Thus, there is need to manage this data. Master Data Management ensures complete, consistent, and accurate data in different areas of organization's activities MDM includes process of data collection, cleansing, consolidation and distribution in the organization ensuring the control of use in various analytical and operational applications [10]. The underlying driver for execution of MDM is to deal with data quality issue frequently emerges across data entities and information systems [13]. MDM is implemented in most of organization with aim to guarantee that master data includes reference details that reflects the present state of business [11].

*Corresponding Author.

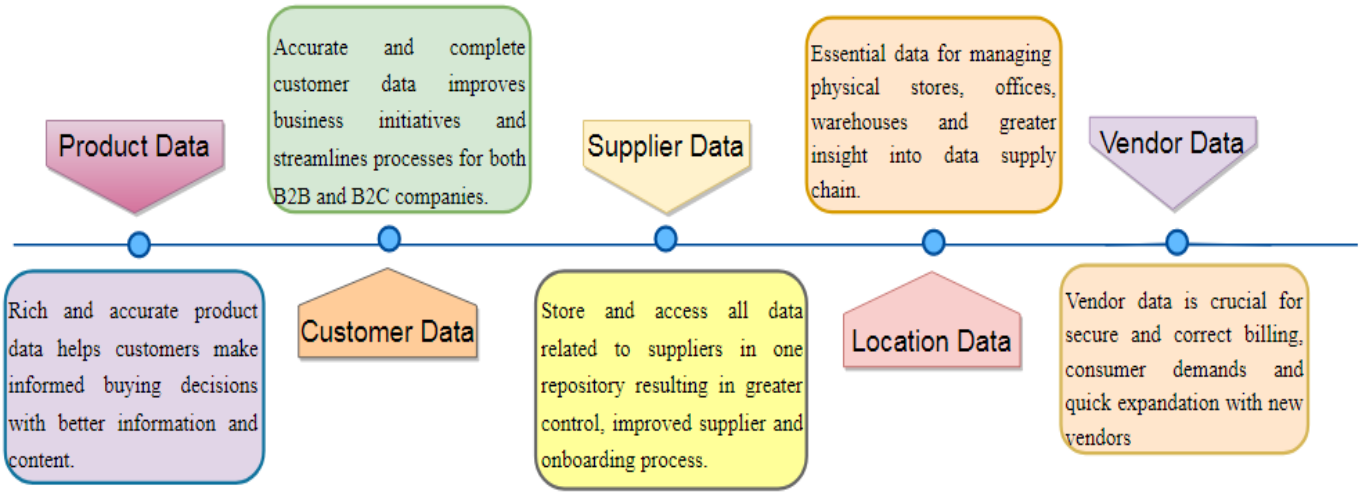


Fig. 1. Categories of Master Data [9].

Fig. 1 illustrates various categories of Master data. On the basis of need of an enterprise, Master data can be an employee data, party data, asset data or ledger data [8].

A Master Data Management program may provide an option to the industry with new ways to handle the data quality issues that the industry has struggled for few years and help prevent the "information rich and data poor" dilemma [14].

The main activities required for implementing MDM in an organization consist of following activities:

- Discover the need of master data management in organization and reference data.
- Categorize the sources and contributors of data.
- Define and maintain data architecture [12].

In light of the examination and investigation, following are the major contributions of this paper.

- To study the management, categories and related terms of Master Data.
- Designing of an algorithm for generating Master Record.
- Estimation and fixing of Data Quality using TALEND MDM tool.
- Use of Pandas Data Frame for creating Master Record.

The remaining part of the paper is organized as follows:

In Section 2, the related literature is reviewed. Section 3 elaborates the research methodology used for present research work. Section 4 is about the design of the proposed algorithm for generating Master records. Section 5 presents the Input dataset. Section 6 shows implementation and results of the proposed algorithm. Last section discusses the findings of implementation of the proposed method for creating qualitative and accurate master data to assist the organizations in decision making.

II. Literature Review

Radachirkova [1] et al. analyzed the usability of the existing data transformation tools for their utilization to achieve the desired quality characteristics in business measures. The focus was to enhance the data quality using existing tools instead of adopting a novel data transformation procedure for every data analytics task. Tending to this issue; Data cleaning, migration and transformation tools were summarized as black box procedure by exposing some properties such as applicability requirements, portion of data modified and constraints satisfied by data over applied procedure. The formal study revealed that these primary outcomes could be applied for accomplishing desired data quality outcomes in data analytics.

PanagiotisLepeniotis [2] integrated the MDM platform with Business Transformation Programme for decision making. The examination revealed that the Management of the Master Data and the never-ending confirmation of the Data Quality are crucial for any organization despite of having a BTP or not. An MDM impacted BTP decision model has been presented in the research. On the basis of case study audits and interviews, the research identified an improved indulgent into decision making process of a BTP concerning MDM and the way, these decisions affect the fruitful execution of a BTP.

Fernando Gualo [3] et al. assessed the "Functional Suitability" of MDM applications by considering useful necessities from section 100 to 140 of ISO 8000 and proposed a solution by considering the appraisal and affirmation of Functional Suitability of MDM applications. In addition to basic requirements, test cases are also designed required for the evaluation. In order to infer the prerequisites from ISO 8000, all the parts are covered except of ISO 8000-115, ISO 8000-116 and ISO 8000-150. Application of assessment method for existing master data based application is also featured.

Shreya Mrigen [4] et al. proposed a method for singular data management in pharmaceutical company. Hierarchies and type of problem faced in managing MDM solution in

pharmaceutical company has been examined in this research. The examination revealed that MDM is an essential activity for creating singularity, improving data consistency and data processing and market examination in pharmaceutical companies.

Dilbag Singh [5] et al. provided an MDM solution for building frameworks. A strategy to pilot the implementation process of MDM has been provided with a Framework, Roadmap and a DFD design. This strategy described complete description of step wise approach to have a better view point on generating, handling, validating and monitoring master data. A study on existing work and Gartner's Hype cycle has been performed to know the latest technical trends in MDM.

Chun Zhao [6] et al. designed a model for assessing the viability and rationality of master data network. Set Pair Analysis (SPA) design of MDM has been presented in the research. Data network based on master data and data keys are established using MDM system. In association with master data, the main concern was to assess the effectiveness and rationality of the network. Contextual analysis showed convenient update of information, distribution and active response are significant elements in cloud fabricating climate.

On the basis of literature review, it can be concluded that designing of policies and standards are required to evaluate the functional suitability of MDM based applications. Further, the identity resolution, business rules and MDM solution is required for data consistency and security. Data governance and stewardship is of utmost importance for managing the data.

III. Research Methodology

Identity resolution technique is significant in MDM as it integrates two or more data identities into one object. To allow data citizens to access the right information, a productive identity resolution technique is required. In MDM, Identity resolution finally determines the master record. To realize this need, a methodology has been proposed in the present research for identity resolution in MDM. Research begins with exploring, analyzing and enrichment of input data set on TALEND tool for legitimate quality of data. The characteristics of relevant attributes of input data sets like matching score, threshold value and merging condition are considered. Finally, the experimental study has been performed to generate Master Table using Pandas data frame. Hence, an exploratory, descriptive and experimental research methodologies are used in this research. This concrete motivation for generating an MDM solution arises from everyday needs of global identification, linking and synchronization of master data across heterogeneous data sources.

IV. Proposed Algorithm for Consolidating Critical Data Entities

In the present research; data is collected from various sources in different formats, therefore, the consolidation has been carried out. Consolidation amalgamates all the data, remove redundancies and inaccuracy before combining it into single place. Critical data consolidation enables 360-degree-view of data assets, efficient plan, implementation and execution of a business process. Data consolidation in MDM initiates with data collection from the significant sources, utilizing business rules to build up a unique data source, data governance and transmission to the concerned departments. To consolidate critical data entities over collected data set, an algorithm has been proposed in this research as shown in Fig. 2.

Input: Data Set collection from relevant sources.
Output: Master Table - Unified version of Truth.
Steps:

1. Collect data from multiple sources
2. Build a standardized data model for integration as per data set requirements.
3. Evaluates data quality with predefined standards and polices.
4. If data quality is not legitimate as per defined rules, do
5. Apply data cleaning and standardization techniques
6. End if
7. Design a model for resultant Master Table
8. Select the desired attributes to be matched from input data files.
9. For each record in input do
10. Apply fuzzy matching technique on corresponding attributes of input files
11. Set m = similarity score of attributes
12. Set k= threshold value
13. If m>=k
14. Set the record for best match and automatic merge.
15. End if
16. If m>=0 && m<k
17. Set the record for manual merge using data stewardship
18. Else
19. If m<0
20. Set the record as worst match and not pertinent for merge.
21. End if
22. End if
23. End for
24. Generate Master Table as T
25. Perform automatic and manual merge of records into T
26. Select a record as Instance I
27. If I is new record
28. Add it to T by following step 3.
29. Else
30. Repeat step 8
31. Apply Data Governance technique on T
32. Deploy T as final master table: Unified version of truth

Fig. 2. Algorithm for Generating Master Data through Consolidation.

The above algorithm is proposed for creating master data. Initially, data is gathered from heterogeneous sources. On the basis of the data set requirements, Data Model-1 is created. The quality of consolidated data is verified by keeping into account the policies and rules of the organization concerned. In case the data is adequate, it will be processed further, failing which quality improvements techniques like Data Cleaning, Enrichment or scrubbing are applied to ensure the data consistency, accuracy and timeliness. Once the data is

enriched, the attributes to be matched are selected from the input files. For each record, a similarity score (m) is generated. The corresponding attributes are matched and a score is generated by using fuzzy matcher. A threshold value (k) is selected on the basis of matching score for categorizing the matching and merging process. The attributes of the input files are compared to find out the match. For each record, a matching score is calculated by applying fuzzy matcher.

On the basis of matching score, a threshold value is selected. Matching score of a record, greater than or equal to threshold value, will lead to the best match and the data is automatically merged. However, if matching score is greater than or equals to zero or less than threshold, it is considered as average match and processed for manual merge using Data stewardship. For worst match, matching score is negative and thus the data is not considered for a merge. In the proposed strategy, master table is generated by applying automatic and manual merge. In case there is a new record, it will be directly added to Master table and if a record already exists in the table then the entire process of matching and merging is again carried out. Data Governance polices are applied continuously on master data to ensure the quality of the data.

The resultant Master record should correctly match and merge using identity Resolution technique. Variation in data will adversely affect the search process and quality of data. For example: A Person may use Ritesh as a name at one place but Ritesh Kumar or Retish K. at another places. The variation in names may be due to use of nick names sometimes, aliases or initials. Change in address of roads, areas, billings, mailing etc may lead to variations. Such variations are overcome while merging the data using proposed algorithm. Table I describes range of variations for designed Identity resolution.

Table I explains the range of Variation for identity Resolution using designed algorithm. Identity matching allows correct result focusing on standard and quality of key data elements such as address, phone and email address. It allows enrichment of customer profiles to improve the accuracy of matching.

TABLE I. RANGE OF VARIATION FOR DESIGNED IDENTITY RESOLUTION TECHNIQUE

Type of Data	Potential Variants on a Data Type	Priority
Name Variation	Ashwani, Ashwini, Ash, Aswani	High
Abbreviation	Mohammad, Muhammad, Mohd, Mhd	High
Phonetics /Spelling Variation	Ritesh, Reetesh, Ritish	High
Date Format	9/01/2020, 09/01/2020, Jan 09, 2020, 09 Jan 2020, 01/09/2020	High
Suffix Variation	Ranjeet Singh, Ranjeet S. , Ranjeet	Medium
Null Values	Not known, Unknown, ?, 000, N/A, [Blank],NaN	Medium
Organization Name	Chaudhary Devi lal University, CDLU, Chaudhary devilal University	Medium
Department Name	Department of Computer Science & Engineering, DCSE, Deptt. of Compt. Sc. &Engg.	Medium
Titles	Mr. Pawan, Dr. Pawan, Pawan MD	Low

V. INPUT DATASETS

To implement the proposed work, data set has been collected from University Computer Science & Engineering department, Library section and Account section. The information of student is submitted to department during admission of a course while a separate form is filled by student for issuing the books in Library. As for master data, two or more sources of same information is required The figure below represents the screenshot of input data files.

	D_Reg_No	D_Name	D_Father_Name	D_Email_Id	Subject	Category	Fees	D_Contact
0	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	MCA-32	GENERAL	8230 Rs/-	96******0229
1	130034	AKSHAY BANSAL	MR.KULBHUSHAN BANSAL	akbansal77@gmail.com	MCA-32	NaN	8230	99******0230
2	130035	PARDEEP KUAMR	JOGENDER SINGH	kumarpar258@gmail.com	MCA-32	SC	SC	99******0231
3	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	MCA-32	SC	750/-	92******0232
4	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	NaN	NaN	750/-	93******0233
5	130037	kawal Preet	Balraj Singh	singhkawal97@gmail.com	MCA-32	SC	750/-	99******0234

Fig. 3. Dataset1- Input_Flow_1 File.

Fig. 3 shows the input dataset1 as Input_flow_1 file containing the data collected from department. The attributes D_Reg_no, D_Name, D_Father_Name represents the Registration No, Name and Father Name of student as recorded in department and so on.

	L_Reg_No	L_Name	L_Father_Name	L_Email_id	L_Contact
0	130004	Amritpal	Tehal Singh	dhaliwal19362@gmail.com	79******0202
1	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	96******0229
2	130031	SUMANDEEP KAUR	LAKHWINDER S.	KAURDA061@GMAIL.COM	75******0228
3	130037	Kawal Preet	BalraJ SINGH	singhkawal97@gmail.com	99******0234
4	130016	Monika Verma	Stapal Singh Verma	monuverma1577@gmail.com	99******0213
5	130030	Vikas	Jitender	vikurampura@gmail.com	77******0227

Fig. 4. Dataset1- Input_Flow_2 File.

Fig. 4 describes second dataset as Input_File_2 taken from library. The attributes L_Reg_no, L_Name, L_Father_Name represents the Registration No, Name and Father Name of student as recorded in Library and so on.

VI. IMPLEMENTATION AND RESULTS

The schema diagram designed for present work comprises of four steps: Data Enrichment, Data Matching, Data Merging and Data Governance as shown below.

Fig. 5 illustrates the schema design for data consolidation process in MDM. A sequential flow of execution from Data enrichment to Data Governance has been employed in this research.

For the implementation, the quality of input data is verified first. As Fig. 3 addresses the problem of Data redundancy (Duplicate records for Registration No 130036), domain integrity constraint violation ('SC' value in FEES attribute) and incompleteness (Null entries in Subject and Category attribute). Before creating Master data, all these issues must be resolved to make data clean. Thus, in present research, data enrichment is performed on TALEND tool. The following sections describe the modules Data Enrichment, Data Matching, Data Merging and Data Governance in detail.

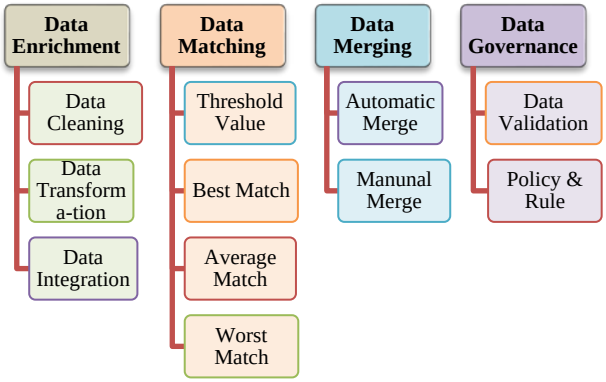


Fig. 5. Schmea for Methodology used in Research.

A. Module 1: Data Enrichment

Data Enrichment is characterized as enhancing or appending the collected data by supplementing incomplete, missing or inaccurate data. The subsequent enriched data empowers organizations in effectively customizing the data. To implement Data Enrichment, TALEND open studio for MDM has been used in present research as shown below.

Fig. 6 illustrates TALEND window with two input data files as Dept_data and Lib_data taken from University department and library. The results of TALEND tool are shown through Python platform for bringing together the results with further implemented result.

1) Data cleaning: Data cleaning is the process of removing or fixing inaccurate, corrupt, outdated, and incorrect formatted, duplicate entries from data sets. Data Cleaning is performed on input files Dept_data and Lib_data using predefined components in TALEND. The following figure represents the screenshot of Clean Data in Output_Flow_1 file.

Fig. 7 signifies that Unique Row and Replacement components are used for removing redundancy; null values, inaccurate and domain integrity constrain violation on input data source. The contents of Output_Flow _1 file shows that the above said problems are resolved in this step. Similarly, the same procedure is applied on library data and clan data is stored in Output_Flow_2 file.

2) Data transformation: The way toward changing over data from one format or construction into another is known as Data Transformation. In the present research, Python language has been used to execute the MDM solution. Python, being case sensitive language; lower case sentences are required to be converted into upper case sentences. Hence, data transformation is performed under mapping section of TALEND tool.

Fig. 8 indicates the Mapping process performing both Data Transformation and Integration. A mapping component takes two input as Main (department) and Lookup (Library) as shown above. The records are mapped on the basis of registration number of student. The attributes: Student Name, Father Name and Email id are converted into upper case under this section.

3) Data integration: Data Integration refers to the process of consolidating data from multiple sources into single or unified view. This technique helps analytic tools to generate effective and actionable business intelligence.

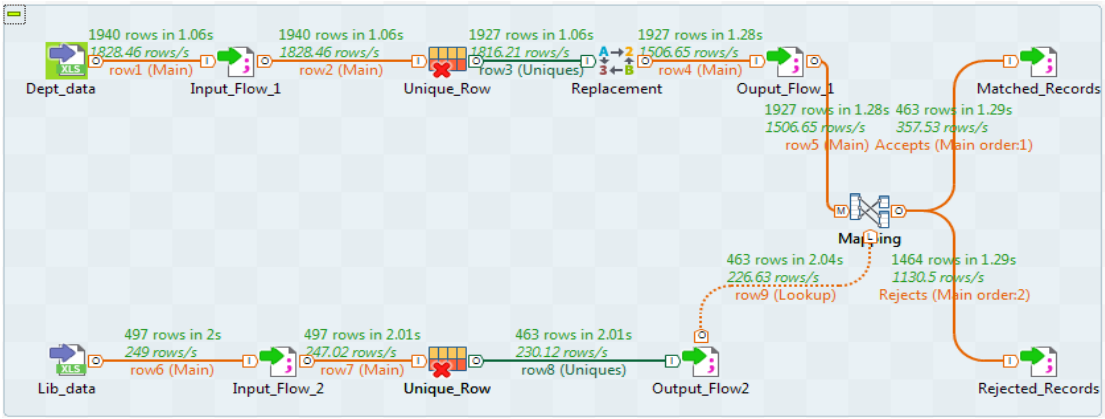


Fig. 6. Data Enrichment using TALEND.

	D_Reg_No	D_Name	D_Father_Name	D_Email_Id	Subject	Category	Fees	D_Contact
0	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	MCA-32	GEN	8230.0	96******0229
1	130034	AKSHAY BANSAL	MR.KULBHUSHAN BANSAL	akbansal77@gmail.com	MCA-32	GEN	8230.0	99******0230
2	130035	PARDEEP KUAMR	JOGENDER SINGH	kumarpar258@gmail.com	MCA-32	SC	750.0	99******0231
3	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	MCA-32	SC	750.0	92******0232
4	130037	kawal Preet	Balraj Singh	singhkawal97@gmail.com	MCA-32	SC	750.0	99******0234
5	130038	SUNIL KUMAR	KRISHAN KUMAR	kumarsunil123@gmail.com	MCA-32	SC	750.0	89******0235

Fig. 7. Clean Dataset- Output_Flow_1.

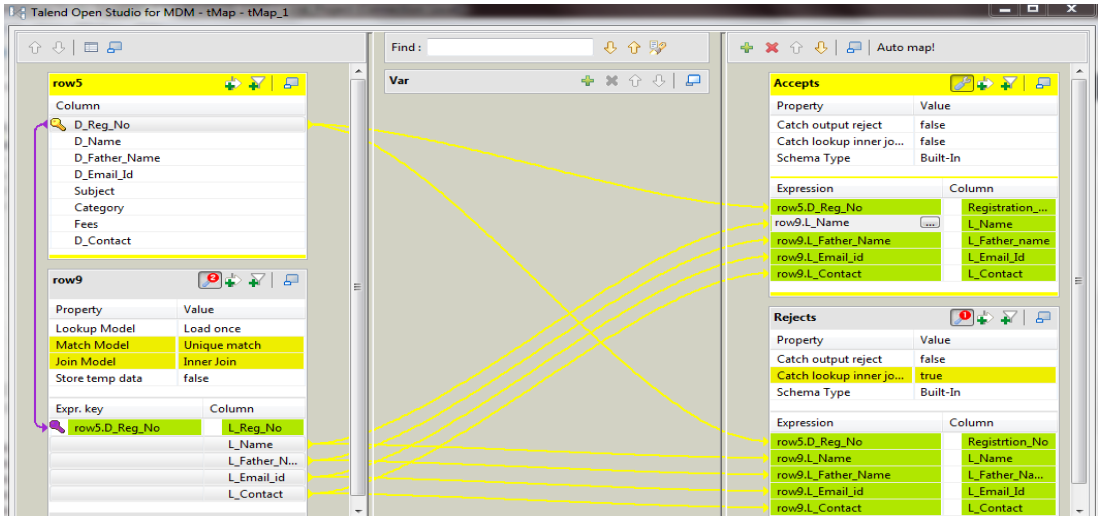


Fig. 8. Data Transformation using Talend Mapping Window

Fig. 9 describes the integrated result of mapping component as Matched records common to both department and library. In mapping, department and library records are mapped to generate a single copy of matched records by considering "Registration No". In above table, the attribute: registration number taken from department, whereas Name, father Name, Email and contact are taken from library record. The matched records are common records to library and department. Rejected records are not taken for analysis in present study.

Fig. 10 shows the Rejected records. The records of student which are not present in the library are rejected.

B. Module 2: Data Matching

Matching refers to the method of comparing different data sets and matching them against each other. The objective is to find the data that refers to same entity. To implement matching process, Python Fuzzy Match library is used to evaluate the matching score of records. A fuzzy match represents a match that is not exact. This technique identifies two elements of string, text or entities that are approximately same but not exact. In order to find a match score, four attributes: Name, Father Name, Email Id and Contact are considered for comparison in present research. Following values are evaluated on the basis of matching score of records.

1) ThresholdValue: Threshold value is considered on the basis of maximum matching of attributes. The Figure below illustrates the threshold value taken in the present research for merging of records in Final master Table.

In Fig. 11, the attribute best_match_score represents the matching score of the records. Matching score value 0.561963 is taken as threshold value. The records above or on threshold are considered as best match whereas below threshold and positive match score are average matched.

	Registration_No	L_Name	L_Father_name	L_Email_Id	L_Contact
0	130029	KAVITA	SUNDER LAL	KAVITABISHNOI97209@GMAIL.COM	89******0226
1	130030	VIKAS	JITENDER	VIKURAMPURA@GMAIL.COM	77******0227
2	130031	SUMANDEEP KAUR	LAKHWINDER S.	KAURDA061@GMAIL.COM	75******0228
3	130032	MOHIT	VED PARKASH	RRMOHIT77@GMAIL.COM	96******0229
4	130037	KAWAL PREET	BALRAJ SINGH	SINGHKAWAL97@GMAIL.COM	99******0234
5	130042	ASHOK KUMAR	MAHENDER SINGH	AGODARA7711@GMAIL.COM	96******0239

Fig. 9. Integrated Matched Records after Mapping.

	Registrtion_No	L_Name	L_Father_Name	L_Email_Id	L_Contact
0	130034	NaN	NaN	NaN	NaN
1	130035	NaN	NaN	NaN	NaN
2	130036	NaN	NaN	NaN	NaN
3	130038	NaN	NaN	NaN	NaN
4	130039	NaN	NaN	NaN	NaN
5	130040	NaN	NaN	NaN	NaN

Fig. 10. Rejected Records after Mapping.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_(
775	0.596529	MAMTA RANI	MAMTA	BHOOP SINGH	BHOOP SINGH	MAMTAK1207@GMAIL.COM	MAMTAK1207@GMAIL.COM	98***
170	0.593535	PARDEEP KAUR	PARDEEP KAUR	KULVINDER SINGH	KULWINDER SINGH	PARDEEPKAUR1313RD@GMAIL.COM	PARDEEPKAUR1313RD@GMAIL.COM	93***
243	0.561963	PRIYANKA RANI	PRIYANKA	KRISHAN KUMAR	KRISHAN KUMAR	PRIYAKAMBOJABC@GMAIL.COM	PRIYAKAMBOJABC@GMAIL.COM	79***
997	0.561095	GAURAV WADHWA	GAURAV WADHWA	PREK KUMAR	PREK KUMAR	GAURAV99@GMAIL.COM	GORUWADHWAW1@GMAIL.COM	81***
999	0.537010	LALITA DEVI	LALITA DEVI	OM PARKASH	OM PARKASH	LALTA89@GMAIL.COM	LALLTINEW01@GMAIL.COM	80***

Fig. 11. Data on the basis of Threshold Value.

2) Best match: The record for which the matching scores value is greater than or equal to 0.561963 are categorized as best match records as shown below.

Fig. 12 shows the best matched records. The highest matching score is 0.947167. All the records above than or equal to threshold value are categorized as best matched records.

3) Average match: The record for which the matching scores value is greater than zero but less than 0.561963 are categorized as average match records as shown below.

Fig. 13 represents the screenshot of records below threshold value and greater than zero. There is need to manually determine which records are to be merged. Hence, these are categorized as Average matched records.

4) Worst match: The record for which the matching scores value is less than zero are categorized as worst match records as shown below.

Fig. 14 describes the worst matched records. All the records with negative matching score are considered as worst match records.

C. Module 3: Data Merging

Merging process allows merging the source data record into destination data record. Two base object records are merged to form a single consolidated base object. For merge operation, the attributes: Name, Father Name, Subject, Category, Fees are taken from department table whereas Email id, Contact are taken from library table.

Considering, k=Threshold Value, m=Matching score

Merge Condition =
- m>=k — Best Match → Automatic Merge
- m>0&&m<k — Average Match → Manual Merge
- m<=0 — Worst Match → Rejected

1) Automatic merge: The records for which matching scores is greater than or equal to threshold value are categorized as best match and considered for Automatic merge. Automatic merge is simply performed by merging corresponding attributes from two datasets. The result is Final Master Table containing Master record or Unified version of Truth for each record.

2) Manual merge: The records for which matching score is below than threshold value or greater than zero are categorized under average match and considered for manual merge. Data Steward helps to manually determine which records should be considered for a match. A data steward is accountable for carrying out data usage and security policies as determined through organization. The resultant records are stored into Final Master Table.

Fig. 15 depicts the Master Table after applying data enrichment, matching and merging. The content of the master table represents the Golden copy of each record stored in a single table. The attributes of table D_Name, D_father_Name, Subject, category, Fees are taken from department and L_Email_Id, L_Contact are taken from library. Thus, it represents a unified version of truth consolidated from multiple sites. It is essential to adopt an MDM strategy as master data represents the most valuable business objects and agreed upon the information that is shared among the organization.

D. Module 4: Data Governance

Data Governance is the set of policies, roles and standards that guarantees high quality throughout the lifecycle of data. It is the process of governing how data is used, by whom and when. Data Governance ensures that data accessibility, security and management rules are followed consistently, every working day. Services provide by data governance are: aligning rules and policies, establishing accountability and decision rights, specifying data quality requirements and performing data stewardship. An optimized data governance program will underpin the business transformation toward operating on digital platform at many levels. MDM requires Data governance activities for agreement on business and technical resources, data to be mastered, business rules and policies, consolidation rules and data quality. Thus, implementation of MDM is irrelevant without doing the significant bit of data governance.

1) Data validation: The purpose of Data Validation is to envelope accuracy: "The degree of similarity of an action to a standard or a genuine worth" and validity: “the degree to which data conforms to defined business rules”. It ensures that information present in the system is correct. Thus, in this study, the scope of Data validation is specified to figures out what data to validate, when to validate data and how to manage data that fails validation. Thus, Data Validation being an integral part of governance improves data quality and efficiency of the system.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Conta
840	0.947167	RANJEET SINGH	RANJEET SINGH	HARPAL SINGH	HARPAL SINGH	BTECH.SULEKHA@GMAIL.COM	BTECH.SULEKHA@GMAIL.COM	99******02
1001	0.929684	ASHWANI AHUJA	ASHWANI AHUJA	DILBAG RAI	DILBAG RAI	ASHWANIAHUJA2009@GMAIL.COM	ASHWANIAHUJA2009@GMAIL.COM	90******02
116	0.920330	SAHIL NARANG	SAHIL NARANG	SHIV DAYAL NARANG	SHIV DYAL NARANG	SAHILN321@GMAIL.COM	SAHILN321@GMAIL.COM	95******02
1000	0.901804	VIJAY KUMAR	VIJAY KUMAR	BACHANA RAM	BACHANA RAM	VIJAYVAID77@GMAIL.COM	VIJAYVAID77@GMAIL.COM	89******02
838	0.896942	DIKSHA JASUJA	DIKSHA JASUJA	BHAGWAN DASS	BHAGWAN DASS	ABHIMONARK@GMAIL.COM	ABHIMONARK@GMAIL.COM	98******02

Fig. 12. Best Match Records.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Cont
243	0.561963	PRIYANKA RANI	PRIYANKA	KRISHAN KUMAR	KRISHAN KUMAR	PRIYAKAMBOJABC@GMAIL.COM	PRIYAKAMBOJABC@GMAIL.COM	79******0
997	0.561095	GAURAV WADHWA	GAURAV WADHWA	PREK KUMAR	PREK KUMAR	GAURAV99@GMAIL.COM	GORUWADHWAW1@GMAIL.COM	81******0
999	0.537010	LALITA DEVI	LALITA DEVI	OM PARKASH	OM PARKASH	LALTA89@GMAIL.COM	LALLTINEW01@GMAIL.COM	80******0
225	0.476170	SAPNA RANI	SAPNA RANI	JARNAIL SINGH	JARNAIL SINGH	SAPNAKAMBOJ9696@GMAIL.COM	SAPNAKAMBOJ9696@GMAIL.COM	93******
998	0.463360	RAJENDER SINGH	RAJENDER SINGH	SHER SINGH	SHER SINGH	RAJSINGHSERA56@GMAIL.COM	RAJUSINGH77@GMAIL.COM	82******0

Fig. 13. Average Match Records.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Con
784	-0.517985	NITIN KUMAR SACHDEVA	RITESH KUMAR	BINDER SACHDEVA	RAVINDER	BAJAJ7201@GMAIL.COM	RITESHKUMAR66@GMAIL.COM	99******0
672	-0.540876	SUNDER	ASHOK KUMAR	KAILASH	MAHENDER SINGH	SUNDER2233@GMAIL.COM	AGODARA7711@GMAIL.COM	88******0
1064	-0.620879	SOUJANYA UPPAL	AJAY	RAKESH UPPAL	RAMESH KUMAR	SOUJANYAUPPAL99@GMAIL.COM	WAYASIJA23@GMAIL.COM	92******0
62	-0.665611	MOHD SOAIB	POOJA	MOHD ILYAS	JAGDISH KUMAR	SOABMOHD9191@GMMIL.COM	POOJAWALECHA490@GMAIL.COM	75******0
355	-0.695725	AKSHAY BANSAL	MOHIT	MR.KULBHUSHAN BANSAL	VED PARKASH	AKBANSAL77@GMAIL.COM	RRMOHIT77@GMAIL.COM	99******0

Fig. 14. Worst Match RecordsBest Match Records.

	D_Name	D_Father_Name	Subject	Category	Fees	L_Email_Id	L_Contact
0	PRIYANKA	DINESH KUMAR	MCA-32	GEN	8230	AGYATT@GMAIL.COM	90******0201
1	POOJA	JAGDISH KUMAR	MCA-32	GEN	8230	POOJAWALECHA490@GMAIL.COM	77******0207
2	SAHIL NARANG	SHIV DAYAL NARANG	MCA-32	GEN	8230	SAHILN321@GMAIL.COM	95******0210
3	MONIKA VERMA	SATPAL SINGH VERMA	MCA-32	GEN	8230	MONUVERMA1577@GMAIL.COM	99******0213
4	PARDEEP KAUR	KULVINDER SINGH	MCA-32	GEN	8230	PARDEEPKAUR1313RD@GMAIL.COM	93******0214
5	KARAMJEET KAUR	JAGJEET SINGH	MCA-32	GEN	8230	KHUNDAL07@GMAIL.COM	92******0215
6	MANPREET KAUR	RANJEET SINGH	MCA-32	SC	750	MK9821752@GMAIL.COM	94******0216
7	GURJASHAN SINGH	GURTEJ SINGH	MCA-32	GEN	8230	GURJASHAN80@GMAIL.COM	93******0223
8	AJAY	RAMESH KUMAR	MCA-32	GEN	8230	WAYASIJA23@GMAIL.COM	99******0225
9	VIKAS	JITENDER	MCA-32	GEN	8230	VIKURAMPURA@GMAIL.COM	77******0227

Fig. 15. Master Table.

2) Policies and rules: The significance of Data governance policy is tied straightforwardly to the significance of a solid data governance program. A committee or governance team establishes policies and rules over collection, storage, usage and security mechanism of organization's data programs. Following functions are articulated in this study for making a governance policy.

- Reliable, proficient and successful administration of the information resources all through the organization and over the long run.
- Designing of Laws and regulation specifically designed for organization's data program.
- Proper assurance and security levels for various classifications of information as set up by the governance team.

In present paper, an algorithm has been designed to create Master data, with input data sources containing number of records: 1940 and 497, respectively. The records of these data sets are matched to find the similarity score using fuzzy matcher. On the basis of score and threshold considered, the input data set is classified into three types as shown.

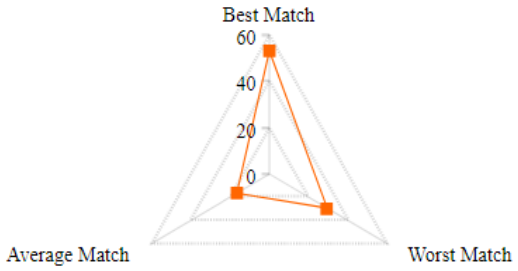


Fig. 16. Threshold based Classification of Input Data Sets.

The above Fig. 16 categories the input data set as: Best match, Average match and Worst match records. Matching is performed using fuzzy matcher to find a similarity score of a record. On the basis of match score, threshold value is considered to define the merge condition. The records above the threshold are taken as best matched records whereas below threshold and greater than zero are average matched records. Matching score with negative value are categorized as worst matched records. It is examined that the records in best, average and worst matching are: 53.2 %, 16.6% and 29.48% respectively for above said input. Master Table has been created using best and average matched records by automatic and manual merge process. It ultimately helps in improving data quality, business operation and analytics.

VII. Conclusion

In present study, an algorithm for critical data consolidation in MDM has been designed for small organizations. As Small and mid-sized organizations do not have the required capabilities to support best practices in MDM. They often underestimate complexity, cost, the flexibility to easily add attributes, and level of collaboration required for MDM program. Thus, to realize this target, a holistic approach for resolving and managing critical data entities in MDM has been presented in the present paper. In addition to data consolidation, Identity Resolution technique is also designed which helps in assessment of records while making "Unified version of truth". To discover and assess the quality of the identifying attributes, TALEND tool for MDM has been used in this research. Data quality over collected dataset is enriched with TALEND tool. By using fuzzy matcher, the matching score of corresponding attributes of input datasets is calculated. On the basis of threshold value, records are categorized as: Best match, Average match and Worst match. Best and average matched records are merged to generate Master Table while worst match records are rejected. Thus it is concluded that this approach is an optimal fit solution in small organization which enables users to integrate and circulate single standard view of master data across the frameworks like ERP, CRM, Apps, and systems. An effective MDM technique helps the business over wide variety of activities like reporting, up-selling and cross-selling of decision making and observance. The chances of being business successful, increases with significant implementation of master data.

References

[1] Rada chirkova, Jon Doyle, Juan Reutter, “Ensuring Data Readiness for Quality Requirements with Help from Procedure Reuse,” Journal of Data and Information Quality, ACM digital library, vol 13, issue 3,April 2021.

[2] Panagiotis Lepeniotis,“Master Data Management: Its importance and reasons for failed implementations,” Sheffield Hallam University, Ph.D thesis, Jan 2020.

[3] Fernando Gualo, Ismael Caballero, Moises Rodriguez, “Towards a software quality certification of master data-based applications,” Software Quality Journal, 28(3), 1019-1042, 2020.

[4] Shreya Mrigen, Dr. Vikram N B, “Relevance of Master Data Management in Pharmaceutical Industries,” International Journal for Research in Applied Science & Engineering Technology, 8(6), 190-197, 2020.

[5] Panagiotis Lepeniotis Master Data Management: Its importance and reasons for failed implementations, Sheffield Hallam University, Ph.D thesis, Jan 2020.

[6] Dilbag Singh, Dupinder kaur, “A Master Data Management solution for building frameworks: a constructive way to pilot the implementation,” in 2nd international conference on Data Analytics and Management (ICDAM), Springer 2021.

[7] Chun Zhao, Lei Ren, Ziqiao Zhang, Zihao Meng, “Master data management for manufacturing big data: a method of evaluation for data network,” World Wide Web, Springer, 23, 1407–1421, 2019.

[8] https://blog.semarchy.com/backtobasics_data_classification last accessed on 05/08/2021.

[9] https://www.stibosystems.com/what-is-master-data-management last accessed on 05/08/2021.

[10] Aditya Rahman A, Gusman Dharmaet al. “Master Data Management Maturity Assessment: A Case Study of a Pasar Rebo Public Hospital. 2019,” International Conference on Advanced Computer Science and information Systems (ICACSIS), IEEE, 497-504 Bali, Indonesia, 2019.

[11] Igor Prokhorov, Nikolai Kolesnik “Development of a master data consolidation system model (on the example of the banking sector),” Post proceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA, 142, 412-417, Elsevier, Prague, Czech Republic, 2018.

[12] F. G. Pratama et al.: Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture. International Conference on Computer, Control, Informatics and its Applications (IC3INA), 1-6, IEEE, Tangerang, Indonesia, 2018.

[13] Z. Murti et al,: Master Data Management Planning: (Case Study of Personnel Information System at XYZ Institute). International Conference on Information Management and Technology (ICIMTech), 160-165, IEEE, Jakarta, 2018.

[14] S. Thomas Ng et al.: A Master Data Management Solution to Unlock the Value of Big Infrastructure Data for Smart, Sustainable and Resilient City Planning. Procedia Engineering, 196, 939-947, Elsevier 2017.

"Digital Influencer": Development and Coexistence with Digital Social Groups

Jirawat Sookkaew[1]
Department of Computer Graphic and Multimedia
School of Information and Communication Technology
(ICT), University of Phayao (UP)

Pipatpong Saephoo[2]
Department of Business Digital
School of Information and Communication Technology
(ICT), University of Phayao (UP)

***Abstract*—Digital identities, also known as virtual influencers, are created by humans through the creation of digital tools that mimic human behavior through the use of creative design. As a result of this, it has resulted in the creation of a group of people who are very fond of and trendy called "virtual influencers", particularly in the modern day. With the rise of virtual influencers, they must be used as tools in marketing and media, particularly in the online world. Because such a character is able to overcome a variety of limitations that humans are unable to provide, character styles, which do not need to have the same look or composition as people, are factors that make these characters popular, but the development of the virtual influencer depends on the social and cultural factors of the people of that era, as well as relying on technology to play a role for humans to be able to apply and use these elements to integrate with the existing virtual influencer to grow and develop more.**

Keywords*—Virtual influencer; online social; virtual character; media

I. Introduction

Today's technology and creative innovations, such as the development of multimedia technologies, including visual and entertainment innovations, have greatly expanded human imagination and creativity. Today's technology makes it possible to create realistic images of people. The general public can create images and video materials using design software or programs from their personal computers, as well as the tools used to create them. Humans began to study and create until humans in today's society encountered and became accustomed to virtual images and 3D designs in everyday life. It appears that these computer-generated images or characters are a part of the social and media lifestyle that we see and consume daily without feeling alienated.

Human society has advanced in computer technology, which has resulted in the advancement of computer graphics, and 3D technology as byproducts of development, as well as technology in the filmmaking, animation, and game industries. Beginning the creation of 3D works, and pushing the field of 3D computer graphics to people, with the release of Star Wars in 1977, a movie that the world recognized as a visual effect work of the world. The art of creating three-dimensional works has progressed to the present day since then.

Creating an imaginary world entails much more than simply building a structure or purchasing an object. Because technology has progressed to the point where tools can generate characters or simulate people from 3D programs. Furthermore, simulated people are now easily achievable, allowing people to create and simulate characters or objects, as well as create a fictional character. "Avatar" is a term for a fictional character that is processed by software to make it more intelligent. Furthermore, this effect contributes to the creation of a more realistic display for imaginative characters. Today's technological advances have created worlds, games, and applications in which people can meet, talk, and socialize without physically being present. Individuals create virtual avatars, or representations of themselves, to represent the user in those environments [2]. Not only can the characters take on the form of people, but they can also take on the form of monsters, imaginative characters, and characters created by the creators.

II. Problems of the Study

The purpose of this research paper was to investigate the popularity of virtual influencers as well as the factors that contributed to the popularity and support of the technology trend. This includes factors that will strengthen the current trend and encourage it to progress and develop further.

III. Method

The researcher used a method to analyze data from popular virtual influencers in various forms, such as analyzing the strengths that are regarded as supporting factors in the modern era in various aspects related to lifestyle and human media use. The current technology study has the potential to help drive the technology of the digital society of virtual influencers and summarize the issues by using this method.

IV. Character Development and Creation of Fictional Characters

Humans have long worked together in designing, thinking, and imagining. We have discovered imaginations of fictional deities dating back to recorded history, ranging from fictional paintings on cave walls to oil paintings of imaginary gods on the walls of famous artists' palaces. Visual stereotypes can appear in many different media formats: photos, movies, paintings, drawings, comics, animated movies, etc. Some of these formats allow different degrees of visual naturalism, which in the context of stereotypes is a feature of interest [1]. Visual Stereotypes and Virtual Pedagogical Agents Rather than using narrative, belief bonding, or even communication, characters are created by imagining gods or characters. The

attentive presence of a virtual character appears to be effective in reducing subjective stress responses, especially when the user believes the virtual other is controlled by a human [9]. This is accomplished by imagining fictional characters and transforming them into images for public broadcast. For a long time, humans have been modeling the shape of creating an ideal appearance, whether it is the nature of the gods and goddesses who have a beautiful appearance according to the ideals of each nation. For a long time, stereotypes or fictional examples have existed in human society to represent the archetype of strength, beauty, and perfection that is not found in the average human being. Virtual influencers bring new scenarios to digital marketing and more; if most analyses point to more convergence than divergence between real and virtual and between human and non-human [2].

In today's world, graphic design has become more straightforward. Creating the ideal character to represent or communicate the creator's imagination or the need for a person with the desired characteristics. The use of media technologies is growing increasingly prevalent around the world; within these technologies, users are finding ways to portray themselves. The more frequently our social interaction occurs via mediation, the more common the use of avatars will become [3].

CGI (computer-generated imagery) has been a popular tool in the entertainment industry for many years. Animated graphics, figures, and characters in films and video games have changed our perceptions of our surroundings [4]. In today's world, graphic design has become more straightforward. Creating the ideal character to represent or communicate the creator's imagination or the need for a person with the desired characteristics. CGI has been a popular tool in the entertainment industry for many years. Animated graphics, figures, and characters have altered our perceptions of our surroundings, whether in film or video games [5]. Character creation, or character design, maybe another way of reflecting one's identity and the need for a human appearance can be conveyed and presented through imagination and innovation in each era.

V. Creating a Digital Identity to Meet the Requirements of the Digital Market

The media in which humans see animated characters, monsters, and fictional characters have become the norm in today's communication and media world with the creation of human-made characters, thanks to the advancement of digital multimedia technology. There was a previous love trend when the trend of characters or cartoons flourished due to the appearance of anime characters. The following is an example of a trend that has sparked interest in the animation industry around the world, particularly in Japan, the East's leader. It is a mascot created for the Japanese voice-editing software Vocaloid, also known as Humanoid. The character's name is Hatsune Miku. A teen girl of 158 cm height and 42 kg weight appears. Her hair was turquoise, and her outfit was a short white skirt. She was adorable. This character was born in the year 2007. The Anime style, which is a technique used to create two-dimensional characters, was used to create her. The characters' popularity has spawned a massive industry phenomenon, including the ability for fans to organize their concerts utilizing 3D holographic character presentations. Furthermore, another response was their superior ability to humans' in communicating in a frequent way. This enhanced the level of intimacy with virtual influencers, showing that there were tendencies into developing bonds with them [6]. It was a way for the character to interact with her fans in person in 2014. This is yet another phenomenon in the digital world that demonstrates and conveys the fandom's affection for young girls. A large number of people also follow her on social media. More characters based on Miku's approaches and prototypes were created to satisfy the growing fan base as the character's popularity grew. This popularity has also spread outside of Japan to several other countries. This is reflected not only in the mashup of musical styles and aesthetics that underpin Hatsune's image as a performer but also in the numerous contributors to her live performance as well as her growing discography, which was created using synthesizer software [7]. Miku's success was most likely due to more than just the use of adorable new characters. Other factors that contribute to this character's popularity include Japanese culture, technology, sound, performance, release timing, appropriate publicity, and a variety of other factors. This phenomenon implies that the characters created are not always realistic, humanoid, or human representations.

The competition format and digital creations have become more realistic as we move into a more advanced era of virtual reality technology. We get to see the film characters, as well as the setting and fictional characters. In the entertainment industry, virtual reality graphics are widely used to create a more realistic appearance layout. Producers and cinematographers can express their imaginations in films by using characters created with graphics rather than human limitations. The images created by using these fictional characters are no different than those created by using real stars, people, or by creating and stimulating locations. The props are more realistic as well. It demonstrates digital characters' ability to transcend physical limitations or meet needs that humans are unable to meet. When such tools are capable of creating or producing people in such a way that it is difficult to tell them apart, creating fictional identities brings them to life. A variety of factors support and drive these factors. It is possible to make friends or form friendships in the modern era, as well as to make oneself known through a variety of social media channels. As a result, birth in the online world is regarded as a channel for presenting or informing birth. To present these digital characters, the majority of digital influencer channels currently emphasize the use of social media. Famous digital influencers were all discovered on Instagram, which is significant in and of itself. People consume content quickly on this platform, and the dimensions of the screen dictate how visual information is processed. At the start of their careers, digital influencers frequently use side channels. The online world is the primary channel for birth notification.

A persona that aligns with the target audience should be thoughtfully developed and executed. The content creators should develop an identity for the CGI that would include all components of a belief system, unique to the CGI. The CGI

should ultimately mock a human, physically, emotionally, mentally, and spiritually. Just as the CGI and brand must align for effective advertising, as in all advertising, the brand and target audience must align. Therefore, the brand, CGI, and target audience must align [8]. Furthermore, Virtual Influencers may emerge from well-known and popular virtual characters in the digital world, such as characters from computer games, animated movies, brand characters, mascots, and so on. "It is clear that the virtual aesthetic of video games dominates this collection." When we consider heroines, or what might constitute the nature of a woman whose actions can be so courageous that she becomes superior and iconic, it becomes clear that a virtual entity integrates with Maison's founding principles. Lightning is the ideal avatar for a global, heroic woman, as well as for a world in which social networks and communications are now inextricably linked. She is also a representation of new pictorial processes. How do you create an image that deviates from traditional photography and design principles? Lightning signals the beginning of a new era of expression." Nicolas Ghesquière [9]. It not only brings popular characters from niche platforms, such as computer games, such as Final Fantasy's Lightning, but it also brings this distinct character to the world of fashion. It is also expected to continue the trend and raise awareness among other consumer groups, as well as establish a link between the brand group and the game player. This is a good way to connect relationships and create marketing channels, such as when creating the Kumamoto mascot, a black bear with red cheeks who represents Kumamoto Prefecture in Japan. Kumanmon is regarded as a pivotal figure in Kumamoto. Kumamon travels all over the world as the prefecture's sales manager, introducing people to the wonders of his hometown and its abundance of nature, seafood, and delectable fruits [10]. When people are recognized as representatives in the character, the character is a representative or image of the place, implying that the character is a representative by default. Items associated with a particular character are well-received and desired by both fans and customers when a fan base develops for that character. Kumamon's popularity has had its impact, with merchandise sales reaching approximately 29.3 billion Yen in the first year of its promotion into the local budget flow. This result was very close to the total income of Kumamoto Prefecture, which is well-known for its agricultural products [11].

VI. The Social Development of Digital Influencer

The number of Virtual Influencers is growing, and the typical influencer model is to penetrate a specific market segment, such as being an influencer in one area such as travel, fashion, gaming, performances, sports, and so on. Countries or regions that wish to present language-specific content, however, the content presented retains the concept of coexistence with the characters' real life by referring to the current and imitating real life to reach or be closer to the person, for example, living like a normal person, interest in the online world like a normal person, expressing feelings or creating imitation behavior in real life based on people's age and character. It is likely that many brands will find the use of CGI as influencers appealing. The benefits are not limited to the potential of increased sales, more reach, and a reputation as an innovator. Should this be the new normal, brands may need to adjust and more CGI influencers may be needed [8]. It is thought that the character should be presented to get as close to and reach the person as possible. Although CGI allows for far more possibilities than photographic or realistic representation, these accounts attempt to depict scenes that are somewhat similar to or resemble real-life environments and scenes (Dovile Dudenaite). Human behavior may be imitated in the performance or presentation. Virtual Influencers want to sense the physical presence of the characters in the real world. After getting to know these characters, even fans and aficionados realize they are not real. The appreciation style in personality, stereotypes, or expressions, on the other hand, is most likely a big part of how people follow or become a fan base and support. CGI influencers are having an impact on the influencer marketing landscape and have the potential to change how brands communicate with their audiences [4].

VII. Strengths of Virtual Influencer that Aid in its Growth

A. Physical and Medical Limitations

Creating a virtual influencer is certainly a character regardless of physical use or physical health, and creators can accordingly customize the character's physical characteristics, as well as the content presentation or gestures. Because computer-generated influencers are not constrained by energy levels, family commitments, or overtime legislation, they are essentially available for brand use 24 hours a day, seven days a week [12]. Clients or employers can then select one of these influencers to provide an unlimited supply of product work. These Virtual Influencers have an advantage in being a good choice for today's entertainment or advertising business due to rest or physical factors. Characters with a creative streak can take risks or create images in dangerous or difficult-to-access environments. Traditionally, brands and advertisers have used real-life models as influencers to promote their brand or product. Fashion models were chosen to be slim and beautiful. These models came with overhead costs such as hair and makeup, lodging, transportation, their own reputations and sometimes a troubled past or future. With the advent of computer-generated images (CGIs), all of those overhead costs and potential public relations crises can be eliminated [8].

B. Workplace, Travel, and Time

In today's world, where traveling or moving is difficult due to the travel health influencers' safety epidemic, visual influencers are digital information that can be displayed on modern computers and mobile phones. As a result, the budget for Virtual Influencer-related expenses has been cut. We can change the character's position to suit the task or information to be presented. The presence of software assistance and computing capabilities allows for a more realistic display, removing the need for scheduling or job queuing. We could meet Virtual Influencers on the other side of the world and then vanish in an instant. These abilities could be one of the characteristics that make influencers more appealing to their followers.

C. Additional Customizations to fit the Content

Creating value or content for an influencer is simpler and less expensive than it is for the average person, as changing outfits and adding accessories around the body can change the composition in a very short time depending on the difficulty of creating a 3D piece for assembly. As a result, the new customizing influencer method saves money and time when creating content. Each image publicity for Influencers or singers, in general, necessitates a significant investment in terms of resources, expenses, and personnel. Creating content for these Virtual Influencers necessitates not only the use of computers to create images but also the character creators' skills and abilities.

D. Bypassing Legal Constraints

The media, labor, and the common man at work are all protected in advertising circles, as is the scope of the population's legal protection. Each region or country has its own set of specifications. They are rules that humans use to control and protect themselves to reduce problems and protect their rights. Many of these concerns are alleviated by virtual influencers: because they are not humans, there is no ethical issue of branding attached; the image can remain consistent; and the risk of indiscretions is minimized because they do not exist offline, allowing their "behavior" and image to be calibrated in the background (J. Tan). The rise of the artificial intelligence influencer: are they simply easier to work with? August 27, 2019, Marketing Interactive [13]. However, there is no specific law that governs Virtual Influencer. Despite their appearance and elements that are difficult to distinguish from the real person, the preliminary images are interpreted as a type of PR medium. The work and the virtual influencer continue to be fictitious characters, akin to cartoons or animated characters. Communication or overwork will work with these characters, so the style of content that appears with Virtual Influencers is much broader and more flexible than we've seen before. However, the requirements of different countries' media laws vary. Virtual Influencer characters are still regarded as media. Laws governing the public presentation of media also govern the rating of published media. Although CGI influencers provide brands with new ways to express themselves, the newfound power must be managed [4]. The creation of Virtual Influencer content necessitates knowledge of the countries and community's laws and customs. Even if the characters are real people, cartoons, or made-up. Although CGI influencers provide new avenues for brands to express themselves, the newfound power must be managed [12]. It is critical to have a regulatory and review body, as well as legal content related to Virtual Influencers, such as the Federal Trade Commission (FTC), whose mission is to protect consumers and competition by preventing anticompetitive, deceptive, and unfair business practices through law enforcement, advocacy, and education while not unduly burdening legitimate business activity. When receiving media, including communication from both Virtual Influencers, it is aware of information presentation and consumer protection.

VIII. Accelerating the Future Virtual Influencer

Although Virtual Influencers are gaining popularity in many parts of the world today. The character creates a distinct identity and persona to gain recognition and love from the online world as a gateway, as well as people's and brands' attention. Virtual, or rather artificial, influencers operate online in the same way that real ones do. Brands want to collaborate with them to design their fan base. Even if they are not intended to be brand ambassadors, their popularity will almost certainly attract companies looking for endorsement deals [14].

Popularity and trend in the online world are thought to be another driving factor of influencer identity, but with current trends and developments in many areas of technology, especially nowadays, it is evolving quickly. The advancement of computer graphics, as previously stated, has fueled the trend of Virtual Influencers. Online media is a factor that contributes to the trend's popularity. The Virtual Influencer's form and characteristics can be modified and developed further through the use and support of technology, which can be supplemented in the following ways.

A. A.I.(Artificial Intelligence) System Development

It is a type of technology that is vital in today's world and is used in all industries. It was created to be a simulation of the human nervous system, which processes information from learning and optimizes itself. Furthermore, it can analyze and process on its own in the process of thinking and self-development by relying on digital tools and computer systems. A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes [15]. The advancement of digital human technology has a positive impact on the AI system development model. The seemingly infinite number of technologies, techniques, and applications that fall under the AI umbrella can be usefully divided into two categories. The first consists of knowledge-based systems that are "committed to the notion of generating behavior through deduction from a set of axioms" [16].

Which AI technology drives innovation in a variety of areas, including ways to accelerate advancement that can be combined with Virtual Influencer, a computer graphics applied science that can bring AI technology to promote this industry and society? In the future, the project Baby X will be developed as a clear example of the use of multimedia technology combined with AI processing. Dr. Mark Sagar created Baby X, a digital artificial intelligence mixed reality installation. It is a neuro-behavioral computational model with emergent behaviors that are actively used for neuroscientific research and, at times, a public media art installation [17]. Based on the Baby X project's experiments, we can anticipate the emergence of a new type of virtual influencer in the future - a framework for interaction that could serve as a model for the next great relevant project. We can predict and see the direction of the development of existing virtual influencers or new virtual influencers from such an intriguing project, which is likely to be a tool to help foster the ability to create interactions and allow virtual influencers to meet and greet fans. Virtual Influencers, on the other hand, do not have this capability. Bringing these incredible AI systems and

capabilities to Virtual Influencer could add augmented reality to the characters, charisma, and senses, bringing the virtual influencers to life in ways never seen before. AI is now being used by digital character creation companies to create characters for businesses, organizations, and even individuals. It was created by Soul Machines. The company is a pioneer in the field of character innovation or the simulation of real-life people.

This is a first step toward bringing Virtual multimedia and AI into the industry in the future by customers who have joined the corporate character creation when combined with the application of AI technology to be able to interact with people and study the needs of interactions to be the most similar to human personality. Vodafone ("Kiri"), NAB's UBank ("Mia"), the Ministry of Primary Industries ("Vai"), ASB Bank ("Josie"), UBS Bank ("Daniel Kalt"), and, more recently, BMW, Southern Cross, and BCG are among its customers (and digital humans) [18] and respond in an emotionally appropriate manner, providing the embodiment of the to the brand and developing brand loyalty and advocacy, bringing "humanness" into digital experiences, increasing sales conversions and customer advocacy (Monica Collier Scott Manion). Artificial Intelligence in Action: Digital Humans [18] as well as creating identity and recognition for the Brand's personality for customers and visitors. Currently, more and more such services are being born in other companies to create these digital characters, increasing the opportunity for Ai Digital Influencer to expand and widen the market, which benefits the consumer side. Consumption rises when competition in these markets rises. Nevertheless – this portrayal of the ideal can be taken one step further with interactive computer media. A key difference lies in what is otherwise seen as a central potential of virtual characters – not the least in pedagogical terms – namely their interactivity: Virtual characters may communicate, respond, and answer, thus establishing a dynamic, mutual social relation [1].

B. Robotics Advancements in Industrial Technology

We are familiar with the visualization and perception of virtual influencers' presence when viewing a digital display through a screen, but what if these influencers could transform into real people? They're just as real as robots or characters from science fiction movies. The task of perception is another important application of AI in robotics. Robots can sense their surroundings using integrated sensors or computer vision [19]. By mentioning the introduction of an AI system to manipulate and work with characters in the Virtual Character format to create a fictional character with a system that mimics human interactions as closely as possible. Virtual influencers who are unable to meet and interact with real people and followers will benefit from the adoption of this technology. Furthermore, the virtual influencer's difficulties in considering the specific context in real life were identified [6] and when considering the study and production of robots as another science that complicates. In fact, the development of these robots entails a plethora of subtleties and techniques. Humanoid robots should share the same working space as humans and should respond in a human-like manner. As a result, they require a light-weight body, high flexibility, a wide range of sensors, and a high level of intelligence [20].

Structuring is an important part of imitating and simulating a wide range of structural features, such as muscles, organs, hair, and skin, all of which necessitate the use of technology, science, engineering, art, and beauty in the creation of the humanoid. This is yet another feature that improves the Virtual Influencer industry while also introducing a new presentation style. In the future, having a unique digital identity or simply seeing these characters on screen may not be enough to meet the needs of fans, so creating a virtual influencer in the form of a robot that looks like a real person, known as a humanoid, may be required. A significant amount of difficulty is also involved, with the key being to create characters that allow for as natural an interaction and movement as possible. Robotics is a challenging field, both in terms of engineering human-like movements and expressions and the challenges that arise when a robot assumes human form. With this format, the social and emotional aspects of interaction take precedence [21]. Realbotic, a technology company focused on developing interactive and immersive virtual robots using AI and machine learning processes, as well as the development of external components, is currently developing augmented reality robotics in the industry. The movement and appearance of the characters create a sense of intimacy and an experience similar to meeting a real person for those who interact with them. Humanoid’s creation market has grown since the company's inception. Humanoid robotics labs around the world are working to develop robots that are one step closer to the androids of science fiction. Building a humanoid robot is a difficult engineering task that necessitates a combination of mechanical, electrical, and software engineering, computer architecture, and real-time control [22]. Humanoid's growing popularity has aided the industry and related technologies in improving their capabilities in these fields. In addition to enhancing and driving the virtual influencer industry, Humanoid is involved in promoting and driving other related industries such as entertainment, film, acting, service, helping, and the puppet industry. In a variety of situations, these virtual machines can play the roles of real people.

There are also various factors that support the opportunity and development of the Virtual Influencer to increase and leap forward, such as The development of Internet sensitivity is laying the foundation for today's global systems such as technology internet 5G, Internet of thing (IoT). The presence of a new online community Social media that will support the development and innovation of multimedia technology to make it more accessible to the way of life such as the world of Metaverse. Some believe that virtual worlds will comprise a metaverse, combining immersive VR with physical actors, objects, interfaces and networks in a future form of Internet [23] a social, virtual world that parallels – and in some respects, replaces – the real world [24]. which is regarded as space and the digital world in a new platform born from the development and pushing of the future simulation technology. It also consists of many factors such as the more people have time to dive into the online world. Everyday activities are more involved in the virtual world. Creating a wide range of technological innovations increased research or the price of technology products is more accessible to each individual, etc.

IX. Discussion

The study's design demonstrates that the popularity of the virtual influencer's identity is driven by the trend of online technology, which provides a means to promote recognition and popularity. Today, each of the channels is becoming more widely distributed. Popular social media platforms of choice for influence are Instagram, Facebook, Snapchat, and YouTube. An influencer has the power to affect the purchase decisions of others because of their authority, knowledge, position, or relationship with their audience. It is important to note that these individuals are not simple marketing tools, but rather social relationship assets with which brands can collaborate to achieve their marketing objectives [25]. It is regarded as a factor in promoting human beings' closer and more connected relationships with virtual influencers. As the number of online media channels grows, so does the number of people who like them. The more opportunities for relationships and interactions are better.

The technology is linked to the character's development style. As processing technology and character creation have advanced, the creation of such fictional characters, such as anime, has become more feasible. However, with the advancement of digital technologies, computer generated images (CGI) are becoming more and more human-like, and virtual agents are increasingly capable of simulating human content [26]. It is thought to be another factor that has contributed to the increase in popularity and likes, as well as the broadening of its target audience, possibly including groups of fans of technology, games, cartoons, sci-fi movies, and so on. Furthermore, other factors in the creative technology industry will help promote and create success through virtual influencers by utilizing technologies throughout the world. The development of information and computer technology allows modern man to live a double or triple life, and therefore, one reality will be superimposed on another. Perhaps in the future, the virtual world and the real world will be inseparable, but these are only predictions [27].

X. Conclusion

A study of the factors and successes of virtual influencers, the phenomenon of popularity and emergence, discovered that what helps virtual influencers create characters is also due to nature's imitation and the presence of a virtual influencer. A real person can become the main character of a story or combine or improve the strengths of a variety of personalities to create a unique character. It could be a design based on the character's appearance that matches popular trends and people's preferences, or it could be designing a character that isn't necessarily based on anything that is considered personal. Virtual influencers are created by interested parties, and followers are drawn to the style and uniqueness of the content presented. Influencers are the factors that contribute to people's preferences being met. These virtual influencers are options for people who like and have an attitude toward a character's lifestyle, despite the fact that virtual influencers are not always realistic, humanistic, or display any form of perfection. That is, it could be a new social group that addresses the needs and attitudes of people who have ideas or suggestions in an era when the line between people and the digital world is almost indistinguishable. Existing influencers may be unable to replace virtual counterparts. In a technologically influenced world, it's simply a new subset of preferences and approaches. Virtual influencers are characters whose attitudes and imaginations are enriched by the attitudes or influence of a society that has adopted human-born attitudes and ideas. According to the study, the factor that contributes to popularity could be the form of popularity in appearance and the liking of the ideal character that has been with humans for a long time, such as what we see in the visualization of gods, mythological and mythological characters created to meet the needs and attitudes of demanding physical traits and desired traits. Overall support in both technology and infrastructure is an important factor that will make a difference or drive a virtual influencer's social group to gain popularity and more likes than before. The use of computers and creative technology has gained popularity for the creation of virtual influencer identities, which create beautiful characters that are similar to real people or can be conveyed as closer to humans. People's interest in information technology today coincides with a period in which both social and technological conditions are becoming increasingly involved in human daily life through active lifestyles. The media has changed, and there is still room for growth due to the rate of growth and the number of internet and electronic device users. A computing system in the field of AI that will assist virtual influencers in learning and becoming smarter. Fast Internet connection systems, IOT (Internet of Things) applications, or the robotics industry are also becoming increasingly intelligent. Furthermore, many future academic and creative innovations will help drive the combination of development with the creation of Virtual Influencer characters to grow and progress. The benefits of a virtual influencer to humans also include a variety of computer tweaks and corrections, reduced stress, and a number of physical and mental limitations that real-life celebrities face. At this point, a virtual influencer may be appropriate for activities or areas where relevant trends are likely to occur. Although a variety of factors contribute to the rise in popularity of virtual influencers, there are still many factors that virtual influencers cannot replace today. A distinct way of life in which the naturalness and presentation of content to those who follow is rather unique and communicated directly to the audience. Human development and growth continue to be so enticing that virtual influencers fail to convey these feelings to the viewers. Although virtual influencers can be customized and corrected in body proportions to be perfect or according to societal trends, the charm of change and the development of human growth is still a strength that virtual influencers cannot replace and enhance the strengths that are the hallmark of human evolution. It is an option for those who admire and sympathize with the character's way of life. That is, it could be a new social group formed to meet the needs and attitudes of people with ideas in an era when the line between people and the digital world has almost vanished. It cannot be separated, which means that the popularity of virtual influencers is derived from social behavior, attitudes, technological processes, social trends, and responses to the needs of people who are satisfied with their character and will be able to influence trends and the development of the virtual influencer

society. In terms of living a way of life, conducting social sciences, and pushing themselves to be modern and comfortable, they are consistent and similar to humans. Because of technological advancements, life has become more convenient. Existing ones may not be able to be replaced by virtual influencers. It's simply a new subset of preferences and approaches added to the audience's options. Virtual influencers can only exist because of the encouragement and support of people who value those characteristics. Perhaps the ideal representation of the individual is where a society of virtual influencers and a variety of communication sciences collide to create new forms of self and character that satisfy and delight humans.

REFERENCES

[1] H. Magnus and G. Agneta, "Visual Stereotypes and Virtual Pedagogical Agents," Journal of Educational Technology & Society, vol. 11, no. 4, pp. 1-15, 2008. http://www.jstor.org/stable/jeductechsoci.11.4.1.

[2] B.D.S.O. Antonio and C. Paula, "Humanized Robots: A Proposition of Categories to Understand Virtual Influencers," Australian Journal of Information Systems, vol. 25, 2021. DOI: https://doi.org/10.3127/ajis.v25i0.3223.

[3] J. Fox and J.A. Sun, "Avatars: portraying, exploring, and changing online and offline identities," In Handbook of research on technoself: identity in a technological society, IGI Global, 2013, pp. 255-271. DOI: 10.4018/978-1-4666-2211-1.ch014.

[4] L. S. Walter, "Changing the instagram game: the rise of a new influencer generation," UNIVERSITY OF TWENTE , Enschede, Netherlands, 2020.

[5] K. M. Rahill and M. M. Sebrechts, "Effects of Avatar player-similarity and player-construction on gaming performance," Computers in Human Behavior Reports, vol. 4, no. 100131, p. 100131, 2021.

[6] S. N. Victoria Molin, "Robot or Human? – The Marketing Phenomenon of Virtual Influencers," Uppsala University, Uppsala, Sweden, 2019.

[7] A. J. R. B. Cross., 10 - Keepin' it real? Life, death, and holograms on the live music stage. Chandos Publishin, 2015.

[8] C.D. Oglesby, "The New Frontier of Advertising: Computer-Generated Images as Influencers, 2019.

[9] Louisvuitton.com. [Online]. Available: https://au.louisvuitton.com/eng-au/articles/series-4-lightning-a-virtual-heroine-. [Accessed: 03-Oct-2021].

[10] Ourkhungbangkachao.com.[Online].Available:http://www.ourkhungbangkachao.com/Uploads/articles/Tourism%20Development%20Case%20Study%20in%20Kumamoto.pdf. [Accessed: 03-Nov-2021].

[11] P. Kusuma and D. W. Soewardikoen, "City mascot as A supporting force in city imaging," in Proceedings of the 4th Bandung Creative Movement International Conference on Creative Industries 2017 (4th BCM 2017), 2018.

[12] A. V. A. Mary Caroline Creasey, "Virtual Influencing: Uncharted Frontier in the Uncanny Valley," LUND UNIVERSITY, Lund, Sweden, 2020.

[13] "The rise of the AI influencer: Are they simply easier to work with?," Marketing-interactive.com, 27-Aug-2019. [Online]. Available: https://www.marketing-interactive.com/rise-of-ai-influencers-how-you-can-get-a-piece-of-this-pie. [Accessed: 03-Nov-2021].

[14] M. H. H. Zdenka Kadekova, "Influencer marketing as a modern phenomenon creating a new frontier of virtual opportunities," Communication Today., pp. 90–104, Jan. 2018.

[15] R. J. Schalkoff, Artificial Intelligence: An Engineering Approach. New York, NY: McGraw-Hill, 1990.

[16] N. Cristianini, "On the current paradigm in artificial intelligence," AI Commun., vol. 27, no. 1, pp. 37–43, 2014.

[17] D. Lawler-Dormer, "BABY X: Digital artificial intelligence, computational neuroscience and empathetic," in ISEA 2013 Conference proceedings.

[18] Org.nz. [Online]. Available: https://aiforum.org.nz/wp-content/uploads/2019/10/FaceMe-Case-Study.pdf. [Accessed: 31-Oct-2021].

[19] Javier Andreu Perez, Fani Deligianni, Daniele Ravi and Guang-Zhong Yang, "Artificial Intelligence and Robotics," 2016.

[20] K. Berns, T. Asfour, and R. Dillmann, "Design and control of the humanoid robot ARMAR," in Romansy 13, Vienna: Springer Vienna, 2000, pp. 307–312.

[21] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," Found. Trends® Hum.–Comput. Interact., vol. 1, no. 3, pp. 203–275, 2007.

[22] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: a new kind of tool," *IEEE Intell. Syst.*, vol. 15, no. 4, pp. 25–31, 2000.

[23] J. Smart *et al.*, "A cross-industry public foresight project," *Metaverseroadmap.org*. [Online]. Available: https://metaverseroadmap.org/MetaverseRoadmapOverview.pdf. [Accessed: 25-Oct-2021].

[24] W. Mason, "Why social VR is a huge priority for Oculus, but their metaverse is still far off," Uploadvr.com, 13-Oct-2015. [Online]. Available: https://uploadvr.com/oculus-social-vr-metaverse-palmer-luckey/. [Accessed: 17-Oct-2021].

[25] S. Olenski, "The Impact Of Live Streaming On Influencer Marketing," Forbes Magazine, 25-Sep-2017.

[26] J. Arsenyan and A. Mirowska, "Almost human? A comparative case study on the social media presence of virtual influencers," Int. J. Hum.Comput. Stud., vol. 155, no. 102694, p. 102694, 2021.

[27] E. Samoylova, "Virtual world of computer games: Reality or illusion?," *Procedia Soc. Behav. Sci.*, vol. 149, pp. 842–845, 2014.

Modified Deep Residual Quantum Computing Optimization Technique for IoT Platform

Rasha M. Abd El-Aziz, Alanazi Rayan, Osama R. Shahin, Ahmed Elhadad, Amr Abozeid, Ahmed I. Taloba
Department of Computer Science, College of Science and Arts in Qurayyat
Jouf University, Saudi Arabia

***Abstract*—Internet of Things (IoT) is defined as millions of interconnections between wireless devices to obtain data globally. The multiple data are targeting to observe the data through a common platform, and then it becomes essential to investigate accuracy for realizing the best IoT platform. To address the growing demand for time-sensitive data analysis and real-time decision-making, accuracy in IoT data collecting has become critical. The Res-HQCNN is a hybrid quantum-classical neural network with deep residual learning. The model is qualified in an end-to-end analog method in a traditional neural network, backpropagation is used. To discover the Res-HQCNN efficiency to perform on the classical computer, there has been a lot of investigation into quantum data with or without noise. Then focus on the application of the artificial neural network to analyze the dangers to these IoT networks. For data recording purposes, to undertake in-depth analysis on the threat severity, kind, and source, a model is trained using recurrent and convolutional neural networks. The intrusion detection system (IDS) explored in this study has a success rate of 99% based on the empirical data supplied to the model. Due to irregularly distributed robust execution, larger affectability for the introduction of authority dimension, steadiness, and the extremely large crucial area, a quantum hash function work has been proposed as an amazing method for secure communication between the IoT and cloud.**

***Keywords*—Internet of things (IoT); cloud; Res-HQCNN; intrusion detection system (IDS); optimization**

I. Introduction

Artificial neural networks (ANN) are one of the most successful computational approaches. Neural network-based machine learning algorithms are improving and advancing [1]. In the machine learning sector, neural networks are currently enjoying remarkable success and have a wide range of applications, including pattern recognition, video analysis, medical diagnosis, and robot control. Quantum neural networks (QNN) appear in parallel with the development of artificial neural networks (ANNs), with the promise of overcoming classical computation limits using quantum computing [2]. The paper shows a quantum feed - forward neural network made up of genuinely quantum neurons. It has a remarkable capacity to study an unknown homogeneous and a high level of robustness when dealing with noisy training data. Due to a decrease in the number of coherent qubits, this process is essential for noisy approximate quantum computers. The adherence among a pure quantum system and an arbitrary quantum state is selected also as cost function in this study. However, as the number of network layers grows larger, the convergence rate of a cost function slows, and the value of convergence even fails to deliver the highest for clean data is shown in Fig. 1(a). In the case of noisy data, Fig. 1(b) shows it as the system gets deeper, the strength for noisy data weakens. As a result, guess if the cost function's efficiency can be enhanced both for clean and noisy data. To show the number of convolution layers in the corresponding layer, use a one-dimensional list of real numbers.

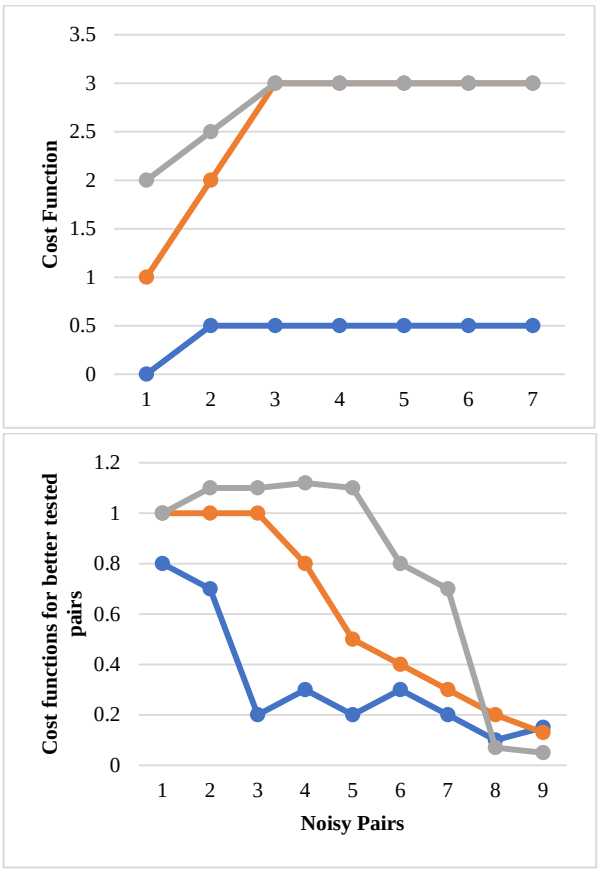


Fig. 1. QNN Numerical Results for Clean and Noisy Data.

To enhance the novel quantum-classical neural network with the deep residual learning (Res-HQCNN) to attain a goal, inspired by the deep residual efficiency learning. This is a novel concept and no work has been attempted as far as know. To find the efficiency incorporate a residual scheme into QNNs. It is not a simple task [3]. The amount of residual block structures, levels of the network count, and whether or not to skip the layer all have an impact on the parameter updating method. Because the informing parameters of the

matrix are derived from the function of derivative formulation, the updating parameters matrix for each network is unique. Res-HQCNN now outperforms previous QNNs on both fresh and loud quantum data while requiring only a conventional machine to build. Discuss a different way for incorporating quantum neural networks with residual block structure so that they can be executed on quantum computers.

Networks intrusion continues to occur. This is even though numerous artificial intelligence techniques have been created throughout time to prevent such incidents [4]. Various advancements and modifications are made to network configuration protocols daily to improve them, but in reality, there is a weakening risk of the protocols with or without understanding. Malware and other intrusions frequently take advantage of minute alterations made to the original core development codes that serve as the foundation for running and maintaining networks. The changes are vital, but they come at a high price. It's time to rethink your threat management strategy [5]. IoT and Cloud Computing benefit in the same way and Cloud Computing is constantly encouraged to improve the introduction to the level of high resource usage, accumulation, necessity, and processing capability.

However, network intrusions continue to occur. This is regardless of the fact that multiple artificial intelligence techniques have been developed over time to prevent such incidents. Various progressions and modifications are made to configuration management protocols on a daily basis in an effort to improve them, but in reality, there is a risk of decimating the procedures with or without knowledge [6]. Malware and other intrusions frequently take benefit of minute changes respect to the design foundational development rules that serve as the foundation for operating and maintaining networks. The changes are needed, but they come at a high cost.

Cloud is a ground-breaking platform that can provide additional features as an information distribution delegate. When an IoT client has valid requests for specific information to be acquired, stored, and accessed, he can simply designate the requests to the cloud whenever with more remarkable comfort. A couple of incites linked to contraption disillusionment are addressed by cloud and IoT applications developed in resource-constrained conditions. A QHF is proposed to address IoT security concerns. It converts an old-style message to a Hilbert space, preventing programmers from obtaining too much information about the old-style message. Safety issues are of extreme importance, and they must be addressed without exacerbating the system's or devices' dimensions [7]. A few calculations concerning safety issues have been published in prior studies. The U-2 hash work is the largest class of hash capacity groups among known hash capacity groups, assuring good safety.

The number of residual block structures, the number of network layers, and whether or not to skip layers all have an impact on the variable propagation algorithm. Because the updating parameters structure is derived from the description of the derivative feature, the updating parameters structure for each network structure is unique. The updating parameters matrix becomes more complex as the network structure becomes more varied. As a result, this investigation is both challenging and intriguing. This hope that our paper will serve as a useful resource in this field of study. The following are some of the contributions made as a result of this paper:

- Develop a new residual learning structure that is focused on QNNs.
- Calculate the current training algorithm using the Res-HQCNN model. Examine the performance from the level of information propagation feedforward and backward, subset of the training algorithm.
- concentrates on using Artificial Neural Networks to evaluate the risks to such IoT networks. For data acquisition reasons, a classifier is constructed using recurrent and convolutional neural networks to perform effective analysis on threat intensity, type, and source.
- Res-HQCNN has better performance across both clean and noisy quantum information than previous QNNs at the cost of implementation.

The remaining part and the aim of this paper have explained the RES-HQCNN optimization technique for IoT; Section 2 defines the highlight of the previous effort that can be done by the scholars in this domain; Section 3 offering the methodology architecture model and its mechanism, Section 4 represents the result and discussion and Section 5 represents the work achieved in conclusion and future work.

II. Related Work

The author in [8] evaluates Quantum Computing (QC) has grown in popularity as a result of its unique characteristics, which, in terms of performance and operation methods, differ from typical computers This research proposes hybrid models and approaches for large-scale mixed-integer programming issues that successfully combine the complementary strengths of deterministic algorithms and quality control techniques to solve a combinatory difficulty. Large-scale instances of these application problems across multiple dimensions, ranging from molecular design to logistics optimization, are computationally demanding for deterministic optimization algorithms on classical computers. To address the computing challenges, hybrid QC-based approaches are suggested, with comprehensive computational experimental results demonstrating their pertinence and productivity. The suggested QC-based solution approaches offer high computational efficiency in terms of solution quality and computation time by leveraging the unique properties of both classical and quantum computers.

The author in [9] introduces a Deep residual network with adequate depth but bounded width has recently been proven to be capable of universal approximation in the sense of the supremum norm. Illustrate to adapt existing deep residual network training methods to establish approximation bounds for the test error in the supremum norm based on the training error using these results. This technique is based on control-theoretic interpretations of these networks in discrete and continuous time, and they show that constraining the set of parameters to be learned in a way that is consistent with most commonly used training procedures is sufficient.

The author in [10] is proposed to use a combination of modified deep learning and reinforcement learning in an incentive-based demand response (DR) algorithm. A modified deep learning model based on recurrent neural network (MDL-RNN) was initially suggested to forecast future environmental uncertainties by projecting day-ahead wholesale energy price, photovoltaic (PV) power output, and power load. Then, using reinforcement learning (RL), researchers looked at the best incentive rates for each hour that would maximize earnings for both ESPs and EUs. When compared to other methods, the findings demonstrated that the proposed upgraded deep learning model can produce more precise forecasting predictions A short-term DR program was developed for peak electricity demand periods, and trial results show that peak electricity demand can be reduced by 17%. This helps to improve power system security by reducing supply-demand imbalances.

The author in [11] improve the feature mapping process, introduce a hybrid quantum-classical convolutional neural network (QCCNN), which is based on convolutional neural networks (CNNs) but is optimized for quantum computing. In terms of both the number of qubits and the depths of the circuits, QCCNN is favorable to existing noisy intermediate-scale quantum computers, while keeping crucial aspects of classical CNN, such as nonlinearity and scalability. Also, offer a methodology for computing the gradients of hybrid quantum-classical loss functions automatically, which may be extended to other hybrid quantum-classical algorithms directly. By using a Tetris dataset to demonstrate the architecture's capabilities, show that QCCNN can perform classification tasks with learning accuracy that exceeds that of standard CNN.

The author in [12] analyze in classical systems, Control parameter optimization is frequently achieved using supervised machine learning and reinforcement learning; however, in quantum systems, parameter optimization is primarily accomplished using gradient-based greedy methods. To use differential evolution methods to avoid the non-convex optimization stagnation problem. To improve quantum control fidelity for noisy systems by averaging across the objective function. To reduce processing costs, this paper proposes methods for early run termination and adaptive search subspace selection. The implementation is massively parallel and vectorized to further reduce execution time. Quantum phase estimation and quantum gate design are two instances where these methods outperform greedy algorithms in terms of fidelity and scalability.

III. Proposed Methodology

In this section, the Res-HQCNN architecture model is defined based on QNN. According to the mechanism Res-HQCNN is defined based on a training algorithm.

A. Architecture Model of Res-HQCNN

In Res-HQCNN describe a residual block structure. The Res-HQCNN structure with many layers is offered. Thus offer Res-HQCNN examples with a unit hidden layer to further understand the mechanism [1]. Finally, examine the difference between the past QNNs and the Res-HQCNN. Thus, Fig. 2 shows the residual block diagram. In Res-QCNNN, a new residual block diagram is defined as follows by including a few assumptions and notations at the start for your convenience.

The procedure for combining the residual block structure with the quantum neural network in Res-HQCNN layer k defines a quantum perceptron as an arbitrary unitary operator with U_{k-1} input qubits and one output qubit. For L=1,2,…U_k the quantum perceptron Q^k_L is a $(U_{k-1}+1)$ for qubit unit. Quantum perceptron's with K hidden layers make up the Res-HQCNN. It uses the layer unitary operator Q^k in the form of a matric product of quantum perceptron's to work on an input state $\rho^{k_{in}}$ of input qubits and obtain a mixed state k+1 out for the output qubits: $Q^k = Q^k_{u1} Q^k_{u1-1} \ldots . Q^k_1$. For 1,2,…$Q_{k,}$ acts on the qubits in layers k-1 and 1, because the unitary operators are arbitrary and do not always commute, the layer unitary order is critical. The residual block structure provides the new input state for layer k+1 by adding the input state with the output state of layer k for k=1,2, …. During the processing of information from 1 into K+1 out and K.

In Fig. 3, "Res" denotes the Res-HQCNN residual block structure. The "Res" can be connected not just layer by layer continuously, but also by skipping one or more levels [13]. Res-architecture HQXNN propagates data from input to output, progressively passing through a network of quantum feeder neurons.

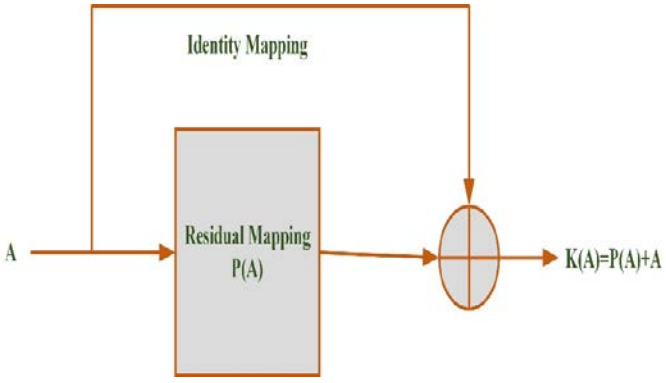


Fig. 2. Residual Block Diagram.

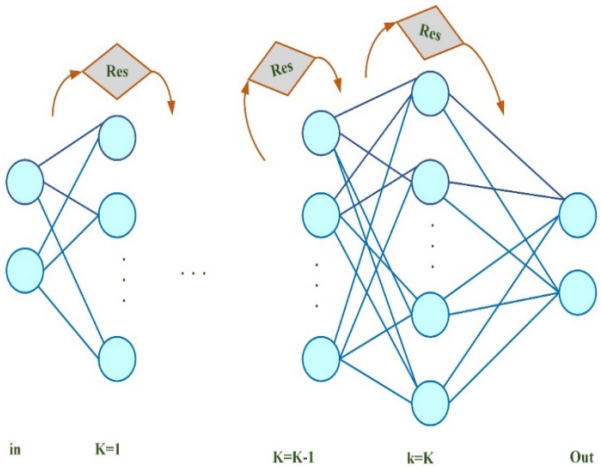


Fig. 3. Res-HQCNN Architecture.

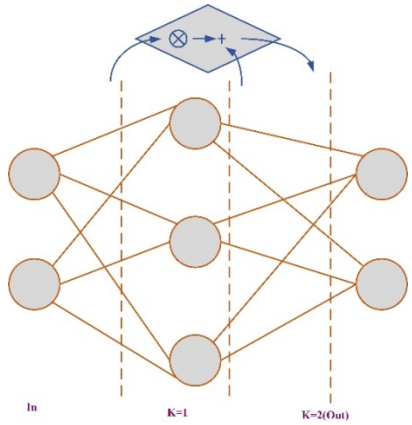


Fig. 4. Res-HQCNN Architecture with unit Hidden Layer.

The mechanism example for Res-HQCNN with one hidden layer in Fig. 4 helps with the comprehension. $Q^1 = Q_3^1 Q_2^1 Q_1^1$, which is a matrix product of quantum perceptron. The layer unitary between the input layer and the hidden layer is defined as $Q^2 = Q_2^2 Q_1^2$. The quantum perceptron's are applied layer by layer from top to bottom in the first stage, and the output state ρ^{1out} of the hidden layer is then computed as;

$$\rho^{1out} = is_{in}(Q^1(\rho^{1in} \otimes |000\rangle_{hid}\langle 000|)Q^{1^+})$$

Then, apply the residual block diagram to ρ^{1in} and ρ^{1out} to get a new input state for the output layer:

$$\rho^{2in} = \rho^{1out} + (\rho^{1in} \otimes |0\rangle\langle 0|)$$

In the following step, to obtain the final output state of Res-HQCNN from Fig. 5:

$$\rho^{1out} = is_{hid}(Q^2(\rho^{2in} \otimes |00\rangle_{out}\langle 00|Q^{2^+})$$

When comparing the previous QNNs to Res-HQCNN, that notice the trace value of the input state ρ^{k+1in} for some k changes as a result of the addition operation in the residual block structure [14]. Indicate k=2, $\rho^{2in} = (\rho^{1in} \otimes |0\rangle_{u_1-u_0}\langle 0|) + \rho^{1out}$, and $\rho^{3in} = (\rho^{2in} \otimes |0\rangle_{u_2-u_1}\langle 0|) + \rho^{2out}$, next trace values of the ρ^{2in} and ρ^{3in} are the 2 and 4, respectively. In theory, ρ^{2in} and ρ^{3in} are not density matrices, hence the training procedure cannot be used in a quantum computer. Every coin, however, has two sides. The residual block structure increases the cost function's performance, especially for deeper networks, as shown in the experiment section. It's also worth noting that the residual block structure can be applied to all concealed layers except the last output layer. Assumed $U_{k-1} \leq U_k$ for k = 1,2,….k and $U_0 = U_{k+1,}$ then the qubits in layer k in general. The final output of the network will be $\rho^{out} = \rho^{k+1in} + \rho^{k+1out}$ if the residual block structure is applied to the ρ^{k+1out} because the dimension of the ρ^{k+1in} is the equal to the dimension, that should use the partial trace on the ρ^{k+1in} to maintain the matrix addition rule.

The previously stated, the Res-HQCNN residual block structure has trouble similar to individuality mapping. But, by doing it will lose some ρ^{k+1in} information, which is incompatible to adopt the residual technique [15]. This also highlights the inefficiencies of applying residual block structure to the last output layer via an experiment.

B. Res-HQCNN Training Algorithm

N number pairs of training statistics, that are possibly unknown by the quantum states, are randomly generalized in the form of $(|\emptyset_a^{in}\rangle, |\emptyset_a^{out}\rangle)$ with a=1, 2,….., and N. It is also permissible to employ adequate copies of a training pair $(|\emptyset_a^{in}\rangle, |\emptyset_a^{out}\rangle)$ of a given a to avoid quantum projection noise, when compared to the cost functions derivative [16]. The intended output $|\emptyset_a^{out}\rangle$ as $|\emptyset_a^{out}\rangle$ = T $|\emptyset_a^{in}\rangle$ is choose to consider with an T as unknown unitary operation.

The cost function is used based on the mean fidelity of the Res output HQCNN and the expected results for all training data. However, to define the Res-HQCNN cost function to divide 2v, where v is the residual block number structures in Res-HQCNN, according to the residual block definition structure and fidelity linear fidelity:

$$R(f) = \frac{1}{2^v N}\sum_{a=1}^{N}(\emptyset_a^{out}|\rho_a^{out}(f)|\emptyset_a^{out}\rangle$$

To know the near network output state and the desired output state are, the higher fidelity. If the cost function equals 1 and 0, consider the Res-HQCNN to be the best performer to be the worst. As a result, the goal in the training process is to maximize the cost function. For each Res-HQCNN layer, that denote $\rho_a^{l_{in}}$ as the layer input state l and $\rho_a^{l_{out}}$ as the output layer state l with l=1,2,….L and a = 1,2,…….N. Consider the scenario in which each layer is added with a residual block structure and there is no skipping layer, then v=L. The Res-HQCNN training algorithm is explained in the following flow chart in Fig. 5.

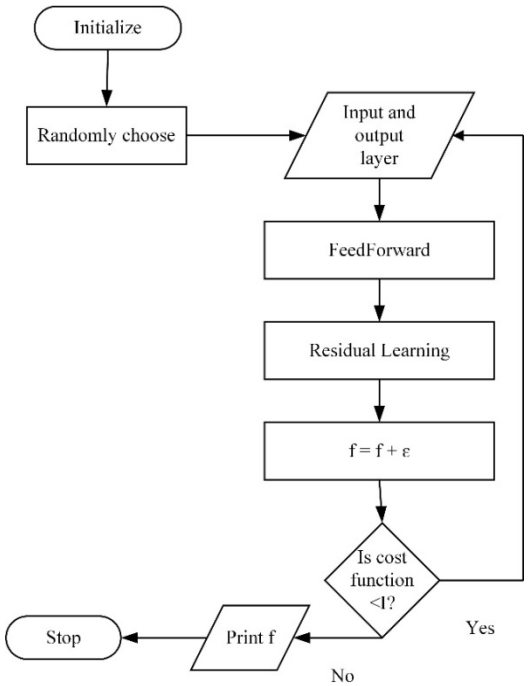


Fig. 5. Training Algorithm Flowchart.

Using a quantum hash function secure the data communication with encryption and decryption. They may occur some threat intrusion. Some types of threats and intrusion are following.

C. Types of Threats

By encrypting and decrypting data using the quantum hash function, data communication can be secured. They may be subjected to some sort of danger incursion [17]. Threats and intrusions of all kinds are on the way. Threats come in various forms:

- Malware.
- Data Loopholes.
- Feeble IoT network outlines.
- Service Denial.

D. Types of Intruders

The types of intruders are trying to intoxicate the network are following:

1) Outer intruder: This is an intruder from a different network from the one they're attempting to intoxicate. They use other networks, but they come to the network to distribute threats and recover data, among other things.

2) Inner network: This is an invader from a network other than the one they're trying to infect. They use other networks, but they come to the network for a variety of reasons, including propagating threats and recover data.

The internet connectivity of intruders on both online and offline:

1) Online Intruder: A danger has been identified as emanating from an internet source. This is particularly prevalent because they take advantage of relatively common IP addresses and can simply steal information from users of the addresses by messing with the network's coded backdrop.

2) Offline intruder: This is an invader who has gained access to the network but does not have internet access. There is virtually little technology available to deal with and counter this type of intrusion threat, yet this is a highly dangerous group of people.

E. Threat Proximity

This information is necessary to demonstrate the degree to which a network user is close to the threat described in the studies [18]. Unfortunately, this method can only be accessed by users of the same network. Due to differences in the functioning of the network, it would be more difficult to draw such a conclusion in the case of an external incursion.

To carry out the threat analysis, the input threat is exposed to a combination architecture of RNN and CNN that chucks the data into bits [19]. The data has a gaussian relationship, and it is assumed that the eventual output, after categorization and regrouping, will be a Gaussian distribution in a very precise manner. The following algorithms were used on the training model:

- Levenberg-Marquardt Algorithm: This approach has been utilized for neural network optimization and is highly useful because the threat is measured on a summation basis [20]. The intrusion is described as a collection of minor threats that add up to a level that is regarded as a threat numerically. Because desire a predefined category of various clusters, the neural network was trained with an input that specifies a certain objective.
- Feed Forward Algorithm: The connections employed in the node do not establish a rotating back dependence, which is ideal for the study. This algorithm is used to train the nonlinear optimization model [21]. It is represented in mathematical as:

  $$f(a) = \frac{1}{1+e^{-a}}$$

  $$f'(a) = f(a)\big(1 - f(a)\big)$$

- Backward learning algorithm: Sensitivity to the impacts of the feed-forward approach for model training. As a result, the feed-forward is primarily reliant on derivative functions, resulting in anticipation. Backward training is a strategy for optimizing a model that involves integration techniques [22]. It minimizes J and so optimizes the cost function for the Jacobian Matrix application.

F. Quantum Hash Function

The hash capacity is presented just in one-way great detail. Selecting the work verification, all single-direction QW work is considered [23]. The single path, solid impact opposition, and fragile crash obstruction are the main characteristics of H-work. The following are the quantum hash attributes capacities:

1) One-direction: It is possible to process the S regard S(G) by giving a data G, but it is computationally impossible to discover the basic data G with a given S regard S(G).

2) Frail crash obstruction: Based on the G data, it is impossible to find another data by computer G1 so that S(G)=S(G1).

3) Solid effect opposition: It is computationally impossible to locate the optional two unmistakably data G and G1 such that S(G)=S(G1). When grasping an S work, these three qualities are key models to consider. Quantum hash work, in comparison to old-style hash work, has more favorable circumstances, such as simple execution and a higher degree of security. Our information verification strategy will become more secure over time. the quantum hash work's nitty-gritty technique is depicted in the diagram below.

Parameter to be selected are [c, $\theta 1, \theta 2, \tau$] under the requirements: c is an odd number and $\{0< \theta 1, \theta 2, \tau < \frac{\pi}{2}\}$ here τ – coin state$|0\rangle = \cos\tau|0\rangle + \sin\tau|1\rangle$, c- number of cycles. In addition, $\theta 1$ and $\theta 2$ are the two controllers of C-QW. The two-coin admin controllers are $\varphi^1 and\ \varphi^2$.

$$\varphi^1 = \begin{bmatrix} \cos\theta 1 & \sin\theta 1 \\ \sin\theta 1 & -\cos\theta 1 \end{bmatrix} \quad \varphi^2 = \begin{bmatrix} \cos\theta 2 & \sin\theta 2 \\ \sin\theta 2 & -\cos\theta 2 \end{bmatrix}$$

The underlying one-information bit selects "0" as its value of φ^1 chooses φ^2. The likelihood dispersion is created by rolling one coin and walking DTQW on a cycle substantially influenced by information G [24]. To frame a twofold H computation, multiply all qualities in the following likelihood circulation by 10i times and maintain only their entire number component modulo 2j with a ≥ b. The S respect has a bit length of mj. This is the methodology used in the most recent QH works conspire, which has a higher level of safety than previous ones. To deservedly chose this QH capability as the approval work.

G. Encryption

The encryption framework works with given data and is a key to creating a figured data that may be delivered through insecure channels without risk of being deciphered by anyone that doesn't have the interpreting key. The key was initially subject to two sets of keys, one open and one private, for security concerns [25]. Initially, to encode, then to untangle, and finally in a different way; this is achievable due to the usage of particular mathematical constraints, which have non-reversible features.

Encryption = S (amount of A/T, S, \open key, A/T)

Decryption = ((m^k) | modified A/T (W)\) * open key

Hash efficiently communicates on little information to produce a string with a known length of G. The IoT sight and sound data are validated by the quantum value, open keys, and restricted irregularity has been able to abuse reduced the S-esteem in light of this worth [26]. The quantum value, open keys, and restricted irregularity have been able to exploit bargain the H-esteem in light of this worth to certify IoT sight and sound data.

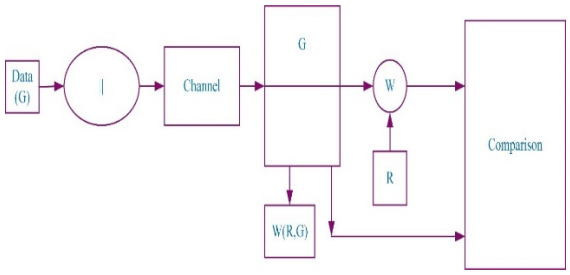


Fig. 6. Hashing Mechanism for Data Authentication.

Fig. 6 demonstrates the general information procedure confirmation. G is the data that will be transferred from the sender to the recipient. W is denoted as verification work is used to scramble the primary information of G. "| |" is a technique for teaching the underlying knowledge as well as the figure script. During correspondence, the square edge is used to symbolize the channel. The key that is utilized to encrypt the underlying data is R.

IV. Result and Discussion

The result examines the Res-HQCNN robustness to noisy quantum data. To compare to employ the same rule to test the robustness. The numerical output from running the two neural networks, RNN and CNN, is shown in the results. As can be seen, the entire input is fed into two robust neural platforms, which optimize the model that has been trained using learning algorithms, and the output is a classification of threats with subfolders indicating the severity of the threats. The level of threats is shown in Fig. 7.

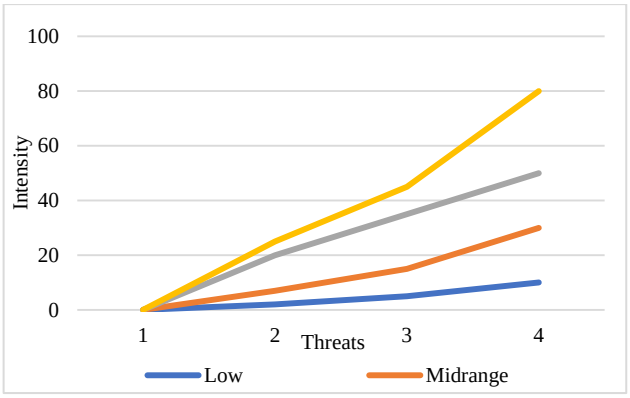
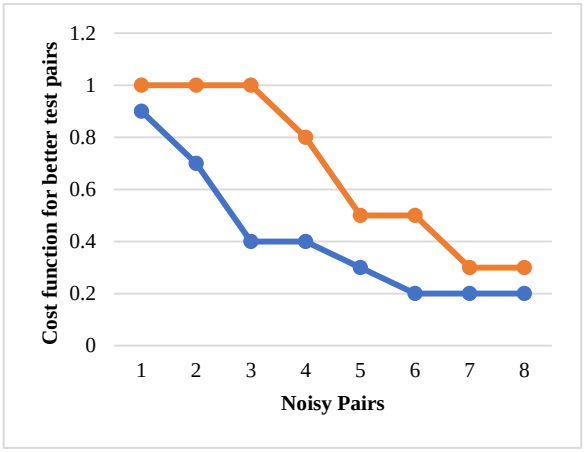
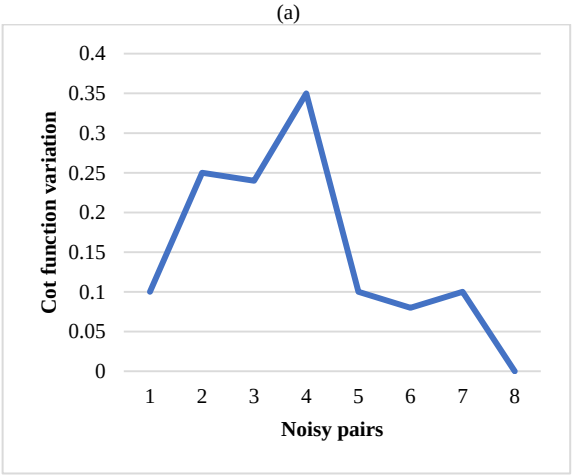


Fig. 7. Threat Ranges.

To produce an N better training pairs as $(|\emptyset_a^{in}\rangle, T|\emptyset_a^{in}\rangle)$, then n destroy by changing them with the training noisy pairs. At each period the changed subgroup is selected arbitrarily. Then, the cost function is measured for all the better test pairs. By select the example, Res-HQCNN [4, $\bar{5}$, 4], with η = 1/1.8 and ε = 0.1 is represented on Fig. 8.



(a)



(b)

Fig. 8. Noisy Training Data Behavior.

In Fig. 8(a), the orange line is the Res-HQCNN results and the blue line are the results, respectively. The Fig. 8(b), plots the cost function variations between the orange lines and blue lines. The x-axis in Fig. 8 demonstrates the number of better training pairs is changed by noisy pairs. Assume a small number of training rounds and pairs, for example, 50 training rounds and 30 training pairs. The cost value for both [4, $\bar{5}$, 4] and [4,5,4] decrease as the amount of noisy pairs raises and the cost value variation is always positive. This demonstrates the [4, $\bar{5}$, 4] superiority for noisy training data with the minimum training rounds and the minimum training pairs. Next, assume the amount of training rounds and pairs are increases as training rounds are 200 and training pairs in 100. Res-HQCNN and QNNs both offer robust toughness to the noisy quantum data when the number of noisy pairings is modest, such as less than 35. The cost values for the orange and blue lines begin to decrease at the same time as the number of noisy pair increases.

When the number of noisy pairings hits 60, the cost variation increases, reaching a maximum when the number of noisy pairs reaches 70. This contains three unstable points (55, -0.0115), (90, -0.0161) and (100, -0.0012) then the variation is negative. There are 21 orange line pairs and blue lines. For every period, the better training data ($|\emptyset_a^{in}\rangle, T|\emptyset_a^{in}\rangle$), and ($|\emptyset_a^{in}\rangle, |\emptyset_a^{out}\rangle$) as noisy training data are produced randomly. The ($|\emptyset_a^{in}\rangle$ *and* $|\emptyset_a^{out}\rangle$) elements are casually chosen out a normal spreading before regularization. The training data randomness produces some uneven ideas, it is shown on comparable results. Then, the Res-HQCNN [4, $\bar{5}$, 4] it shows better robustness to the noisy data than [4, 5, 4] QNNs.

To detect the deeper network as [4,$\bar{5}$,$\bar{6}$, 4] to noisy data is shown in Fig. 9. When the amount of training pairs and rounds are minimum like training rounds as 150 and training pairs as 30 on the figure. To notice a sign of improvement from the figure. The cost function variances are always positive. With an increase in the number of noisy pairs, the variation reduces. There is an amount of training rounds and pairs are large, like training rounds as 600 and training pairs 100 in Fig. 10. It's great to have all cost function variances are always positive and there are no unstable points. The greatest variance value is greater than 0.35, but the one in the figure is less than 0.12. This noisy data is deeper as [4,$\bar{5}$,$\bar{6}$, 4] it shows great improvement than [4,$\bar{5}$, 4]. It going via the studies for Res-HQCNN with or without noise and found that it outperforms QNNs in terms of cost function performance. Although this does not exhibit an outcome for Res-HQCNN with the four or more hidden layers, believe that due to the mechanism of its training method, deeper Res-HQCNN would increase cost function performance more.

Fig. 10 and Table I depicts the final results of the presented parameters in the evaluation. To find the encryption size, disentangling size, memory, and execution time as a function of record size. The result shows that as the archive size grows, so does the encryption and unscrambling size. As a result, the execution time grows as well. The result shows that as the archive size grows, so does the encryption and unscrambling size. As a result, the execution time grows as well. In any event, the given paradigm, which differs from various methodologies, secures IoT data in a high-level manner.

Fig. 11 depicts the throughput rate as a function of database size. For every information base size, the QH work provides an ideal level of safety. In QH work, the level throughput is normally excessive, averaging 90%.

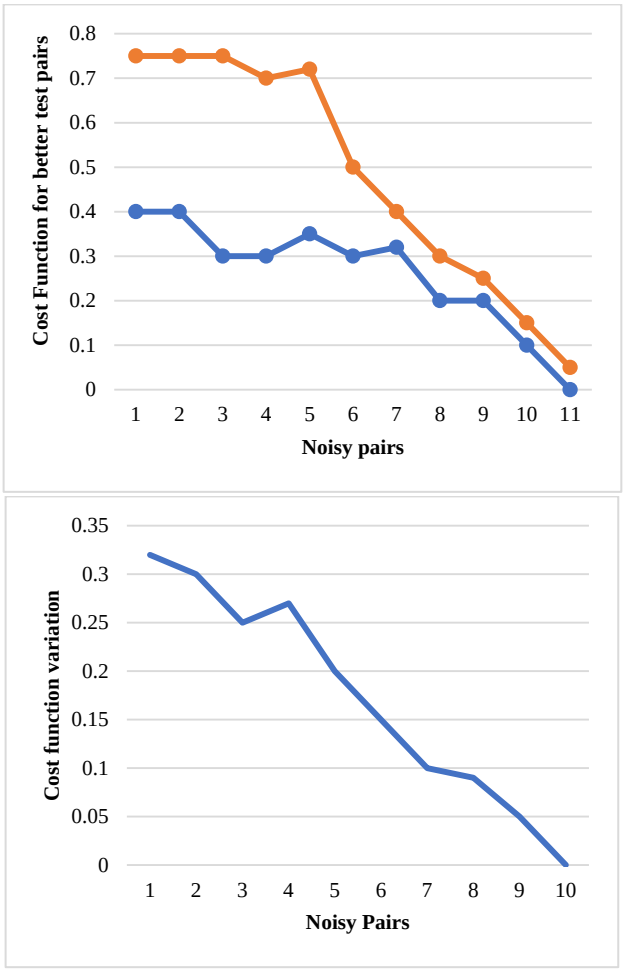


Fig. 9. Deeper Network Detection to Noisy Data.

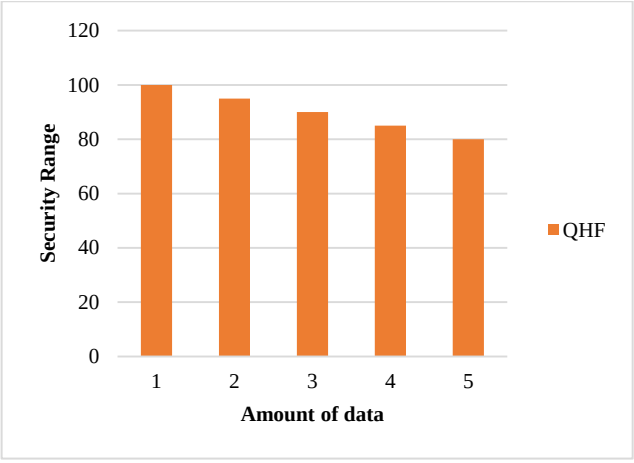


Fig. 10. Amount of Data with Values of Quantum Hash.

TABLE I. QUANTUM HASH FUNCTION

Size of files	Encrypted	Decrypted	Memory	Processing time (ms)
20	27	20	2156432	87231
40	36	40	468769	9423758
60	45	60	476545	10978
80	50	80	576653	113547
100	58	100	563523	115764

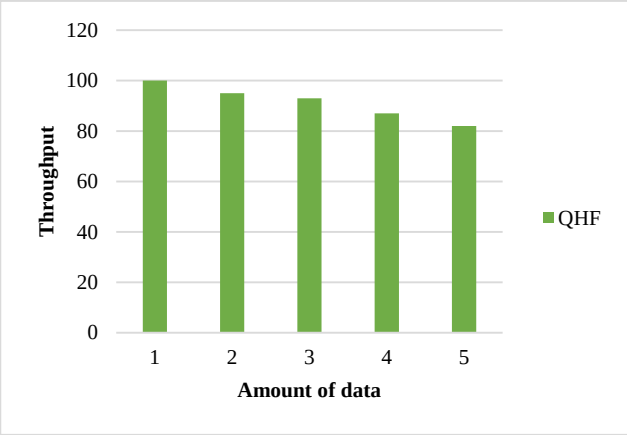


Fig. 11. Amount of Data with the Value of Throughput.

V. CONCLUSION

In this research, to enhance the routine of the cost function for the deeper networks, a quantum-conventional hybrid neural network with deep residual learning was used. Based on the QNNs, a new structure of residual blocks in the quantum concept was developed. Then, the Res-HQCNN training algorithm was also made for different cases. The residual block structure, from the standpoint of information propagation, is similar to the ANN mechanism with deep residual learning in that it permits information to travel from the input layer to any deeper layer. The replications are demonstrated by Res-HQCNN's although it can only work on a regular computer. Due to its non-linear disordered dynamic execution and large key space, quantum hashing work has been proposed as a phenomenal tool for secure IoT communication. The benefits of quantum hashing work have been presented in this research effort as the latest breakthroughs in achieving secure data distribution and information assurance based on Q advancements. For the IDS system, a solution result was modelized using RNN and CNN. It consists of all learning models requested by various network providers. The provided approaches are characterized in terms of increased precision, safety, throughput, and toughness over a few well-known assaults, making them suitable for use in a variety of IoT and cloud applications. In the future, use simulation to investigate the QCNN model, which is more cost effective and has best performance. It is necessary to develop an effective data encoding principle for quantum systems and real information. Finding a way to evaluate threats authorized by offline cyber-attacks is a future suggestion. This study was limited to just online attacks that would be heavily discussed.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research at Jouf University for supporting this work by Grant Code: (DSR-2021-02-0378).

Funding Statement: This work was funded by the Deanship of Scientific Research at Jouf University under grant No (DSR-2021-02-0378).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] Y. Liang, W. Peng, Z.-J. Zheng, O. Silvén, and G. Zhao, “A hybrid quantum–classical neural network with deep residual learning,” Neural Networks, vol. 143, pp. 133–147, Nov. 2021, doi: 10.1016/j.neunet.2021.05.028.

[2] D. Mu, Z. Guan, and H. Zhang, “Learning algorithm and application of quantum neural networks with quantum weights,” International Journal of Computer Theory and Engineering, vol. 5, no. 5, p. 788, 2013.

[3] S. A. Stein et al., “QuClassi: A Hybrid Deep Neural Network Architecture based on Quantum State Fidelity,” arXiv preprint arXiv:2103.11307, 2021.

[4] T. Michael, “CNN Intrusion Detection for Threat Analysis of a Network,” TURCOMAT, vol. 12, no. 3, pp. 3945–3949, Apr. 2021, doi: 10.17762/turcomat.v12i3.1683.

[5] R. Majumder et al., “Hybrid Classical-Quantum Deep Learning Models for Autonomous Vehicle Traffic Image Classification Under Adversarial Attack,” arXiv preprint arXiv:2108.01125, 2021.

[6] K. Shankar, “Improving the Security and Authentication of the Cloud with IoT using Hybrid Optimization Based Quantum Hash Function,” Feb. 2020, doi: 10.5281/ZENODO.3689761.

[7] Y. Yang, Y. Zhang, G. Xu, X. Chen, Y.-H. Zhou, and W. Shi, “Improving the efficiency of quantum hash function by dense coding of coin operators in discrete-time quantum walk,” SCIENCE CHINA Physics, Mechanics & Astronomy, vol. 61, no. 3, pp. 1–8, 2018.

[8] A. Ajagekar, T. Humble, and F. You, “Quantum computing based hybrid solution strategies for large-scale discrete-continuous optimization problems,” Computers & Chemical Engineering, vol. 132, p. 106630, Jan. 2020, doi: 10.1016/j.compchemeng.2019.106630.

[9] M. Marchi, B. Gharesifard, and P. Tabuada, “Training deep residual networks for uniform approximation guarantees,” p. 12, 2021.

[10] L. Wen, K. Zhou, J. Li, and S. Wang, “Modified deep learning and reinforcement learning for an incentive-based demand response model,” Energy, vol. 205, p. 118019, Aug. 2020, doi: 10.1016/j.energy.2020.118019.

[11] J. Liu, K. H. Lim, K. L. Wood, W. Huang, C. Guo, and H.-L. Huang, “Hybrid Quantum-Classical Convolutional Neural Networks,” arXiv:1911.02998 [quant-ph], Aug. 2021, doi: 10.1007/s11433-021-1734-3.

[12] P. Palittapongarnpim, P. Wittek, E. Zahedinejad, S. Vedaie, and B. C. Sanders, “Learning in Quantum Control: High-Dimensional Global Optimization for Noisy Quantum Dynamics,” Neurocomputing, vol. 268, pp. 116–126, Dec. 2017, doi: 10.1016/j.neucom.2016.12.087.

[13] J. Shi et al., “An approach to cryptography based on continuous-variable quantum neural network,” Scientific reports, vol. 10, no. 1, pp. 1–13, 2020.

[14] S. Hou, G. Yang, and H. Xie, “Optimized initial weight in quantum-inspired neural network for compressing computer-generated holograms,” Optical Engineering, vol. 58, no. 5, p. 053105, 2019.

[15] M. P. Heinrich, M. Stille, and T. M. Buzug, “Residual U-net convolutional neural network architecture for low-dose CT denoising,” Current Directions in Biomedical Engineering, vol. 4, no. 1, pp. 297–300, 2018.

[16] W. Jia, Y. Tian, R. Luo, Z. Zhang, J. Lian, and Y. Zheng, "Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot," Computers and Electronics in Agriculture, vol. 172, p. 105380, 2020.

[17] Y. H. Hwang, "Iot security & privacy: threats and challenges," in Proceedings of the 1st ACM workshop on IoT privacy, trust, and security, 2015, pp. 1–1.

[18] M. Abomhara and G. M. Køien, "Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks," Journal of Cyber Security and Mobility, pp. 65–88, 2015.

[19] Y. Hu and X. Lu, "Learning spatial-temporal features for video copy detection by the combination of CNN and RNN," Journal of Visual Communication and Image Representation, vol. 55, pp. 21–29, 2018.

[20] H. Yu and B. M. Wilamowski, "Levenberg–marquardt training," in Intelligent systems, CRC Press, 2018, pp. 12–1.

[21] H. F. Lui and W. R. Wolf, "Construction of reduced-order models for fluid flows using deep feedforward neural networks," Journal of Fluid Mechanics, vol. 872, pp. 963–994, 2019.

[22] D. A. Lorenz and T. Pock, "An inertial forward-backward algorithm for monotone inclusions," Journal of Mathematical Imaging and Vision, vol. 51, no. 2, pp. 311–325, 2015.

[23] D. Li, Y.-G. Yang, J.-L. Bi, J.-B. Yuan, and J. Xu, "Controlled alternate quantum walks based quantum hash function," Scientific reports, vol. 8, no. 1, pp. 1–7, 2018.

[24] R. Bernardo-Gavito et al., "Extracting random numbers from quantum tunnelling through a single diode," Scientific reports, vol. 7, no. 1, pp. 1–6, 2017.

[25] Y.-G. Yang, P. Xu, R. Yang, Y.-H. Zhou, and W.-M. Shi, "Quantum Hash function and its application to privacy amplification in quantum key distribution, pseudo-random number generation and image encryption," Scientific reports, vol. 6, no. 1, pp. 1–14, 2016.

[26] B. Abd-El-Atty, A. A. Abd El-Latif, and S. E. Venegas-Andraca, "An encryption protocol for NEQR images based on one-particle quantum walks on a circle," Quantum Information Processing, vol. 18, no. 9, pp. 1–26, 2019.

Adding Water Path Capabilities to QWAT Databases

Bogdan Vaduva, Honoriu Valean
Automation Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

***Abstract*—The main purpose of this article is to show how to extend an existing open source database, namely QWAT (Acronym from Quantum GIS Water Plugin), by using pgRouting (PostgreSQL routing extension) in order to achieve the ability to find the water flow in a water network. The water path in a water network is a key information needed by any water supplying company for different activities such as customer identification, meter the water flow or isolating areas of the water network. In our environment an open source database was used and that database didn't have any means to identify the water path, so our research is intended into that direction. Once a water path is found, our next goal was to show that identifying customers for a water supplying company is just a click away (by using no directional graphs). Another key information needed by the water supplying companies is to know which valves should be closed in order to shut off the water for an area of the water network. As result, the second purpose of the article is to show how to identify the necessary valves, to be closed or open, in order to shut off or on the water (within the pipe network).**

***Keywords*—Relational database; graphs; water network; water path; open source; QWAT**

I. Introduction

Water supplying companies all over the world needs to know who their customers are and this task is a very important one, for reasons like: knowing to whom they will invoice, knowing how much water was used and if a leak occurs to identify which valves needs to be closed in order to shut off the water. In a previous paper we presented a way of predicting leakage in QWAT[1] databases, but identifying and predicting leakage is not enough, because, any existing or future leakage has a negative impact on the water supplying company's image [1].

QWAT databases are relational databases [2] that models water network pipes, by keeping information like: pipe attributes and geographical position, valve attributes and geographical position, meters, hydrants, network elements, subscribers, leaks. Pipes are kept in a table (within a schema called qwat_od) and have a set of attributes from whom we will focus on two, first node (fk_node_a) and second node (fk_node_b).

In the abstract of this paper, we said that we want to show a way of finding the water path within QWAT databases, but if the pipe table does have a first and a second node, why QWAT model doesn't have a water path function until the final consumer? Or does it? The answer is given by QWAT designers, into their documentation, where they presented how the QWAT model should be used. In that documentation they recommend that pipes should be broken whenever the pipes change their material, function, type or diameter. Other recommendations are:

- Pipes should only be coupled to the right of each intersection with another pipe (Fig. 1).
- Pipes should only be coupled to the right of each intersection with a branch (Fig. 1).
- Pipes should not be coupled in any case to the right of a private branch (Fig. 1).

Those recommendations were put into a picture as in Fig. 1.

If a QWAT database is filled the way described above, the water path should be available, but not actually to the final user/consumer, because, usually, water supplying companies have their own software for customer management and invoicing. We have to outline that, at the time this article was written, there wasn't any function within QWAT database, that allowed to find the water path or extracts the customers for some parts/sections of the water network and we came up with some proposals formulated into a document in 2017. [3].

When it comes to valves they also did some recommendations (Fig. 2):

- Do not cut the pipe on the right of each valve.
- Place the valves on the vertices of the pipe line.

Fig. 1. Example of How to Break Pipes at Intersections in QWAT Databases.

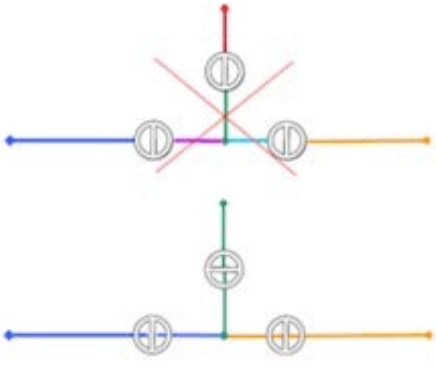


Fig. 2. Example of How to Place Valves in QWAT Databases.

In our case we begin using the QWAT model and we start entering data into that model, which uses a PostgreSQL [4]

[1] QWAT – Acronym for QGIS Water Plugin – http://www.qwat.org

database server. After a few months, we found that even if we somehow link customers, to a water distribution pipe, we can't actually differentiate between customers and all of that, because of the initial recommendations and the way our water network was drew. In our case, the colleagues from GIS (Geographic Information Systems) department, responsible with drawing the water network, were not splitting the pipes every time a pipe intersects a private branch.

We have to clarify what an intersection means from a GIS standing point: two or more pipes that actually connect to each other and not pipes that traverse other pipes (on top or under).

On other hand, getting outside on the field and looking at an actual private branch connection, we found that private branches are directly coupled to the distribution pipe (Fig. 3).

Fig. 3. Example of How a Private Branch is Actually Connected to the Distribution Pipe.

The current paper will continue with the following chapters: Database Background, Database Proposed Changes, Applied Research, Results and Conclusions.

II. Database Background

We started to study the QWAT model and see how and if, can be changed in order to accommodate our needs.

The QWAT model has a table called "pipe" which holds information for each water pipe. That information refers to attributes like: pipe id, pipe function, install method, material, bedding, precision, installation year, node a, node b, parent id, geometry and a few other that are not relevant to our paper. Let's consider a real street, which has a distribution pipe and each private branch is connected to that distribution pipe the way presented in Fig. 3. From a QWAT database standing point, the data is as in Fig. 4 and Table I.

Fig. 4 presents a distribution pipe, on which we placed, one node of the private branches, on its vertices (distribution pipe's vertices). That means that distribution pipe is not fragmented as in Fig. 1 right section and it means that we can't have a path (water flow) from node 38578 (placed on a private branch) to node 38529 (placed on distribution pipe), for example. In order to be able to do that a parallel water network should be created and kept within QWAT database. By having that parallel water network, we will actually create a graph that will contain nodes and edges for describing the water flow.

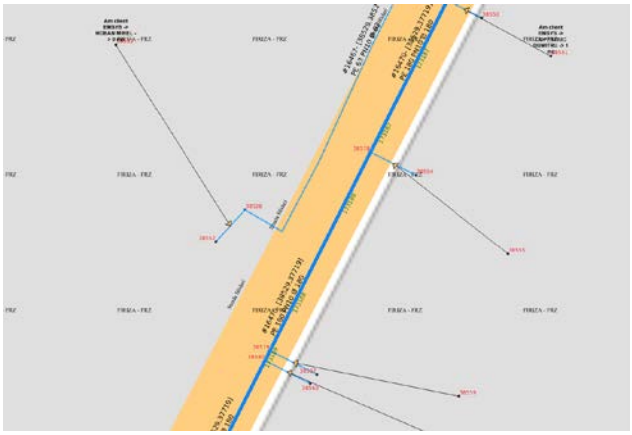


Fig. 4. Example of How a Private Branch is Actually Connected to the main Pipe.

TABLE I. Example of Data Kept in the Pipe Table

Pipe ID	**Pipe Table**				
	Function	***Material***	***Node A***	***Node B***	**…**
16470	Distribution pipe	Steel	38529	37719	…
18900	Private branch	Polyethylene	38580	38560	…
17898	Private branch	Polyethylene	38579	38557	…
14567	Private branch	Polyethylene	38578	38554	…
….					

In the current database model, we can achieve this water path goal only by fragmenting the distribution pipe every time it intersects a private branch. Doing so in a real-life scenario is unrealistic because of the amount of work and can't be achieved.

III. Database Proposed Changes

Analyzing the above data, we came with some initial conclusions and to do list, such as:

1) We need to (somehow) connect private branches to the distribution pipe without redrawing the water network pipes.

2) We need to know/change the valve state in order to be able to have a water path to the final consumer.

3) We need to match a private branch to a consumer.

4) We need to be able to show the water path from the source(s) to any consumer. Knowing that, we will be able to determine which valves should be closed in order to shut off or on the water to a specific customer.

We stated above that we need to address at least four database improvements in order to make it water path friendly. The first improvement on that list: "connect private branches to the distribution pipe without redrawing the water network pipes", it is the one that will allow us to determine the water flow within the water network and further to achieve the other goals on our list.

The first thought we had was to somehow split the distribution pipe, but in real world those distribution pipes are not split (Fig. 3) and our next idea was to build a parallel water

network that will have the desired format. The newly parallel water network will be fragmented as in Fig. 5. Furthermore, that parallel network should be created without asking the user to redraw the network in the format presented in Fig. 1.

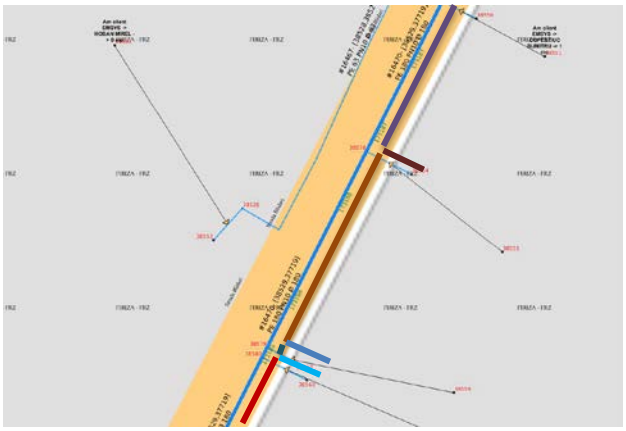


Fig. 5. Example of How we want to Build the Parallel Water Network. We used different Colors just to show that we will have Different Segments.

The parallel network should be built at the same time the user draws the water network in it's usually way, Fig. 4 and 5, where the user does not split the distribution pipe every time it intersects a private branch but places one end of a private branch on the vertices of the distribution pipe. In addition to this goal we figured out that a refresh button that creates that parallel water network for the current bounding box is acceptable and probably necessary.

Once the parallel network will be built we planned to add an existing PostgreSQL extension, called pgRouting [5] to QWAT database.

The first change to the QWAT model was to add a new table, which we called it "pipe_reference" and has the following fields:

- id, integer – an id for each pipe segment that will be generated/created,
- fk_pipe, integer – a reference to the initial pipe,
- fk_node_a, integer – a reference to the first node of the new segment,
- fk_node_b, integer – a reference to the second node of the new segment,
- geometry, geometry – the geometry of the newly created segment.

The reason behind the "pipe_reference" table is to keep the parallel water network saved, for a later use. At this point, we designed the "pipe_reference" table, but that one needed data and we planned to populate it with the help of a new function, called "fn_pipe_reference_update". Further we will present what that function does.

We imagined the function to have as input parameter the ID of the pipe (distribution pipe) that we want to split (ex. 16470 – Table I, Fig. 5). The function works as follow:

- For the input ID we delete all the records from "pipe_reference" having the fk_pipe field equal to the input parameter.
- We select all the pipes (usually private branches) that have one end placed on or near the vertices of the inputted pipe. For all the selected pipes we keep only the nodes placed on the vertices of the inputted pipe.
- Using all the nodes placed on the inputted pipe (distribution pipe), together with the first and second node of it (distribution pipe), we insert the missing rows into "pipe_reference" table (Table II).

TABLE II. EXAMPLE OF WHAT THE PIPE_REFERENCE TABLE WILL HOLD

Pipe Reference ID	**Pipe Reference**			
	Fk_pipe	***Node A***	***Node B***	***Geometry***
1	16470	38529	38577	…
2	16470	38577	38578	…
3	16470	38578	38579	…
4	16470	38579	38580	…
5	16470	38580	38581	…
6	16470	38581	38582	…
7	16470	38582	42086	…
8	16470	40286	37719	…
…				

The next envisioned step was to make a union between the content of the "pipe_reference" table with the records from the "pipe" table for whom the ids cannot be found in any of "fk_pipe" values of "pipe_reference" table. By doing so, we will have the parallel water network (automatically generated), labeled in green, that will be pgRouting friendly (Fig. 6).

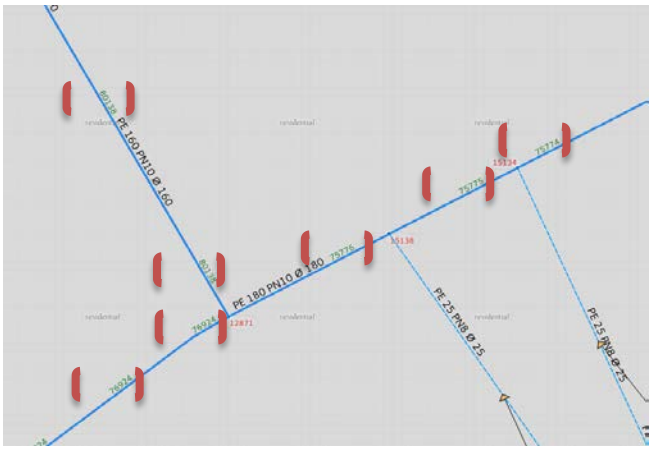


Fig. 6. Example of How the Parralel Water Network was Overlapped with the Pipe Table. Observe the Green Labels in between Orange Brackets on Top of some Pipes.

IV. Applied Research

After envisioning the parallel water network design, we moved forward by adding the new "pipe_reference" table into the QWAT database and the function named "fn_pipe_reference_update" that will construct / fill the parallel network. At the same time, we installed pgRouting in our database server. The new extension, gave us some new routing functions, from which we used "pgr_dijkstra". This function returns the shortest path using Dijkstra algorithm. In 1956 Edsger Dijkstra created a graph search algorithm (graph with non-negative edge path costs) for determining the shortest path [6]. Because the function works on graphs with non-negative edge path costs and because there are valves on pipes (which could be open or closed) we also added a function which we called "fn_element_valve_status". This function checks, if there are valves on a pipe and if one valve is closed on that pipe we consider the cost as being negative.

The example of running the function between node 38557 and node 38529 can be viewed in Fig. 7.

To this point, we addressed only the first and second item on our to do list, but we said that we want have a match between a private branch and a consumer and thus the path from the water source to the final user.

On our map (Fig. 7), we have outlined that each private branch serves a household. At some point the owners can change and we have to take that into consideration. On other hand the information about consumers is kept in a different database that is not by default accessible to QWAT database.

One of the last changes done to our QWAT database was to link the data about consumers into QWAT by adding a new table called "qwat_od_subscriber_location" which has two fields fk_subscriber and fk_location. The first field is a foreign key to "subscriber" table, which is native to QWAT model. The second field is a foreign key to "location" view, in a foreign database that keeps the consumer information and invoices. Regarding the foreign databases that keeps the consumer information and could be linked to QWAT, the only restrictions are related to the fact that those databases should be addressable by a PostgreSQL feature called Foreign Data Wrappers. The "location" view should always give us the information about the current consumer for a specific location.

We built that function that displays basic information about the current consumer at a specific location and the result can be seen in Fig. 7.

We succeed in solving another point in our to do list, but one item was still on that list. It is about the ability to find out if the water flows from one source of the water network to one final consumer.

Knowing which the clients are will allow the water supplying company, to send customized information, to its customers and thus improving its image. Water can flow in either directions on a pipe and as result, we have a no directional graph when it comes to water network pipes. We have implemented the extraction of this graph, in JAVA, on the server side of our web-based application. The result of this tool can be seen in Fig. 8.

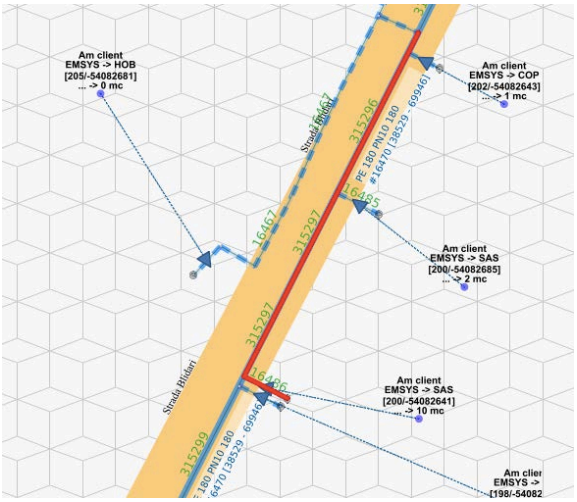


Fig. 7. Example of the Path between Node 38557 and 38529.

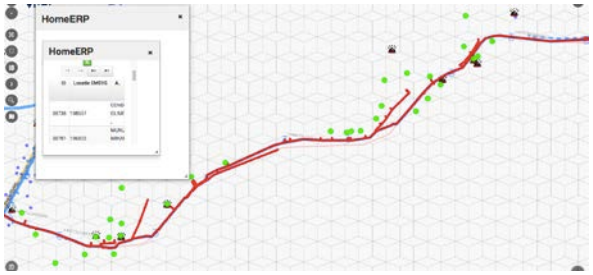


Fig. 8. Example of Determing/Identifying Clients/Customers.

To solve the last item in our to do list, we created a new function within QWAT database that we called "fn_valve_to_close". The new function takes as input parameter, the subscriber, which translates as the end node of consumer's private branch and determines all the water paths from the sources, defined for that water network, to that subscriber/consumer. For all the paths we outline the valves and we color the closest one with green. The result of running the function, on subscriber corresponding to the node 38557 can be seen in Fig. 9.

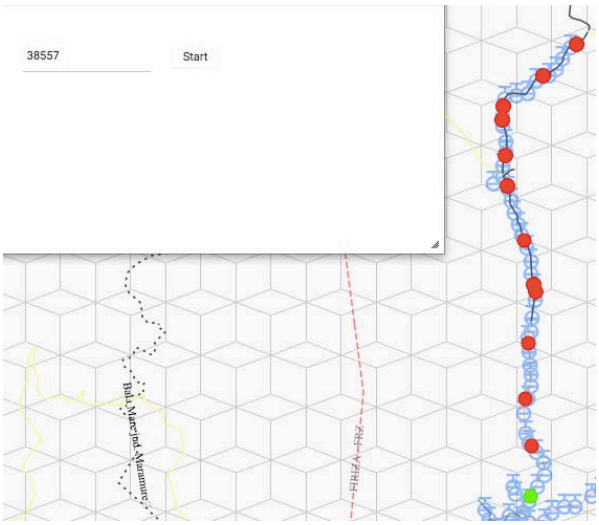


Fig. 9. Example of the Valves that Needs to be Closed for Node 38557.

We have to make a note here, related to "pipe_reference" table. The "pipe_reference" table has to be kept in synchronization with the "pipe" table, which means, that every time a user makes changes, to a pipe geometry, the "pipe_reference" table has also to be updated, to keep up with the modifications.

Regarding the hydraulic analysis that can be done on the water network [8] we have to note that are applications that already do that very well. From those applications we will mention the one called EPANET[2] (software application used throughout the world to model water distribution systems) given freely by the government of USA [9] [10].

V. Result

We showed in our article how an existing open source water network model (QWAT model) can be changed in order to accommodate new requirements, but it was only one of the steps we've done to improve the QWAT model.

Our way of extending the open source model, namely QWAT, it's not the only way, and other persons embraced our ideas and extended QWAT model in a slightly different way, which brought us joy to see that our work has been appreciated [7].

Lately, another idea came to our attention, which is the possibility to use graph databases for our processes [11] [12]. We know that graph databases are suitable for recommendation engines and those can be used in our environment, namely QWAT. Before presenting our envisioned workflow, we have to outline that our relational database data is already in a friendly graph format and already uses a graph extension called pgRouting. All these will probably allow us in the future, to export our data to graph databases and to extract more information out of it. At this point and this time, we formulated the following algorithm to be used to export data to a graph database [13]:

- choose a graph database – only once.
- define case scenarios.
- specify relations or define new views to be exported.
- export the data to graph databases.
- run case scenarios to find out results.
- import results into the relational database and make use of them.

By doing the above steps in an automated fashion we will allow users to focus onto the results and thus maximizing the amount of information extracted from a relational database. We plan to test our proposed use of graph databases in the nearest future and see how much we can benefit out of it.

The work presented in the current paper shows how to find the water path in a water network, modeled in an open source database (QWAT), but there are commercial databases that also do that. One of those commercial databases is proposed by ESRI (Environmental Systems Research Institute). Regarding the performance of our model compared to ESRI's model we can't tell much because of the costs involved in installing the ESRI application.

We used our model for identifying the customers in a Romanian city called Sighetu Marmatiei located in the northern part of Romania. The city of Sighetu Marmatiei has about 20000 customers and from those customers only about 9000 are placed on the map and linked to private branches. The time to extract those city customers using our model is around 2-3 minutes which is a reasonable time.

VI. Conclusion

The current paper shows that an open source database model of a water network can be extended to accommodate new functions and those are similar to the functions from a commercial database. We achieved that purpose by adding tables, functions to an existing relational database, namely QWAT. The new tables and functions allowed us to identify the basic water flow of a water network but that flow did not take in consideration pressure zones or pumps [14]. On other hand our paper didn't go into the hydraulic analysis of the water network.

An improvement that easily can be done by any water supplying company, is to add electric valve actuators, either to every consumer or to valves within the water network and to each actuator add a Wi-Fi circuit breaker. Connecting those circuit breakers to the Internet and further to QWAT model, the water supplying companies will be able to shut on or off water to its clients, remotely [15].

In conclusion our paper presented a way of solving a problem for a water supplying company by using open source databases and their features.

[2] Internet – https://www.epa.gov

References

[1] Bogdan Vaduva, Honoriu Valean - Water pipes leak prediction in QWAT databases, ICSTCC 2021, Iasi, Romania, 2021.

[2] H. Darwen, An Introduction to Relational Database Theory, United Kingdom: Bookboon. com, 2010.

[3] Internet - https://github.com/qwat/qwat-data-model/issues/171 - 2017.

[4] Hans-Jurgen Schonig - Mastering PostgreSQL 11: Expert Techniques to Build Scalable, Reliable, and Fault-tolerant Database Applications, 2nd Edition.

[5] Internet – www.pgrouting.org – last access in September 2021.

[6] Schulz, Frank & Wagner, Dorothea & Weihe, Karsten. (1999). Dijkstra's Algorithm On-Line: An Empirical Case Study from Public Railroad Transport. Algorithm Engineering.

[7] Internet - https://github.com/benoitblanc - last accessed in October 2021.

[8] Naser Moosavian, Mohammad Reza Jaefarzadeh, "Hydraulic Analysis of Water Distribution Network Using Shuffled Complex Evolution", Journal of Fluids, vol. 2014, Article ID 979706, 12 pages, 2014. https://doi.org/10.1155/2014/979706.

[9] G. VenkataRamanaDr.aCh.V.S.S.SudheerbB.Rajasekhar - Network Analysis of Water Distribution System in Rural Areas using EPANET - Procedia Engineering, Volume 119, 2015, Pages 496-505.

[10] Rai, R. K. and Lingayat, Prashant, Analysis of Water Distribution Network Using EPANET (February 22, 2019). Proceedings of Sustainable Infrastructure Development & Management (SIDM) 2019, Available at SSRN: https://ssrn.com/abstract=3375289 or http://dx.doi.org/10.2139/ssrn.3375289.

[11] R. De Virgilio, A. Maccioni, and R. Torlone. Converting relational to graph databases. In GRADES, 2013.

[12] Konstantinos Xirogiannopoulos, Udayan Khurana, Amol Deshpande - GraphGen: Exploring Interesting Graphs in Relational Data, Proceedings of the VLDB Endowment, Volume 8, 2014-2015.

[13] Ahmad Shahzad, Frans Coenen - Automated Generation of Graphs from Relational Sources to Optimise Queries for Collaborative Filtering – 2020.

[14] Mrs. Vaidya Deepali R., Mali Sandip T. - Pressure driven approach in water distribution network analysis: A Review - The International journal of analytical and experimental modal analysis, Volume XI, Issue VII, July/2019, ISSN NO: 0886-9367.

[15] Rosiberto Gonçalves, Jesse J. M. Soares and Ricardo M. F. Lima - An IoT-Based Framework for Smart Water Supply Systems Management - Future Internet 2020, 12, 114; doi:10.3390/fi12070114.

An Integrated Reinforcement DQNN Algorithm to Detect Crime Anomaly Objects in Smart Cities

Dr. Jyothi Mandala[1]
Assistant Professor
Department of CSE, School of Engineering & Technology, CHRIST (Deemed to be University), Bengaluru, India

Pragada Akhila[2]
Assistant Professor
Department of CSE, Gayatri Vidya Parishad College of Engineering (A) Visakhapatnam, Andhra Pradesh, India

Vulapula Sridhar Reddy[3]
Assistant Professor
Department of IT, VBIT
Telangana, India

***Abstract*—In olden days it is difficult to identify the unsusceptible forces happening in the society but with the advancement of smart devices, government has started constructing smart cities with the help of IoT devices, to capture the susceptible events happening in and around the surroundings to reduce the crime rate. But, unfortunately hackers or criminals are accessing these devices to protect themselves by remotely stopping these devices. So, the society need strong security environment, this can be achieved with the usage of reinforcement algorithms, which can detect the anomaly activities. The main reason for choosing the reinforcement algorithms is it efficiently handles a sequence of decisions based on the input captured from the videos. In the proposed system, the major objective is defined as minimum identification time from each frame by defining if then decision rules. It is a sort of autonomous system, where the system tries to learn from the penalties posed on it during the training phase. The proposed system has obtained an accuracy of 98.34% and the time to encrypt the attributes is also less.**

***Keywords*—HybridFly; Advanced Encryption Standard (AES); reinforcement; anomaly detection; crime rate prediction; security attacks; RCNN**

I. Introduction

Anomalies always refer to the abnormalities or deviations that occur in regular flow. Since, all the devices in the IoT are arranged in the distributed network manner, the implementation of anomaly has its impact on the designed system in terms of root cause analysis detection, cost and threat reduction. The various kinds of mechanisms that can identify anomaly are discussed in Fig. 1.

The Visual Discovery is the trending approach for anomaly detection process, where IoT connected video surveillances is connected in network. During this process, huge amount of data are captured and it is difficult to work on those data streams with high and multi-dimensions. So, an anomaly marker system is developed, in which threshold is marked with the help of neighborhood estimation. So, an anomaly marker system is developed, in which threshold is marked with the help of neighborhood estimation.

In this existing system, the visualization process helps to form the segments which are occasional and peculiar and it also finds the correlation between different entities that are captured in the data streaming process with in the marked area of regions. The threshold values in this mechanism are dynamic in nature, which are automatically adjustable based on the application data in the marked region. But, the system suffers with unbalanced noisy data.

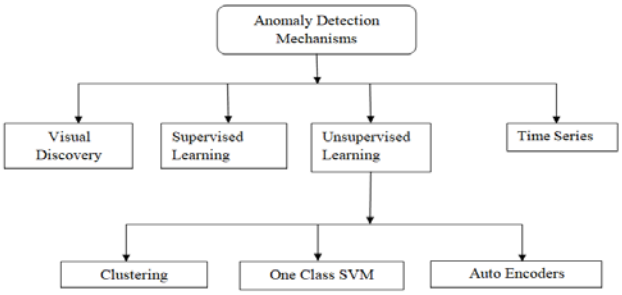


Fig. 1. AD Mechanisms for Crime Predictions.

II. Literature Survey

In [1] the primary goal of an abnormality vulnerable model is to classify the system's characteristics into usual and untruthful behaviors. To assess the tendency of occurrences, smart city managers must use anomaly observation engines to safeguard information from being compromised by flaws or breaches. This paper focused on conducting an analogy of various ML methodologies over DNN by performing a research study to identify the best method that can work with abnormalities in data. This study has chosen the issue of attack type by the rarity in IoT. This study was carried out on various attacks that may occur by the abnormalities and an entire study was made out. The researchers have also developed a grid structure that can determine the various types of interventions and foresee the possibility of happening an intrusion. This research helps in choosing the best methodology for others to perform with their own study from the proposed method depending on their issue as this study explains each method that was used layer by layer.

In [2] an innovative and expert motive that was developed to help in the growth of the resources or for the people using the advanced technologies that connects various electronic gadgets like sensors and work on together by generating a cumulative output is the development of smart cities. This amazing development was deployed in most of the places round the globe, but to make the devices derive better precise

values that helps in yielding best accurate outcomes and to make the developed machine more reliable, DL concepts were opted. This paper relies on various DL techniques and derives a cumulative study depending on the structural enhancement on the smart city information. Along with the usage of these preferred technologies, their usage, utilizations, and post enhancements were also explained.

The standard methodologies used to deploy the values that usually deviate out from the actual information, but it will be useful in some conditions, with mathematical strategies [3]. Such type of data also satisfies a few constraints so that ignoring this information paves data loss. This data can also be included to make the machine perfect. So, these papers focused on deploying a method that trains itself given a condition with all the possibilities available and learn from its experience. This method has shown better results in gaming, as stated and to promote it to next level, this model was applied in real time. This model was deployed and compared with those outcomes of ML techniques with various performance calculations and stated that their neural grid has derived more precise results with less issue and keep tracking of all the abnormalities.

With the increasing growth of intellectual areas and connecting the electronic devices the information is gathered and is shared among the united devices which lead in misleading of the information and exposure of risks is high [4]. So, rather than focusing on the raising technologies, advancing the growth of security devices or systems, analyzing them continuously and to eradicate the unnecessary conditions are also critical. For this issue, ML methods were opted to identify the rate of efficiency of the services provided within the intelligent city with security problems. The researchers have developed a self-learning system that learns from the given constraints or the past experiences from the information that supervises the overall activities that were organized. A neural grid examines the incoming activities and identifies the suspicious activities by breaking down them into chunks of parts since it reduces the difficulties and enhances the performance of the grid.

The problems that may rise due to the connectivity of electronic devices within an intellectual city are as similar as those that arise in an intelligent home rather on a smaller scale, but the risks that may occur to the associated accessories with security and for handling of the information [5]. The consequences of a security flaw are not restricted to online; it could also influence or be assisted in spatial context, for as by speech. Vulnerability assessment in this ecosystem should not rely only on estimation methods that retain the same throughout times and for all participants. The researchers have presented a system that automatically adjusts to is ecosystem whenever a new commodity is introduced that also distinguishes the necessary abnormalities. This method was induced in a reward technique to its attributes under its related ecosystem of untagged data and finds the anomalies.

With the rise in advanced technologies and their applications in real life have changed into a smarter life like form urban areas to intellectual cities [6]. This also helped in usage of various sources and services in hand to very common people as well. But even these advancements have challenges, one of which is power issue. These electronic devices needed to be associated with the commodities and should always be in a communication in share but have limited power storage devices along with the issues in networks and information protection. This paper deployed a statistical self - learning model that focuses on avoiding the DDos aggression on IoT accessories and their networks. These models were deployed based on the correlations of the various models within the layers of the network grids that also focuses on the security uncertainties.

Recently, advancements in connecting the intellectual accessories have improved with various type of grids and development [7]. One of such enhanced topologies with a harmonized structure was implemented in WMN which had derived several powerful attributes for development of the intelligent smart grids NAN with the advancing research on the devices that record the capabilities, working and usage of a smart system. With the raise of such systems, the uncertainties that may happen have dragged the scientist's attention with are related to ignorance of uncertainties. To find those abnormalities patterns, a model that finds out them locally on each site was established throughout the power distinguished model. A self-learning technology with block chain was developed to associate the local patterns of the outliers on a larger image.

The advancements in the connected system under a single ecosystem model were increased regularly and are in a great demand in multiple fields to its varied applications [8]. But increasing technologies also increases that risks with those related to various modules in between. The recent advancements in ML with named or tagged data for distinguishing the objects have helped in reaching out the problems. Detecting the attacker is one of the major problems when dealing with protection of devices in association. This paper has established a methodology focusing on this issue with a model that learns by itself based on the given attributes, constraints set and information of past. This method on working on the same data iteratively develops itself and can distinguish a risk or suspicious action when it is introduced. This research was applied to a live data with tagged names and stated that it has shown a better performance in contrasting with ML models.

The study of 19 unsupervised anomaly detection algorithms is with evaluation for multiple domains [9]. These evaluations focused on strengths and weakness of different approaches. This research is applied on 10 different datasets which is focused on global/local anomaly behaviour on real-world applications.

Prevention from cyber-attacks is very much needed for secure operation [10]. This research paper is about online anomaly detection problem which mentioned a solution for this using model free reinforcement learning. The results generated by this approach show the timely and accurate prevention by detecting the cyber-attacks which are targetting on smart grid.

III. Proposed Methodology

To reduce the crime rate in the smart cities, the proposed system has developed an anomaly detector integrated with reinforcement techniques especially in the remote places during the mid-night hours. The Fig. 2 represents different data frames associated with fire crime accident happened at a business location.

Fig. 2. Fire Accident Captured in Video Surveillance.

Proposed system works on the dataset known as "UCF-Crime Anomaly", which has multi cases video captured records of different crime incidents occurred in various locations in India. The major goal of this system is to identify the criminals and also sends notification to the nearest police station about the incident. The reinforcement technique helps the model in identifying the multiple objects simultaneously and generates a sequence of actions by generating the dynamic rules based on the captured entities. The overall experimental setup is illustrated in Fig. 3.

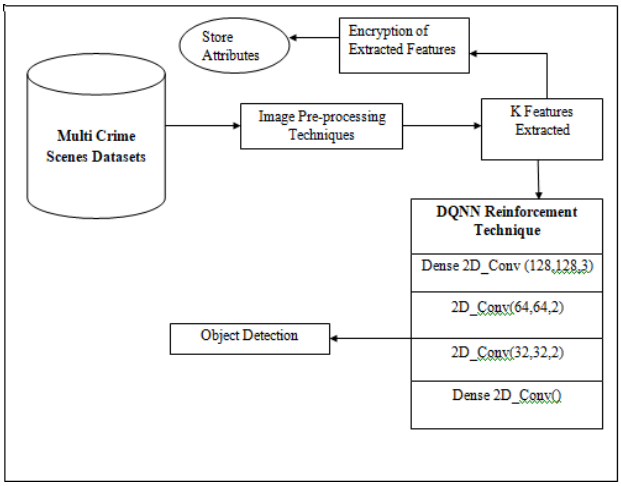


Fig. 3. Object Detection Process in Crime Anomaly Detection.

The proposed model, starts processing the objects and persons captured by the video by finding the facial land marks for the persons and RCNN for the objects. The proposed model uses traditional processing techniques like brightness and geometric transformation.

In brightness transformation, it implements histogram equalization to enhance the quality of the pixels it uses cumulative distributive frequency and produces normalized histogram. The reason for selecting histogram equalizer is it is efficient in dealing non-monotonic and non-linear structures. The crime scene contains more number of dissimilar and complex structured figures. So histogram equalizer is the apt brightness technique. In geometric transformations, the proposed algorithm majorly focuses on interpolation based on Grey scale factors. The change of direction or enhancements to the images produce new co-ordinates, these values may not be accommodated. The system needs interpolation mechanism to fit these new co-ordinates system. The proposed system implements linear interpolation technique, in which the neighboring pixel values are examined and considers the value with maximum brightness function as output.

Then for the person identified images are fine tuned to remove the noisy data during the capturing process and extract the important features. The proposed system uses the concept of dynamic fly to select the adaptive features and extract the needed features. The concept of genetic algorithm is opposite to the correlation because the attractiveness property states that distance have inverse impact on the attraction. The extracted features are encoded using Hybrid Fly algorithm to protect the information for further mishandling by the attackers or hackers.

Algorithm: Pseudocode for Attribute Encryption using Hybrid Fly Algorithm:

Begin

1. Define an objective function of K-extracted dimensions
2. Set up initial dimensions as firefly population, k
3. Determine the intensity for each firefly group based on the threshold
4. for each i 0 to k
5. for each j 0 to i

i. if (F[i]>F[j])

Then update the firefly to next fold

ii. Else

Calculate attractiveness distance based on the threshold

iii. best_feature[i] Update rank value for each feature by finding the best score

6. new_feature AES(best_feature[i],key=256)

End

In the proposed system, model free Deep Q-Learning Neural Network(DQNN) technique is implemented by defining an action value function in terms of current state and action to be performed on the current state by the agent based on the rules generated by the AO* algorithm. The major goal of this NN model is to maximize the rewards in every iteration

so that the time to detect the object decreases. The reward function is defined as shown in (1).

DQNN(S_New, A_New)←R(S_New,A_New)+α* max(S_old,A_New) (1)

In every iteration DQNN, updates the table associated with state and action. The states are passed as input to the neural network, which is designed as the auto encoder. All the Q-values are obtained as continuous output. The "tanh" activation helps the regressor in predicting the output variable. The output of the objection is represented in Fig. 4.

Fig. 4. Person Identification in Crime Scene.

For object detection, the CNN creates the feature map based on the selective search to create new regions of interests in the network. The visual attention mechanism helps the system to consider the weighted average mechanism, to obtain the new vectors.

IV. Result and Discussion

The proposed system to prove its efficiency, it has performed a comparative study on the different existing mechanisms and is illustrated in Fig. 5.

Fig. 5 represents different algorithms compared on the x-axis and accuracy percentages on the y-axis. There is a clear evidence that proposed model has exhibited best accuracy among all the others. The model has also obtained recall and precision values in terms of identifying the true positive and false negative labels. The model also discusses the comparison based on the encryption time of different security algorithms and the proposed algorithm and is illustrated in Table I.

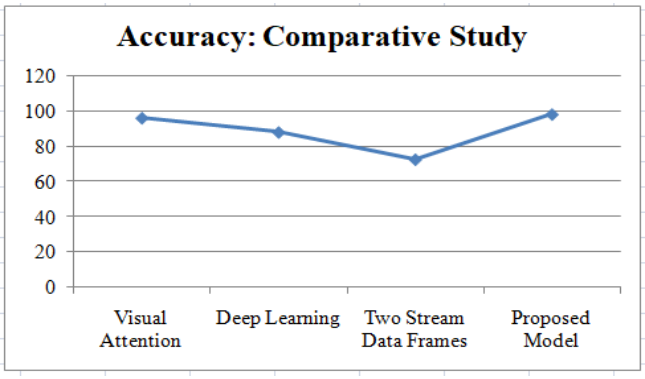


Fig. 5. Fire Comparative Study on Accuracy in Predicting the Object Detection.

TABLE I. File Encryption Time

S.No	Input Size in MB	RSA	DES	Proposed
1	25	0.485	0.41	0.334
2	50	0.535	0.443	0.352
3	75	0.563	0.499	0.431
4	100	0.625	0.542	0.46

In Fig. 6, X-axis represents size of the file in MB as mentioned in Table I and Y-axis represents time in nano seconds.

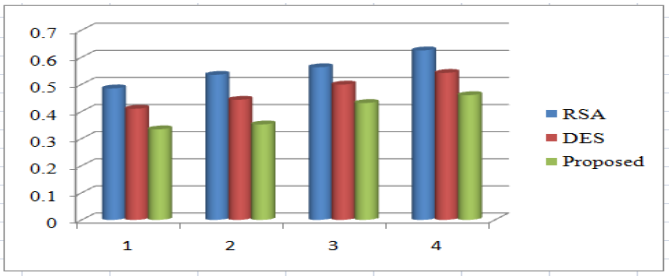


Fig. 6. Encryption Time- Comparative Study.

When compared to RSA and DES algorithms, proposed algorithm has got less time to encrypt the file and the encryption time increases when the file size increases. This description is illustrated in Fig. 6.

V. Conclusion

With the reinforcement technique it is observe that high precision and accuracy values are obtained with very short training of the system. One of the advantages of the proposed system is assigning ranks to the peculiar objects that has been captured and also reduces the ranking losses that incur due to the miss-classification of the objects. The background subtraction process helps the system to improve the quality of the software and integrated visual attention mechanism can also works fine in identification of moving objects. The involvement of the computer vision has enhanced the quality and training phase automatically without any human intervention. The proposed system has implemented model free technique because it consumes less space, since it does not involve any storage of states and actions but it applies brute forces techniques. So, the knowledge updation is time consuming task. In the future work, research can be extended based on the time lines and historical data that happened frequently in the span of time which involves model based techniques.

References

[1] Reddy, D. K., Behera, H. S., Nayak, J., Vijayakumar, P., Naik, B., & Singh, P. K. (2020). Deep neural network based anomaly detection in Internet of Things network traffic tracking for the applications of future smart cities. Transactions on Emerging Telecommunications Technologies, 32(7). https://doi.org/10.1002/ett.4121.

[2] Bhattacharya, S., Somayaji, S. R. K., Gadekallu, T. R., Alazab, M., &Maddikunta, P. K. R. (2020). A review on deep learning for future smart cities. Internet Technology Letters. https://doi.org/10.1002/itl2.187.

[3] Zhou, K., Wang, W., Hu, T., & Deng, K. (2021). Application of Improved Asynchronous Advantage Actor Critic Reinforcement

Learning Model on Anomaly Detection. Entropy, 23(3), 274. https://doi.org/10.3390/e23030274.

[4] Zhang, Mengqi et al. 'Machine Learning Techniques Based on Security Management in Smart Cities Using Robots'. 1 Jan. 2021 : 891 – 902.

[5] R. Heartfield, G. Loukas, A. Bezemskij and E. Panaousis, "Self-Configurable Cyber-Physical Intrusion Detection for Smart Homes Using Reinforcement Learning," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1720-1735, 2021, doi: 10.1109/TIFS.2020.3042049.

[6] Ashraf, J., Keshk, M., Moustafa, N., Abdel-Basset, M., Khurshid, H., Bakhshi, A. D., & Mostafa, R. R. (2021). IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities. Sustainable Cities and Society, 72, 103041. https://doi.org/10.1016/j.scs.2021.103041.

[7] Belhadi, A., Djenouri, Y., Srivastava, G., Jolfaei, A., & Lin, J. C.-W. (2021). Privacy reinforcement learning for faults detection in the smart grid. Ad Hoc Networks, 119, 102541. https://doi.org/10.1016/j.adhoc.2021.102541.

[8] Q. -V. Dang and T. -H. Vo, "Studying the Reinforcement Learning techniques for the problem of intrusion detection," 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), 2021, pp. 87-91, doi: 10.1109/ICAIBD51990.2021.9459006.

[9] Goldstein, M. Uchida, S. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data." PLoS ONE 2016, 11, e0152173, doi: https://doi.org/10.1371/journal.pone.0152173.

[10] Kurt, M.N., Ogundijo, O.; Li, C., Wang, X. "Online Cyber-Attack Detection in Smart Grid: A Reinforcement Learning Approach." IEEE Trans. Smart Grid 2018, 10, 5174–5185. doi: https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.1109%2FTSG.2018.2878570&v=17393103.

Monitoring the Growth of Tomatoes in Real Time with Deep Learning-based Image Segmentation

Sigit Widiyanto[1], Dheo Prasetyo Nugroho[2], Ady Daryanto[3], Moh Yunus[4], Dini Tri Wardani[5]
Department of Computer Science, Gunadarma University, Jakarta, Indonesia[1, 2]
Department of Agrotechnology, Gunadarma University, Jakarta, Indonesia[3]
Department of Information System, Gunadarma University, Jakarta, Indonesia[4, 5]

***Abstract*—Increasing agricultural productivity such as tomatoes needs to be increased, considering the consumption growth reaches 6.34% per year. Efforts to increase productivity can be made through several methods, such as counting and predicting the time of fruit to be harvested. This information is a a visual problem, so computer vision should solve it as an automation method in the industry world. With this information, the farmer can monitor the tomato fruit growth. The proposed method is a framework that has been implemented in real-time processing. To obtain growth information of tomatoes, the tomato area can be used as a region of interest (ROI) every week or another scheduled time. As the challenge of this research, this ROI can be extracted using segmentation analysis. The segmentation method used is Mask Region-Convolutional Network (R-CNN) with ResNet101 architecture. The accuracy of this method is obtained from the similarity value between the proposed method and the ground truth used, namely 97.34% using the Dice Coefficient and 94.83% using the Jaccard Coefficient. This result indicates that the method can extract the ROI information with high accuracy. So, the result can be used as a reference for the farmer to treat each tomato plant.**

***Keywords*—Deep learning; Mask R-CNN; segmentation; tomato; growth**

I. Introduction

Tomato consumption from 2016 - 2020 increased from 883.23 thousand tons to 1,084.99 thousand tons, or an increase of 6.34% from 2019 [1]. Despite an increase in production, tomato fruit has perishable properties and is classified as a climacteric fruit, where the peak of respiration and ethylene production occurs at the beginning of fruit ripening [2]. This can cause tomatoes to be easily damaged, causing a reduced supply of tomatoes at the consumer level and food insecurity. In addition, tomato production also faces future challenges in scarcity of water resources, soil salinization, and other abiotic stresses. The growth and development of tomato plants are influenced by several factors, including temperature, humidity, and altitude. If the environment does not support the growth and development of tomato plants, it will affect the productivity of tomato plants.

Increasing agricultural productivity through several methods, such as fruit counting and prediction of fruit to be harvested and early detection of environmental diseases and weeds, can be done at this time. Solutions for utilizing the latest technologies such as deep learning, the internet of things, and robotics are very effective and efficient for plant management [3]. Smart farming systems can reduce waste, increase productivity, and allow the management of more resources through remote sensing [4]. Remote monitoring through intelligent farming systems allows production yields to increase because farmers need a lot of time to solve pests, soil conditions, rain, or weeds, which can now be done through remote sensing and automation. A survey that has been conducted [5] has analyzed various articles related to deep learning technologies, and each work is compared with existing techniques. Deep learning methods have been widely used in various fields of agriculture, such as plant disease detection, crop classification, weed identification, fruit counting, land classification, obstacle detection, image translation, weather forecasting, yield prediction, and animal behavior classification [6].

Deep learning technology that has been applied to the horticultural domain for variety recognition, yield estimation, quality detection, growth, surveillance, and other detection, where this review aims to assist researchers and provide guidance to them in the use of deep learning. This guide is to understand the strengths and weaknesses that may occur when implementing deep learning. From the results of a review that has been carried out [7], there is a deep learning technique for object detection that is commonly used in horticultural crops, namely Convolutional Neural Network (CNN). There are three types of object-based detection with CNN; the first is object recognition such as LeNet, AlexNet, VGGNet, GoogLeNet, and ResNet, while the second is a combination of the first method with two-stage detection to achieve improved and accelerated detection. This method includes R-CNN, Faster R-CNN, and Mask R-CNN. As for the third type, detection is for one stage, which can immediately give the results of object boundaries according to their position. Using these models, fruit yields can be estimated automatically, and plant stress levels can be detected early. An agriculture science model with the help of deep learning techniques can be used to detect leaf diseases with images of corn, peaches, grapes, potatoes, tomatoes, and strawberries [8][9]. In this case, the image processing technique used is the CNN model for plant disease detection. The tests that have been carried out give an accuracy rate of 94.29%.

The Faster R-CNN method in the classification process of apple objects can give very accurate results [10]. A citrus fruit yield mapping system has been developed using a robotics platform [11]. The results of presenting the Fast R-CNN method used to detect citrus fruits that have been taken from different conditions (distance to fruit, camera angle, and slight

variation) with the implementation of this method give better results than the human process with an Average Precision score (AP)=0.76. The blueberry fruit calculation process has been automatically developed based on ripeness and maturity classification with deep learning methods. In this case, what is used for blueberry fruit segmentation is the R-CNN model by producing an average precision for validation and test datasets is 78.3% and 71.6% below the 0.5 intersections over union (IOU) threshold with an accuracy of 90.6% and 90.4%, respectively [12].

There is another framework, namely the YOLOv3 framework, which has been modified to YOLO-tomato, which can show better results than other advanced methods so that the YOLO-tomato method is better for detecting tomatoes in real-time with complex tomato environmental conditions [13]. The deep learning architecture used in general is a semantic segmentation architecture such as Deep Neural Network (DNN) in the field of Computer Vision (CV) [14]. The most popular DNN architectures are AlexNet, GoogleNet, VGGNet, and Resnet. Implementation of semantic segmentation techniques with DNN architecture has several problems, one of which is caused by the many parameters involved or overfitting. DNN requires high-quality labeled data and large-scale data. So, an effective solution is to build large, high-quality data sets that are difficult to achieve. Semantic detection in real-time is very important because it can be useful in autonomous systems and robotic interactions. Therefore, several new methods are adopted to increase computational efficiency, accuracy, and background noise. The semantic segmentation architecture is a fully convolutional network (FCN), ParseNet, deconvolution network, U-Net, feature pyramid network (FPN), and Mask R-CNN. The survey results show that there are many scopes of improvement in terms of accuracy, speed, complexity, and overfitting problems, so that new methods or combinations of semantic segmentation architecture are needed to increase efficiency and accuracy. In one extension of this model, [15] proposed a Mask R-CNN for object instance segmentation, which beats all previous benchmarks on many COCO challenges. This model efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask R-CNN is essentially a Faster RCNN with three output branches. The first computes the bounding box coordinates, the second computes the associated classes, and the third computes the binary mask to segment the object. The Mask R-CNN loss function combines the losses of the bounding box coordinates, the predicted class, and the segmentation mask and trains jointly. In this research, Mask R-CNN is used to segment the tomato area. The segmentation is more challenging than object detection or classification because all subject areas should be identified, and the pixel that false detection would have become an error of the method. Mask R-CNN more comprehensively detecting the region rather than R-CNN with masking method.

II. METHODS

This research proposes a framework consisting of several stages, namely image data collection, image data labeling, data set separation for training and model validation, then continued by building a model and implementing segmentation using Mask R-CNN and ending with similarity testing between segmentation results using the deep learning method with manual segmentation results. In detail, the stages in the proposed framework can be seen in Fig. 1.

A. Data Collecting

All images were obtained from the primary data captured from tomatoes in a greenhouse. The data was collected from tomatoes is planted in the period from July to September 2021. The commercial tomato varieties planted and used as the data are Tora, Servo, and Tatyana. These varieties can live in the lowlands. The image was taken using a mobile phone camera and mobile robot camera with resolution 2592×1944 (5MP). The image contains multiple objects inside. The total dataset collected is about 600 images (includes 210 images Tora, 185 Servo, and 205 Tatyana). From these images, there are 2476 objects of tomato that can be identified visually. In this study, no preprocessing stage was applied to the image to be tested. Only train data needs to be labeled to support the Mask R-CNN method, and this is related to system development in actual conditions, where lighting and image transformation obtained are unavoidable. An example of an image data set can be seen in Fig. 2.

B. Image Labeling

Image segmentation aims to recognize and understand what's in the image at the pixel level. Every pixel in an image belongs to a single class, as opposed to object detection, where the bounding boxes of objects can overlap.

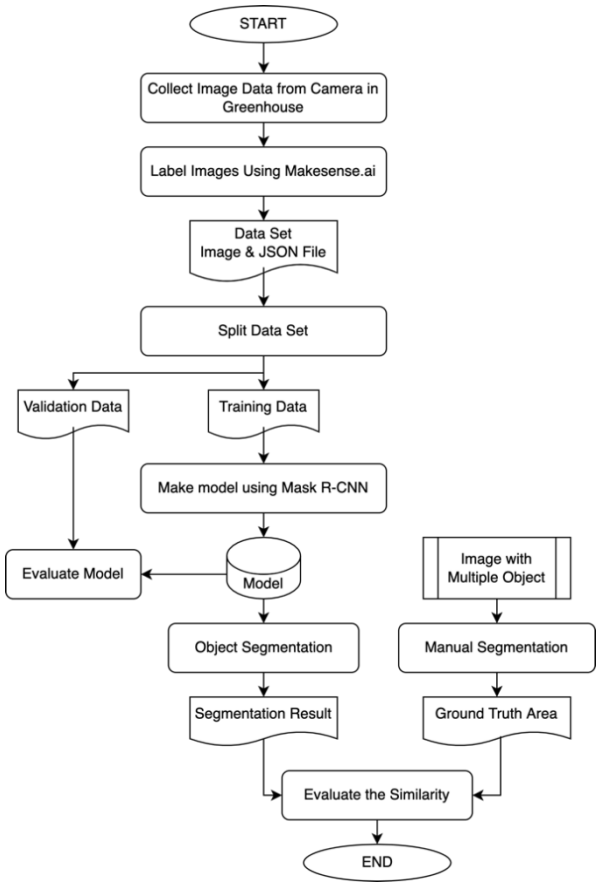


Fig. 1. Proposed Framework.

Fig. 2. Sample of Image Data Sets.

Fig. 3. Image Labeling Process using Maksense.ai.

Therefore, one preprocessing stage is needed in image segmentation, namely, image labeling. As shown in Fig. 3, an example image labeling, there are two tomatoes, so the model should be identifying all the pixels belonging to each tomato. That's where instance segmentation comes in. With instance segmentation, multiple disparate regions can belong to a single instance of a class. Each tomato in the example is annotated as an instance of the tomato class and is also given an ID such as "Tomato 1," "Tomato 2," and so forth. Therefore, the model can identify all the pixels belonging to an individual tomato even if the instance contains multiple regions.

The labeling process gives a point on the edge of the object. These points will be connected to form a polygon line. The coordinates of these points will be stored in a JSON file which must be called when executing the training model method.

C. Modelling and Testing using Mask R-CNN

Mask R-CNN is a deep learning framework that can detect objects in an image that generates a segmentation mask for each instance or commonly called instance segmentation [15]. This method runs on Faster R-CNN [16] (can be seen in Fig. 4), so that in performing mask detection, the R-CNN can be divided into three parts, namely: (i) feature extraction network, (ii) region-proposal network, and (iii) instance detection and segmentation networks.

1) Feature extraction: Mask R-CNN applies multiple backbone architecture. Some of the backbones used for Mask R-CNN are ResNet, ResNet, and FPN. In the Region Proposal Network (RPN) process, a Region of Interest (RoI) will be generated through an alignment process which will then be input for the instance detection and segmentation networks.

Mask R-CNN uses a combination of ResNet101 architecture and FPN (Feature Pyramid Network) to generate RoI features when feature extraction is performed. FPN is a basic component in the recognition system to detect objects at different scales using the same image.

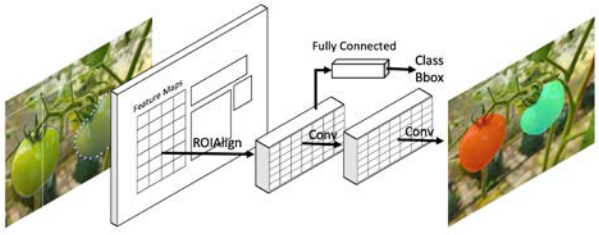


Fig. 4. Mask R-CNN Architecture.

FPN uses a variety of feature maps to produce higher-quality information. FPN is a feature extractor designed using the pyramid concept but is superior in speed and accuracy. FPN processes information in two ways, namely bottom-up (bottom-up) and top-down (top-down).

Bottom-up data processing extracts features using ResNet, in which the spatial dimension of each layer decreases while the semantic value increases. Top-down processing increases the resolution of the semantic layer, but the location of objects is not precise. FPN adds lateral connections between reconstructed layers and a corresponding feature map to help detectors better predict locations. Lateral connection is the convolution and addition operation between the two corresponding levels of the two pathways. FPN surpasses the single ConvNet because it retains semantic features at various resolutions.

From ResNet101 four feature maps were extracted (layer-1, layer-2, layer-3 and layer-4). An approach called the top-bottom pathway is used to produce the final feature map. The top-bottom pathway approach starts from the top feature map (x/32, y/32, 256) and starts down to a larger feature map by applying the up-sample operation. A 1 x 1 convolution was performed to reduce the number of channels to 256 before sampling and then added elements to the up-sample output from the previous iteration. All up-sample outputs were applied to a 3 x 3 convolution layer to produce the last four feature maps, while the fifth feature map was generated from the max-pooling operation.

2) Region proposal network: Each feature map generated in the feature extraction process will go through a 3 x 3 convolution layer. But before that, the feature map is scanned using anchor boxes with various scales and ratios. The resulting output is then forwarded to two branches, one relating to the objectivity or confidence score and the other to the bounding box regressor. A confidence score can be obtained by calculating the IoU (intersection over union) between the bounding box and the ground truth. IoU is obtained by dividing the overlapping area by the combined area of the ground truth and bounding box. The confidence score ranges from 0 to 1. The greater the Confidence score means, the higher the system confidence that the object contained in the bounding box is the object to be detected.

3) Instance detection and semantic segmentation: The segmentation instance process is carried out using a fully connected network that takes RoI as input to detect the presence of objects, bounding boxes, class labels, and confidence values. A fully Convolutional Network (FCN) is

used to perform semantic segmentation on the image by predicting the semantic class of each pixel in the bounding box. This causes each instance to display a different color according to its bounding box.

D. Manual Segmentation

A manual segmentation process is needed to generate Ground Truth data. This data is used to compare whether the segmentation process using the Mask R-CNN method is successful in approaching expert judgment. Manual segmentation is carried out using an image processing application to remove or change the background with black color or a value of 0 (see Fig. 5.b) and then convert to a binary image (see Fig. 5.c).

E. Similarity Evaluation using Dice and Jaccard Coefficient

Dice and Jaccard coefficient have been utilized to measure the segmentation accuracy. Rockefeller used these coefficients in the segmentation in cucumber seed [17]. The Jaccard coefficient measures similarity between sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. Sample sets are obtained by mask (black, white) image on every region from segmentation result and ground truth image, to be detailed see Equation 1. Samples sets contain, i.e., the region of interest from algorithm segmentation result and manual segmentation by an expert judgment as ground truth data. The Jaccard coefficient also called Intersection over Union (IoU), is used in the Mask R-CNN evaluation model.

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

Where X is the set of every region from segmentation and Y is the set of every region from the ground truth image, which has a true value (1 in binary image).

Then, almost like Jaccard, Dice distance is also utilized to measure the similarity, but it has different properties, Equation 2. Dice are used as an optimist measurement which gives the double weighting value on the true positive value (intersection of two regions).

$$D(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

Where 2 |X ∩ Y | is the double weighting of number value which is intersected on two data sets. While |X| and |Y | are the number of data sets which are having true value (1).

(a) (b) (c)

Fig. 5. Manual Segmentation Process. (a) Original Image, (b) Remove or Change Backgorund, (c) Convert to Binary Image.

III. Results

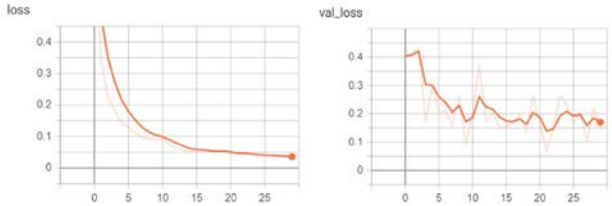
A. Model Evaluation

To conduct training and validation, 530 data sets were used consisting of various types of age and tomatoes. Of all the data sets, 70% are used as training data sets and 30% for testing data sets. The evaluation of the results of the model form can be seen in Fig. 6. Fig. 6.a shows the training loss, which goes towards 0 as the epoch value increases, while Fig. 6.b shows the validation loss, which is 0, although it is not as good as the training loss. However, these two graphs are convergent, showing that the model formed is fit, meaning that the model can predict the data set outside the training data.

B. Segmentation Result

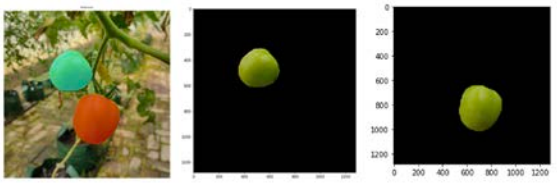
After the model is formed and thorough evaluation, it is declared fit. Then the model is then used to perform image testing segmentation. Mask R-CNN can perform segmentation by providing a mask, and at the same time assigning class labels as class numbering "Tomato-1" and "Tomato-2" so that objects can be separated. In addition, this method can also provide a bounding box accompanied by a confident value from the results of the class classification. In Fig. 7.a can be seen the segmentation results, where the confident value can also be seen. Fig. 7.b and Fig. 7.c are the results of extracting fruit areas "Tomato-1" and "Tomato-2".

The next process is to calculate the number of pixels from the area of each tomato. The first process is to convert the image into a binary image or a value of 0 (black) and 1 (white). Furthermore, the total number 1 in the image can show the number of pixels of each object. Fig. 8 shows the result of converting an image into a binary image.



(a) (b)

Fig. 6. Model Evaluation. (a) Training Loss, (b) Validation Loss.



(a) (b) (c)

Fig. 7. Segmentation Result. (a) Image Segmented with Mask, (b) Tomato-1 Area, (c) Tomato-2 Area.

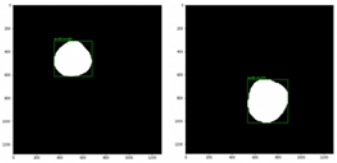


Fig. 8. The Result of Binary Conversion Process.

C. Segmentation Evaluation

Segmentation results are evaluated using the Dice and Jaccard coefficient. This method is used to assess the similarity between the proposed method result and the ground truth obtained through manual segmentation. In Fig. 9, the segmentation results using the proposed method do not have smooth or imperfect edges like manual segmentation results. Thus, this deviation makes the difference in the number of pixels between the two. The evaluation results show that the first image has a similarity value based on the Dice coefficient of 97.90. This value shows 97.90% of the pixels in the segmentation area with the proposed method are the same as the manual segmentation results, likewise, with the Jaccard coefficient, which shows a value of 95.90. Jaccard always gives less value than Dice.

In Fig. 10, the graph of the similarity evaluation results on 30 data sets used as testing data can be seen. The graph shows that the average evaluation result is above 90%. Only one image can be segmented with a similarity value of about 87%.

The average evaluation results for 30 images with the Dice coefficient is 97.34% and using the Jaccard Coefficient is 94.83%. These results indicate that the Mask R-CNN can segment well on the image of tomatoes from various types of tomatoes and tomato colors. Thus, the calculation results in the form of some pixels from the tomato area can be declared valid so that the research can be continued with analyzing the development of tomatoes.

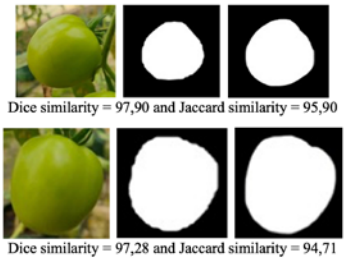


Fig. 9. Segmentation Evaluation. Left: Original Image, Center: Proposed Method Result, Right: Manual Segmentation Result.

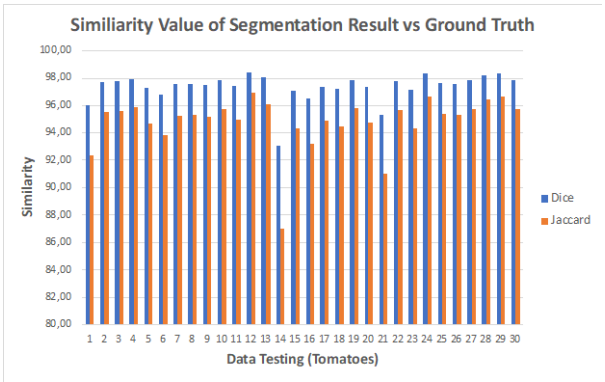


Fig. 10. Similarity Value of Segmentation Result vs Ground Truth.

IV. Discussion

One of the indicators related to the increase in tomato production is the growth of tomato fruit after flowering. The camera can capture fruits formed for a week as objects that have characteristics, namely round or oval or other shapes according to the type of plant. Tomato fruit growth can be seen from the area of the fruit. Although the picture of the fruit must be taken from a consistent direction, for example, in the first week after fruit formation is taken using a camera from an angle of 180 degree or straight from the top of the tree, then in the following week, it is taken from above as well. Likewise, if taken from the side of 90 degree.

In Fig. 11, you can see a graph of the increase in the ten observed fruits (not all of them are shown due to the clarity of the graph). Some fruits have different growth patterns. Knowing each fruit's growth pattern or average can be recommended regarding environmental engineering and proper fertigation. Thus, fruit growth can be evenly distributed, and harvest targets can be achieved.

Tomato fruit development during four weeks of the observation showed good development and was detected by increasing the number of pixels area. Tomato fruit develops every week marked by increasing fruit dimensions and fruit color that changes from light green to a red tinge and finally red. The time required by tomatoes from flowering to harvesting was 5-6 weeks [18]. In Fig. 12 can be seen the average of tomato fruit growth in a week, which shows how each tomato grows and require special handling for such fruit like tomato 5 and 6. According to [19], the weight of small vegetable tomatoes has an average weight of < 50 g per fruit, medium size with a weight range per fruit of 50 - 70 g, and large size with a weight per fruit of > 70 g. This size unit must be converted in some pixels by volume calibration, so the unit can match the image processing standard.

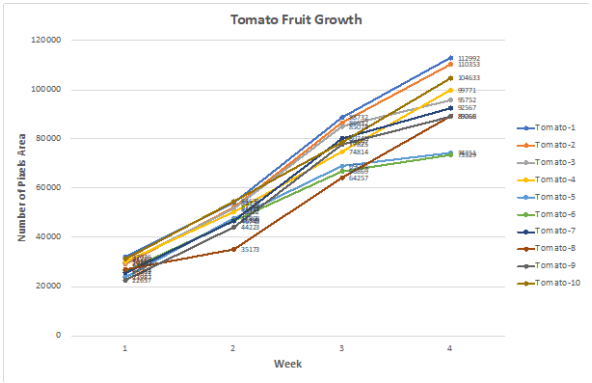


Fig. 11. Tomato Fruit Growth in 4 weeks after Fruit Formation

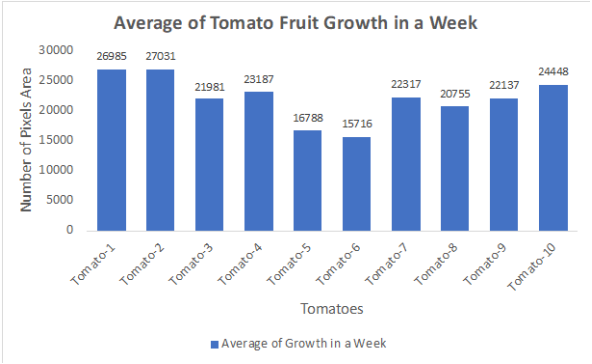


Fig. 12. Tomato Fruit Growth in 4 weeks after Fruit Formation.

V. Conclusion and Future Work

The Mask R-CNN method can perform properly and has accuracy as indicated by the similarity value measured by the Dice Coefficient and Jaccard Coefficient with an average value of 97.34% and 94.83%, respectively. This similarity value indicates that this method can be used to find the Region of Interest area of tomato objects so that it can be used to measure tomato growth. The entire system has worked in real-time with various problems such as lighting, morphology, and transforming fruit shapes and other objects. The real-time segmentation method is difficult to implement using K-Means, SVM, or Neural Network methods. For future work, it is necessary to calibrate the fruit weight unit (kg) into the volume, while in terms of the proposed method, a volume estimation process must be added. Thus, higher accuracy will be obtained regarding the analysis of tomato fruit growth based on the image.

Acknowledgment

This research was funded through the Higher Education Applied Research grant program under the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia with contract numbers 309/E4.1/AK.04.PT/2021 and 09.17/LP/UG/VII/2021.

References

[1] Food and Agriculture Organization of United Nations, Summary Report, Joint Fao/Who Meeting on Pesticide Residues, October 2021.

[2] Chen Y, et.al., “Ethylene receptors and related proteins in climacteric and non-climacteric fruits,” Plant Sci, vol. 276, pp. 63–72, 2018.

[3] Kavitha, “ Deep Learning for Smart Agriculture,” Int. Journal of Engineering Research & Technology, vol. 9, no. 5, 2021.

[4] M. Vengateshwaran, N. Sumithra, S. P. Rani, and B. Pravalika, “A Deep Learner based Smart Precision Agriculture System using Machine Learning Techniques,” Int. Journal of Engineering Research & Technology, vol. 9, no. 5, 2021.

[5] C. Ren, D. K. Kim, and D. Jeong, “A Survey of Deep Learning in Agriculture: Techniques and Their Applications,” Journal of Information Processing Systems J Inf Process Syst, vol. 16, no. 5, pp.1015-1033, October 2020.

[6] N. Zhu, et.al., “Deep learning for smart agriculture: Concepts, tools, applications, and opportunities,” Int. Journal of Agricultural and Biological Engineering, vol. 11, no. 4, pp. 21-28, 2018.

[7] B. Yang and Y. Xu, “Applications of deep-learning approaches in horticultural research: a review,” Horticulture Research, vol. 8, no.123, 2018.

[8] Md. Tariqul Islam, “Plant Disease Detection using CNN Model and Image Processing” in in International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 10, October 2020.

[9] S. Widiyanto, R. Fitrianto, and D. T. Wardani, “Implementation of Convolutional Neural Network Method for Classification of Diseases in Tomato Leaves,” Fourth International Conference on Informatics and Computing (ICIC), pp. 1-5, 2019.

[10] H. Shaikh, Y. Wagh, S. Shinde, and S. M. Patil, “ Classification of Affected Fruits using Machine Learning,” Int. Journal of Engineering Research & Technology (IJERT), vol. 9, no. 3, 2020.

[11] J. P. M. Galdames, C. E. Milhor, and M. Becker, “Citrus fruit detection using Faster R-CNN algorithm under real outdoor conditions,” Proceedings of the 14th International Conference on Precision Agriculture, June 24 – June 27 2018.

[12] X. Ni, C. Li, H. Jiang, and F. Takeda, “Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield,” Horticulture Research, vol. 7, no. 110, 2020.

[13] M. O. Lawal, “Tomato detection based on modified YOLOv3 framework,” Scientific Reports, vol. 11. no. 1, 2021.

[14] Tapasvi, N. U. Kumar, and E. Gnanamanoharan, “A Survey on Semantic Segmentation using Deep Learning Techniques,” Int. Journal of Engineering Research & Technology (IJERT) vol. 9, no. 5, 2021.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” Proceedings of the IEEE international conference on computer vision, pp. 2961–2969, 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, pp. 91–99, 2015.

[17] R. T. Rockafellar, Variational Analysis, Springer-Verlag, 2005, Number ISBN 3-540-62772-3.

[18] A. Daryanto, M. R. A. Istiqlal, Kalsum, and R. Kurniasih, “Penampilan karakter hortikultura beberapa varietas tomat hibrida di rumah kaca dataran rendah,” Journal Agron Indonesia, vol. 48, no. 2, pp. 157–164, 2020.

[19] M. Syukur M, E. Helfi, and R. Hermanto, Bertanam Tomat di Musim Hujan, Jakarta (ID): Penebar Swadaya, 2015.

Personalized Recommender System for Arabic News on Twitter

Bashaier Almotairi, Mayada Alrige, Salha Abdullah
Department of Information Systems
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Reading online news is the most popular way to read articles from news sources worldwide. Nowadays, we have observed a mass increase of information that is shared through social media and specially news. Many researchers have proposed different techniques that focus on providing recommendations to news articles, but most of these researches focused on presenting solution for English text. This research aimed to develop a personalized news recommender system that can be used by Arabic newsreaders; to display news articles based on readers' interests instead of presenting them only in order of their occurrence. To develop the system we have created an Arabic dataset of tweets and a set of Arabic news articles to serve as the source of recommendations. Then we have used CAMeL tools for Arabic natural language processing to preprocess the collected data. After that, we have built a hybrid recommender system through combining two filtering approaches: First, using a content-based filtering approach to consider the user's profile to recommend news articles to the user. Second, using collaborative filtering approach to consider the article's popularity with the support of Twitter. The system's performance was evaluated using two evaluation metrics. We have conducted a user experimental study of 25 respondents to perform an assessment to get the users' feedbacks. Also, we have used Mean Absolute Error (MAE) metrics as another way to evaluate the system accuracy. Based on evaluation results we found that hybrid recommender systems would recommend more relevant articles to users compared to the other two types of recommender system.**

Keywords*—Hybrid recommender system; online social network; Arabic news recommendation

I. Introduction

The magnitude of data generated on the Internet by different business communities, administrations, industrial sectors, scientific research and general data has increased immeasurably [1]. The research of [2] has reported that the world produces about 2.5 quintillion bytes of internet data daily, and almost 90% of this data is unstructured. Similarly, on social media sites, there is a massive bulk of un-related data and other non-authentic news which cause the users several issues related to privacy, psychological problems and many more. Organizations and some users need different new methods to process the extensive, massive data, on social media, into meaningful comprehensions. A report by [3] stated that more than half of the world's population is presently using the Internet and more than 250 million population use Arabic.

Understanding Arabic words and their vowels and consonants is still lacking in machine learning and getting the proper theme of the words written in Arabic. This situation is due to the significant number of Arabic languages subcategories with different accents and meanings, which cause the listener to misconception the idea [4]. The text categorization to assign a category to Arabic language content on the Internet and in social media lags behind the actual outcomes, since there are no firm rules for Arabic language understanding by machine learning systems [5]. Part of these massive Arabic data on social media and other platforms is news. The relation between journalists and audience is shifting towards the use of new technologies and algorithms in digital news. Several platforms display your content in this digital world as YouTube, Instagram, Facebook and different news sites and journals. Different other platforms are being used in news field, but not all of these platforms are authentic and reliable for the news and data collected.

This research's motivation is the need for a valuable tool or system to help readers in such information overload situations. Thus, the news at any site can be displayed in an order based on the reader's interests, instead of using the same order of presentation for all readers based on a publisher's opinion. The motivation for the solution comes from [1], where the researcher has developed a news recommender system with the help of micro-blogging services. News articles are sorted based on the popularity of the article, which is identified using tweets from the public timeline of Twitter. The researcher has also built the user profile based on the user's interests, so the news articles are sorted by matching the user profile's characteristics. The research of [6] has developed a personalized recommender system for calculating the closeness between Twitter users in a social circle. The system suggests topics or interests that users may have to analyze social information. This solution and other similar propositions have already been presented English medium, but not in the Arabic language.

The research aims at developing an Arabic recommender system to display news or articles to each reader based on their interests instead of presenting them only in the order of their occurrence using Twitter.

The research is organized as follows. Section 2 gives a comprehensive overview of recommender systems and discusses other researchers' work in this field. Section 3 discusses the process of data gathering and preprocessing to move to Section 4, where it describes the design and implementation of our Arabic news recommender system.

Section 5 discusses the evaluation of the proposed system's result using one of recommendation systems evaluation metrics. Finally, Section 6 summarizes the research and suggests future works.

II. Literature Review

A. Recommendation Systems

The rapid increase in the amount of accessible digital data and the vast numbers of Internet users has generated a possible knowledge overload problem that sometimes hinders timely accessibility to online information or needed items. Data gathering systems, including Google, Altavista and DevilFinder, have slightly fixed the issue, but the personalization and prioritization of the information have been missing. It has generated a much higher demand for recommendation systems compared to older times. These systems are considered as knowledge retrieval systems used to tackle the issue of overloading information [7]. This is done via filtering essential fragments of information from a substantial amount of dynamically produced information as per user interest, actions or preferences noticed regarding specific items [8]. Using information from user profiles, recommendation systems possess the potential to predict if a user will choose something or not [9].

While several strategies have been established in the past, research is still ongoing as it is mainly used in many apps that configure suggestions and deal with overloading information [10]. These systems benefit both service providers and users [11], they substantially reduce transaction costs for detecting and choosing products in an online retail environment [12]. These systems have been shown to enhance the reliability of the decision-making process. When setting up e-commerce, the recommended systems boost sales since they successfully sell more products [11]. Throughout libraries science, these systems assist consumers by encouraging them to step past index searches. Recommending programs assist consumers of research libraries by encouraging them to pass beyond catalogue searches. Consequently, the need to utilize precise and effective recommendations strategies inside a program that can offer reliable and relevant advice to consumers could not be overemphasized [9].

B. Recommendation Systems' Approaches

Using systems of recommendation that are both reliable and efficient is necessary for a program that aims to offer excellent and practical guidance to its user, demonstrating the value of knowing the characteristics and potential of different domain approaches. Classifying recommendation system is usually based on the rating estimation [10]. Generally, there are three approaches to recommendation systems:

- Collaborative filtering approach.
- Content-Based filtering approach.
- Hybrid Filtering approach.

Collaborative filtering is a domain-independent predictive strategy where the information cannot be conveniently and accurately classified via metadata, e.g., music and movies. This approach operates by constructing a database with user preferences for specific items. Afterwards, it groups the users with related preferences and interests to make recommendations by calculating how similar their profiles are [13]. Content-based filtering technique uses an algorithm which is dependent on domain and relies primarily on evaluating the properties of objects for producing predictions. It is most suitable where documents such as publications, web pages and news, have to be recommended. Throughout this method, recommendations are provided using user identities formed by using features that have been derived from the contents of those items that previously have been reviewed by the particular user, so the system will recommend the user of items that are so related to the items he/she liked before [14,15].

To understand the difference, Content-Based filtering requires information about items' features, instead of using user's interactions and feedback. Good examples are movie attributes such as actor genre, year, director or textual articles content. On the other side, Collaborative filtering doesn't require anything else except the user's historical preference on a set of items to recommend from. So it assumes that user who has agreed in the past will also tend to agree in the future [14,15].

Traditional filtering techniques have distinct weaknesses and strengths. For instance, Collaborative filtering suffers from cold start and sparsity issues, whereas content-based suffers from need and narrowness explanation. Nevertheless, hybrid solution that uses one method to render recommendations where the other fails lead to a more reliable recommendation framework [16,17]. Hybrid filtering combines various recommendation techniques to optimize the system's optimization to avoid specific difficulties and challenges with the systems of recommendation [18]. By using more than one filtering technique, the limitations can be minimized and the recommendations will be more accurate in a hybrid model [19].

C. Current Researches in Recommendation Systems

Some researchers have proposed the hybrid recommendation system technique as an excellent solution to solve or enhance their research issue. The research [20] has discussed the impact of visual information, i.e., customers' photos and put on some blogs, to predict favourite restaurants for any given user. By considering the visual information as an intermediate, the researcher suggested integrating two common recommender system approaches, collaborative filtering and content-based filtering, to show the proposed hybrid system's effectiveness with considering visual information. Another research of [21] has described a recommendation system built on a probabilistic programming language and discussed the benefits and challenges of explaining the generated recommendations to users. Using an online user survey, the research evaluated the explanations for hybrid algorithms in a set of text, visual and graph formats, which are either new designs or derived from existing hybrid recommendation systems. Moreover, the research showed that hybrid systems demonstrate better accuracy than recommendation strategies that use a single source.

A study of [22] had focused on creating an engine for a product recommendation, which can be accessed by third-party software to obtain certain product recommendations using some input data. The study's overall aim was to make novel contributions on how the user events can be gathered and processed and how they can be used by the methods of data mining to make product recommendations. The study has solved product recommendations appropriate for big data by investigating methods to process user event data.

D. News Recommendation Systems

The relation between journalists and audiences is shifting towards new technologies and algorithms in digital news [23]. Several platforms display the digital world's content as YouTube, Instagram, Facebook and different news sites and journals. Different approaches are being used in the news field, but not all platforms are authentic and reliable for the news and data collected. These news fields serve the expanded eyes of the viewers or listeners. The study of [24] showed how the news topics could be used to recommend English news articles. The researcher used supervised learning methods such as Naive Bayes, linear regression and logistic regression. All of these machine learning models have a different nature to determine the user ratings for an article. The research of [25] has presented a robust system for providing English real-time news recommendations to the user based on the user's history of the last visits to the website, popularity of stories and current user's context.

E. Arabic Language and Recommendation Systems

Arabic language is one of the most widespread languages globally [26], spoken by more than 422 million people. Its speakers are distributed in the region known as the Arab world and many other neighboring regions such as Turkey, Chad, Mali, Senegal and Eritrea. Arabic language considered as one of the Semitic languages that are rich in syntax (words arrangement to make phrases and sentences) and its morphology (the structure of internal word) [27]. There are three forms of Arabic language: classical Arabic [28], which is the language of Islam's Holy Book (Qur'an) that is not used in daily life, modern standard Arabic (MSA) as well as dialect Arabic. MSA is the standard form that is used in official news, education and media.

Unfortunately, the Arabic language's Recommendation Systems did not acquire enough attention [29]. It happened due to the limited number of tools and other challenges related to the language's nature. According to [30], the Arabic language became a challenge for researchers and machine learning developers due to language richness, language complexity, ambiguous structure of Arabic, and available Arabic types dialects.

F. The Microblogging Service Twitter

Microblogging is a general term of any web service that allows the users to broadcast short messages to other users of the service quickly from mobile devices. One of the first microblogging services is Twitter.com [31]. Twitter is one of the digital platforms where users can read news. It is a microblogging or a social networking service on which users can post and interact with messages known as tweets [32]. Registered users can like, post and retweet any tweets, while unregistered users can read them only. Any user can access Twitter through the website interface, through short message service SMS or its mobile application.

Now-a-days, Twitter has become one of the most popular services with students, academics, politicians, policymakers and the general public. In the past, many users find some difficult in understanding what Twitter is and how they can use it. However, now it has becomes the social media platform of choice for many.

In collaboration with research company DB5, the American Press Institute with Twitter has created a new study that explores and examines the relationship between news use and the Twitter environment [33]. The study involved an online survey of more than 4,700 active social media users. It found that Twitter users tend to be heavier newsreaders than other social media users. The study also found that reading news is one of the main activities they engage in on the network.

While these users on the service in general, or sometimes do so just as a way of passing the time, they behave differently when following breaking news. They participate, comment, post and share at moments more when events are moving fastest. These signals can lead news publishers to make more effective use of social media in general and Twitter in more particular.

Due to the limitation discussed above, we propose a hybrid personalized Arabic news recommender system that recommends interesting articles or news to users using Twitter service. The proposed tool sorts the news in two ways: Firstly, using a content-based filtering algorithm to consider the user's profile to recommend news articles to the user. Secondly, using collaborative filtering algorithms to consider the article's popularity with tweets from Twitter's timeline. We propose this approach to help users find interesting news articles to read by combining the above two filtering techniques of sorting articles [34].

III. Data Gathering and Preprocessing

A. Data Collection

1) Obtaining news articles: In order to collect Arabic News articles, we accessed RSS (Really Simple Syndication) feeds of online Arabic news sites such as Alarabiya (https://www.alarabiya.net/) and CNNArabic (https://arabic.cnn.com/). RSS is an XML file format created by news website editors for distributing and sharing web content, such as news headlines and summaries. Using an RSS reader, users can view data feeds from many news sources [35]. Alarabiya and CNNArabic organize their news articles by categories, e.g., Politics, Economics, Sport, etc. Therefore we know the associated topic for every article in the collection. The article collection includes 150 different Arabic news articles collected from six different main categories (Sports, Economics, Technology, Entertainment, Health and Politics).

2) Obtaining tweets dataset: In order to obtain a dataset of tweets, a streaming API is set up to store incoming tweets as

soon as they are posted in the Twitter public timeline using Tweepy Library. The data returned from Twitter streaming API are formatted using JSON (JavaScript Object Notation). In this research we only need the text attribute that contains the actual tweet text, the other attributes in the tweet object were ignored. The dataset for this research is a collection of 100,000 of last published tweets which are collected along with the news articles dataset on the same day.

B. Data Preprocessing

To create all data preprocessing steps for Arabic text, we decide to use CAMeL tools [36]. It is an open-source toolkit for Arabic natural language processing in Python. CAMeL tools offer utilities for preprocessing, dialect identification, sentiment analysis, named entity recognition and morphological modelling. What makes CAMeL the most suitable tool for our proposed system is that it is implemented in Python which is our research programming language, whereas other available tools are only supporting Java. The tools also provide APIs and command line interfaces in order to cover CAMeL's utilities, where others provide only command-line tools. This can lead to overhead writing glue code that includes interfaces to different packages. Through using CAMeL tools, we have followed the steps below in order to preprocess the obtained data.

1) News articles cleaning: The RSS news articles need to be preprocessed before making them ready for implementation. We needed to remove unnecessary content such as numbers, special characters and HTML tags.

2) Tweets cleaning: On the other side the tweets dataset need to be cleaned as well to eliminate unwanted noise and preserve textual content only, this can be done be by addressing the issues of character replication, abbreviations, hashtag, URLs, usernames, whitespace and removing repeated tweets.

3) Tokenization: The first step after data cleaning is tokenization; this step consists of splitting the text into tokens or words separated by punctuation characters or whitespaces. The result of this phase is a set of words.

4) Normalization: It is the process to change all the forms of a word into a common form. To get texts common forms, the added normalizer removes the 'tatweel' character '_'such as changing the word "واســـع" which means "wide"to look like "واسع ",and tashkeel such as "هُدنَةٍ" which means "truce" to make it as "هدنة". There are also other steps include removing symbols and special characters such as ! ? (), numbers such as 6which is 6 in English , non-Arabic words such as طبعا "NO", and Al "ال التعريف" which is used as the definite article 'the' in English. We will also replace the letters "أ" and "إ" with the letter "ا" and the letter "ة" with the letter "ه".

5) Stop words removal: Stop words such as prepositions and pronouns are used frequently in tweets and news articles dataset. We decided to remove them since they are not significant in the research. For instance, these words have been removed: "من" of, "على" on, "انت" you and so on.

Fig. 1 shows the general dataset cleaning and preprocessing phases.

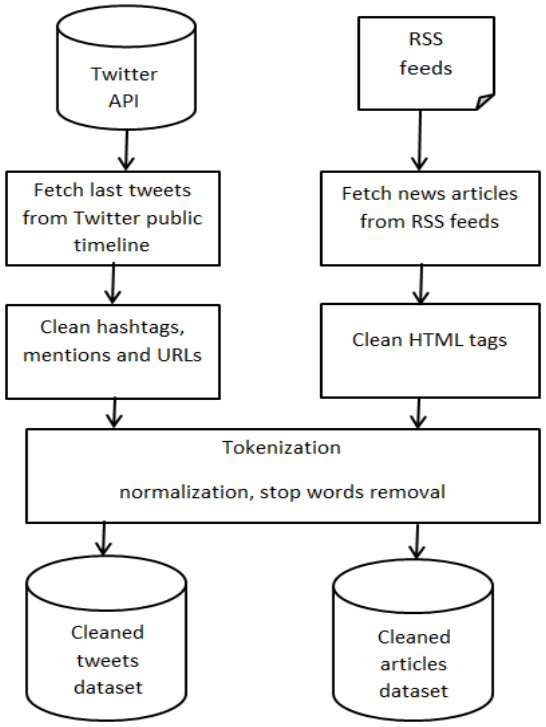


Fig. 1. Dataset Cleaning and Preprocessing.

IV. Design and Implementation

A. High-Level Design

Our proposed system consists of three parts: collaborative recommender system which is popularity-based, content-based recommender system which is profile-based and hybrid recommender system which is the combination of the two previous systems.

The first recommender selects news articles based on the popularity of the article. In this research, the article popularity is identified with the help of the Twitter service. So tweets are gathered from Twitter public timeline and preprocessed to identify the articles that users around the world are tweeting. The tweets are then compared to the news articles based on the co-occurring terms in the articles and the tweets. Articles that mentioned frequently in the tweets will be considered popular or hot.

The second recommender sorts the news articles based on their similarity to the user's profile. In this research, the users build their profiles by providing input about their level of interest in each of six news categories. The incoming news articles are classified into same set of categories using K Nearest Neighbors text classifier. This classifier works by finding the K nearest matches in the training data and then using the label of closest matches for predicting [37]. The articles are then sorted based on the similarity between the categories the user showed interest in his/her profile and the categories which the article belongs to.

Finally the hybrid recommender combines the results from the two recommenders to recommend news articles to the user. It sorts the new articles based on the combination of their popularity ranking and the similarity to the user's profile.

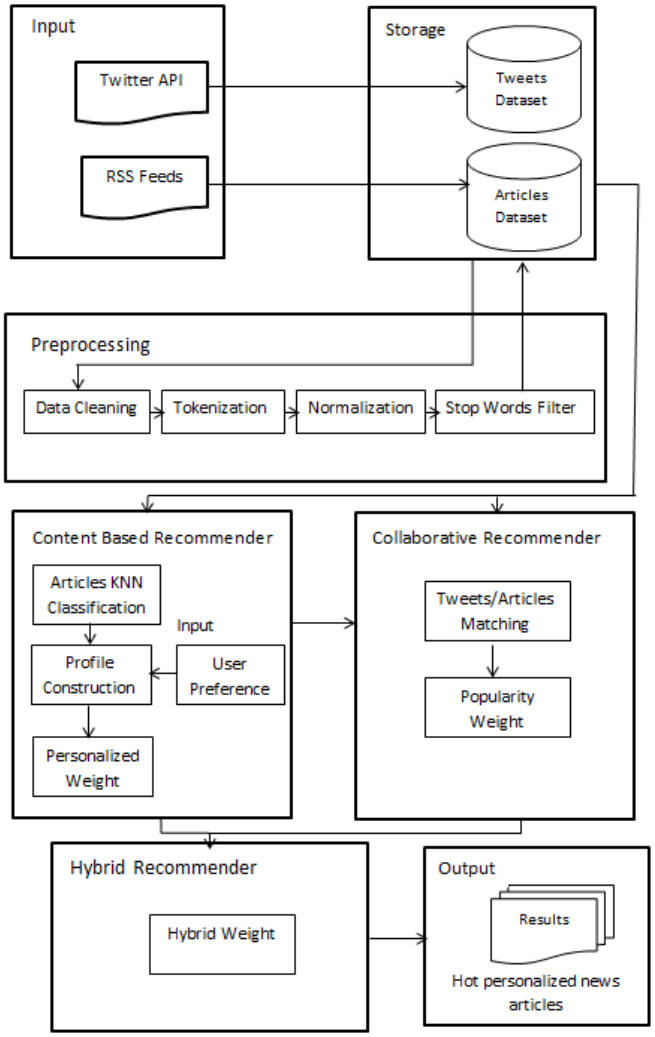


Fig. 2. General Architecture of the System Components.

Fig. 2 shows an architectural diagram of the proposed hybrid recommender system.

B. Collaborative News Recommendation

After the RSS news articles are gathered and preprocessed, the news articles are then indexed using ElasticSearch. ElasticSearch is a free open source search engine built on top of Apache Lucene [38]. It can index many types of content and can be used for several cases such as website search, application search, monitoring application performance, business analytics, etc. It is known for its simple REST APIs which make it flexible to use from any programming language. In addition, it supports 34 languages including Arabic and provides analyzers for every language.

To identify the most popular news articles, the processed tweets are passed as queries to ElasticSearch server. Every tweet is queried against the server to return the articles that correspond to the tweet content. The server returns the articles associated with the tweet content together with a weight that refers to the similarity between the tweet as query and the article. So the articles are sorted according to their accumulated weights across all tweets. The news articles will be ranked based on their popularity using cosine similarity measures between articles and tweets as shown in (1) [39].

$$\text{Popularity Weight} = \text{CosineSimilarity}(\text{Article}, \text{Tweet}) \quad (1)$$

C. Content-Based News Recommendation

This recommender uses the same news articles collection as in the collaborative recommender. Although, news articles are placed in just one category by the website editors, they actually might belong to more than one category. To allow for that in the recommender, each news article is classified into seven potential categories using K Nearest Neighbor classifier. With this algorithm we can classify the articles into categories based on the articles similarity to the training documents for each category.

K-Nearest Neighbors classification algorithm [37] identifies K most similar training documents to the article and then uses these similarity scores as a vote for the category that the training documents belong to. Thus the article similarity to each category is the sum of the scores of the article's similarity to the training documents in that category which placed in overall top K most similar documents. So then the categories are sorted through their accumulated score. Then we use ElasticSearch again to create a second index that maps from category IDs to documents IDs and weights. So this index stores news articles based on the category that they belong to instead of the keywords in the article.

Every user creates his profile based on his interest in different categories, every user has an interface to build his profile manually by scoring the categories. Fig. 3 shows the interface that allows the user to create his unique profile. Each user will enter scores in the range of (0-10) so that the total at the end will equal to 10.

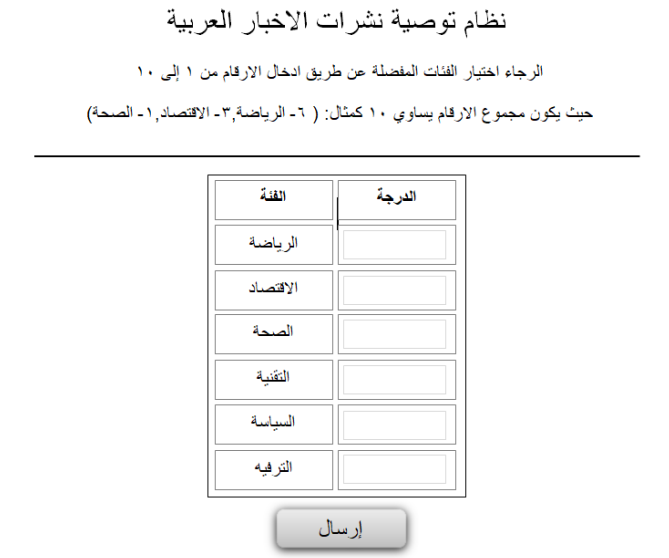


Fig. 3. User Profile Interface.

After that, this user profile is used to identify the documents which best match his profile. The articles and profiles can be considered as feature vectors where every category is a feature. In (2) we have used cosine similarity measure [39] to get the similarity of each article to the user's profile as shown below.

Personalized Weight = CosineSimilarity(Article, Profile) (2)

D. Hybrid News Recommendation

This system combines the scores result by each of the previous two recommenders to generate a new recommendation that integrates the news articles' match to the user's interest with the news articles popularity everywhere. It calculates the hybrid weight through multiplying popularity weight by personalized weight as shown in (3) [40].

HWeight= Popularity Weight* Personalized Weight (3)

V. Evaluation

One of the important phases in building any recommender system is evaluation. There are different metrics to evaluate the performance; the most common are accuracy and experimental studies [41]. Accuracy metrics are divided into decision support and statistical metrics. Decision support metrics show prediction procedure as a binary operation (0 and 1) to distinguish good items from the items that are not good while statistical metrics evaluate the accuracy through comparing the predicted rankings or ratings directly with actual rating.

Deciding about the suitable metric depends on the dataset features and the type of tasks that the recommendation system will do. Based on our research and collected dataset we have evaluated our system through the two metrics, calculating the statistical accuracy through using Mean Absolute Error (MAE). The other metrics is done through performing a user experimental study to measure user satisfaction.

A. Mean Absolute Error (MAE)

It is the most commonly used metric in recommender systems [42]. This metric provides measures and illustrates the difference between the actual and estimated prediction over many items and users. It is given in (4):

$$MAE = \frac{\sum_{i=1}^{d}|a_i - p_i|}{d} \quad (4)$$

Where a is actual observation and p is predicted value divided by d which is the total number of actual ratings or ranks in an item set, the smaller value of MAE the better accuracy we get.

To calculate MAE, we have divided the dataset into training and testing sets. 70% of the data has been used as training set and the rest 30% of data has been used as test set, so variable R in Table I represents the train/test ratio that illustrates the percentage of data used in training set. We have compared MAE results across the three filtering approaches as shown in Table I.

Fig. 4 shows the comparison results of the three filtering approaches, we can observe that the greater value of R is the Lower value of MAE. Besides the most important thing we can notice is that hybrid filtering has the best MAE results compared to the other two approaches.

B. User Experimental Study

We have created a random sample of 25 respondents (n=5) who are native Arabic Speakers from different educational levels. The respondents have created their user profiles manually by providing their interests in the six news categories to form the user profile. Every respondent has a web page to enter weights for every category and the weights must be sum to ten. We also have made three results sets from news collection for every respondent based on the three filtering approaches. The presentation order of news articles was random so the respondent could not realize which system recommended the articles. The respondent has asked to rate each article in a 3-point likert scale as it is relevant to his interest, somehow relevant or not relevant. When the respondent finishes, all information such as filtering approach, article's rank, article's weight, and respondent's rating are logged into a file. Fig. 5 shows a snapshot example of one of the respondents' response.

TABLE I. MAE Comparison Results

Training Ration	Content-Based filtering	Collaborative filtering	Hybrid filtering
R=0.1	0.646	0.618	0.520
R=0.2	0.634	0.597	0.502
R=0.3	0.619	0.569	0.497
R=0.4	0.602	0.547	0.471
R=0.5	0.578	0.524	0.454
R=0.6	0.554	0.502	0.422
R=0.7	0.521	0.478	0.387

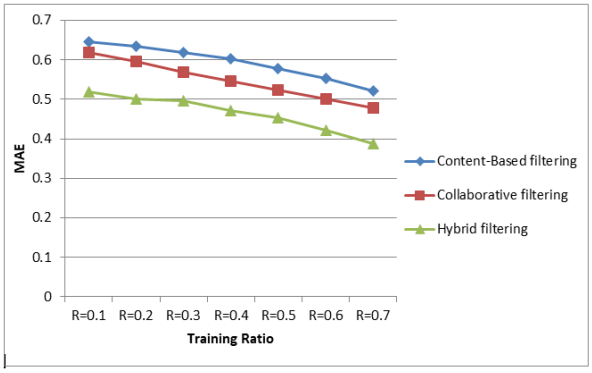


Fig. 4. MAE Comparison Result.

ID	S1 Rank	S1 WT	S2 Rank	S2 WT	S3 Rank	S3 WT	Resp Rating
122	70	0.16634	7	0.91926	20	0.15291	1
621	10	0.47541	138	0.00028	143	0.00013	0
540	25	0.37411	11	0.91634	6	0.34281	2
131	43	0.27787	6	0.92714	13	0.25763	1
618	3	0.68637	0	0	0	0	0
604	6	0.57957	33	0.69677	2	0.40383	2
439	37	0.32415	14	0.91517	10	0.29666	1
531	22	0.41517	5	0.92755	4	0.38509	2
117	8	0.50476	101	0.02954	96	0.01491	0
315	9	0.48671	52	0.68637	8	0.33407	1

Fig. 5. Example of Respondents' Response.

In Fig. 5, S1 Rank refer to where the article was ranked by the collaborative recommender and S1 WT indicate the normalized weight of that article, the same information is available for S2 (the content-based recommender) and S3 (the hybrid recommender). The Resp Rating column shows how the respondent rated each article, where 2 is for relevant, 1 is for somehow relevant and 0 for not relevant. To analyze the result we have calculated the average rank for all respondents over the three filtering approaches as shown in Table II.

TABLE II. AVERAGE RANK

	Not relevant	**Somehow relevant**	**Very relevant**
Collaborative	58.3	70.0	59.1
Content-Based	75.6	63.5	57.4
Hybrid	56.1	54.2	40.5

The table displays the average ranking of the top ten articles presented in each approach. For instance, the articles rated by the respondents as very relevant, the collaborative system ranked those articles on average as 59^{th}. The content-based system ranked the very relevant articles 57^{th} while the hybrid ranked them as 40^{th}. So it is shown clearly the hybrid system ranked the articles higher than the other systems. This study supports MAE results where we have seen that hybrid system get less MAE value than other two systems.

VI. DISCUSSION AND CONCLUSION

This research aims to develop an Arabic recommender system to display news or articles to each reader based on their interests instead of presenting them only in the order of their occurrence using Twitter. To achieve that, we started with collecting the news articles and the dataset of tweets from Twitter and preprocessing them to prepare them to be used in building the recommender system. Then we explained the design and implementation of all the components of the proposed recommender system. Lastly, we have evaluated the recommender system through the two metrics, calculating the statistical accuracy through using Mean Absolute Error (MAE). The other metrics is done through performing a user experimental study to measure the user satisfaction.

Based on the result of the evaluation we found that the hybrid system performs better than the collaborative and content-based recommendations. This means that news articles recommended by the hybrid recommender system are more relevant to the user compared to the other two systems.

Although this research has successfully achieved the goal of developing a personalized Arabic news recommender system, it has certain limitations that are inherent in most of the recommender system researches. First, the size of the created dataset was relatively small, which imposes the need to evaluate the robustness of proposed system against large datasets. Second, creating a new application for Twitter developer account was another challenge. Setting up a Twitter API wasn't easy as it used to be, we have tried several times with Twitter team until we got final application approval.

In future, we have many directions to enhance this research. The accuracy of the proposed news recommender system can be further improved through incorporating other features such as location. In this case the system can recommend news articles based on the user's geographical location. We can increase the efficiency of the system on a wider range of Arabic news articles in order to scale the system and enhance its algorithm so it can apply in different categories of news articles.

REFERENCES

[1] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang and Q. Mei, "Beyond Ranking: Optimizing Whole-Page Presentation," in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, 2016.

[2] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," in Proceedings of the 10th International Conference on World Wide Web, Hong Kong, 2015.

[3] S. Kemp, "Digital 2020 Global Overview Repor," The Next Web, Amsterdam, 2020.

[4] K. Saranya and S. Sudha, "A Personalized Online News Recommendation System," International Journal of Computer Applications, pp. 6-7, 2012.

[5] F. Garcin, C. Dimitrakakis and B. Faltings, "Personalized news recommendation with context trees," in Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, 2013.

[6] C. García , V. García-Díaz, D. Meana-Llorián and E. Núñez-Valdez , "Social Recommender System: A Recommender System Tweets for points of interest," in 4th Multidisciplinary International Social Networks Conference, Bangkok, 2017.

[7] J. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," User Modeling and User-Adapted Interaction, pp. 101-103, 2012.

[8] C. Pan and W. Li, "Research paper recommendation with topic analysis," in 2010 International Conference On Computer Design and Applications, Qinhuangdao, 2010.

[9] F.O. Isinkaye, Y.O. Folajimi and B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian Informatics Journal, pp. 261-263, 2015.

[10] B. Bhatt, P. J. Patel and H. Gaudani, "A Review Paper on Machine Learning Based Recommendation System," International Journal of Engineering Development and Research, p. 3955, 2014.

[11] P. Pu, L. Chen and R. Hu, "A user-centric evaluation framework for recommender systems," in RecSys '11: Proceedings of the fifth ACM conference on Recommender systems, New York, 2011.

[12] R. Hu and P. Pu, "Acceptance issues of personality-based recommender systems," in RecSys '09: Proceedings of the third ACM conference on Recommender systems, New York, 2009.

[13] J. Herlocker, J. Konstan, L. Terveen and J. Riedl, "Evaluating collaborative filtering recommender systems," Association for Computing Machinery, pp. 5-7, 2004.

[14] C. C. Aggarwal, Recommender Systems, Switzerland : Springer, Cham, 2016.

[15] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez, "Recommender systems survey," Knowledge-Based Systems, pp. 109-111, 2013.

[16] S. Bostandjiev, J. O'Donovan and T. Höllerer, "TasteWeights: a visual interactive hybrid recommender system," in RecSys '12: Proceedings of the sixth ACM conference on Recommender systems, New York, 2012.

[17] A. Popescul, L. H. Ungar, D. M. Pennock and S. Lawrence, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," arXiv, p. 437, 2013.

[18] G. Adomavicius and J. Zhang, "Impact of data characteristics on recommender systems performance," ACM Transactions on Management Information Systems, pp. 1-3, 2012.

[19] D. H. Stern, R. Herbrich and h. Graepel, "Matchbox: large scale online bayesian recommendations," in Proceedings of the 18th international conference on World wide web, New York, 2009.

[20] C. Wei-Ta and T. Ya-Lun , "A hybrid recommendation system considering visual information for predicting favorite restaurants," Springer US, New York, 2017.

[21] K. Pigi , S. James , P. Jay and O. John , "User Preferences for Hybrid Explanations," in RecSys '17 Proceedings of the Eleventh ACM Conference on Recommender Systems, Como, 2017.

[22] A. Flodin, "A scalable product recommendation engine suitable for transaction," Mid Sweden University, p. 4, 2018.

[23] J. Jens, “Interactive television: new genres, new format, new content,” in Proceedings of the second Australasian conference on Interactive entertainment, Sydney, 2015.

[24] C. Akshay, "Recommender System for News Articles Using Supervised Learning," Universitat Pompeu Fabra, Barcelona, 2014.

[25] B. Fortuna, C. Fortuna and D. Mladenić, "Real-Time News Recommender System," in Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases, Barcelona, 2010.

[26] A. Alharsh, "Computer and Learning Arabic language," Mentouri University, pp. 217-218, 2007.

[27] N. Madi and H. Al-Khalifa, "Error Detection for Arabic Text Using Neural," MDPI, pp. 2-3, 2020.

[28] A. Wahhab and A. Hassan, "Proposed aspect extraction algorithm for Arabic text reviews," Journal of AL-Qadisiyah for computer science and mathematics, p. 80, 2018.

[29] B. Hawashin, S. Alzubi and T. Kanan, "An efficient semantic recommender method," Emerald Insight, p. 263, 2019.

[30] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM, pp. 1-2, 2009.

[31] P. Jackson , "Web 2.0 tools and context," in Web 2.0 Knowledge Technologies and the Enterprise, Manchester, Chandos Publishing, 2010, pp. 11-12.

[32] J. An, M. Cha, K. Gummadi and J. Crowcroft, "Media Landscape in Twitter: A World of New Conventions and Political Diversity," in Proceedings of the International AAAI Conference on Web and Social Media, Barcelona, 2021.

[33] T. Rosenstiel, J. Sonderman, K. Loker, M. Ivancin and N. Kjarval, "Twitter and the News: How people use the social network to learn about the world," American Press Institute, U.S, 2015.

[34] D. Surabhi, "Recommender system in education," in 5th National Conference on E-Learning & E-Learning Technologies, India,, 2017.

[35] S. Agarwal, "Classification of RSS feed news items using ontology," in 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, 2012.

[36] O. Ossama, "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, 2020.

[37] M. Azam, T. Ahmed, F. Sabah and M. I. Hussain, "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm," IJCSNS International Journal of Computer Science and Network Security, vol. 18, no. 12, pp. 95-100, 2018.

[38] C. Gormley and Z. Tong, "CHAPTER 1: You Know,For Search," in Elasticsearch The Definitive Guide: A Distributed Real-Time Search and Analytics Engine, Sebastopol, O'Reilly Media, 2015, pp. 3-4.

[39] M. B. Magara, S. O. Ojo and T. Zuva, "A comparative analysis of text similarity measures and algorithms in research paper recommender systems," in 2018 Conference on Information Communications Technology and Society (ICTAS) , Durban, 2018.

[40] Suriati1, M. Dwiastuti and Tulus, "Weighted Hybrid Technique for Recommender System," in International Conference on Information and Communication Technology (IconICT), 2017.

[41] F. Isinkaye, B. Ojokoh and Y. Folajimi, "Recommendation systems: Principles, methods and," Egyptian Informatics Journal, p. 270, 2015.

[42] F. Harrag, A. Al-Salman and A. Alquahtani, "Arabic Opinion Mining Using a Hybrid Recommender System Approach," ArXiv, pp. 17-18, 2020.

Machine Learning Model through Ensemble Bagged Trees in Predictive Analysis of University Teaching Performance

Omar Chamorro-Atalaya[1]
Facultad de Ingeniería y Gestión
Universidad Nacional Tecnológica de Lima Sur
Lima- Perú

Carlos Chávez-Herrera[2]
Facultad de Ingeniería de Sistemas e Informática Universidad Nacional Mayor de San Marcos
Lima-Perú

Marco Anton-De los Santos[3]
Juan Anton-De los Santos[4]
Facultad de Ciencias Económicas
Universidad Nacional Federico Villarreal, Lima-Perú

Almintor Torres-Quiroz[5]
Facultad de Ciencias Económicas
Universidad Nacional del Callao
Lima-Perú

Antenor Leva-Apaza[6]
Facultad de Ciencias
Universidad Tecnológica del Perú
Lima-Perú

Abel Tasayco-Jala[7]
Gutember Peralta-Eugenio[8]
Facultad de Ciencias Empresariales and Facultad de Ciencias de la Salud
Universidad César Vallejo
Lima-Perú

***Abstract*—The objective of this study is to analyze and discuss the metrics of the Machine Learning model through the Ensemble Bagged Trees algorithm, which will be applied to data on satisfaction with teaching performance in the virtual environment. Initially the classification analysis through the Matlab R2021a software, identified an Accuracy of 81.3%, for the Ensemble Bagged Trees algorithm. When performing the validation of the collected data, and proceeding with the obtaining of the predictive model, for the 4 classes (satisfaction levels), total precision values of 82.21%, Sensitivity of 73.40%, Specificity of 91.02% and of 90.63% Accuracy. In turn, the highest level of the area under the curve (AUC) by means of the Receiver operating characteristic (ROC) is 0.93, thus considering a sensitivity of the predictive model of 93%. The validation of these results will allow the directors of the higher institution to have a database, to be used in the process of improving the quality of the educational service in relation to teaching performance.**

Keywords*—Machine learning; ensemble; bagged trees; predictive analysis; teaching performance

I. Introduction

The information and communication technology (ICT) sector is currently a leader in the analysis of data from different media [1], [2], such as virtual platforms, survey administration software, among other technological tools [3], [4], which capture or acquire information to be processed and analyzed in descriptive statistical research or in research on predictive models applicable to various areas of knowledge [5].

The advantages that the introduction of ICT has generated in the education sector is based on the importance of technology to develop research that previously could not be carried out, [6], [7] as is the case of the identification of predictive models for the analysis or monitoring of university teaching performance, student performance, among other relevant factors for the education sector [8]-[10].

Worldwide, the education sector has undergone changes and transformations, due to the virtualization of the teaching-learning mode, [11], [12], [13], as a consequence of this scenario, universities face new challenges, to safeguard the quality of education that goes hand in hand with the advancement of technology [14]-[16].

Given this, in the education sector, an increasing amount of data has been generated with greater relevance, product of the iterations of the different actors of the educational process, these being the teacher, the students and the institution, through the application of tools technological, such as survey software, which generate a database [17], [18]. As indicated, the data that are stored, are used in order to improve the efficiency of the educational process through predictive models, among the factors to optimize are academic performance, student dropout, teaching performance, graduate follow-up [19].

There are various technologies used to obtain predictive models, which use data from virtual platforms and survey administration software, applied to students by universities [20]. Within these technologies is the branch of Artificial Intelligence that within its fields houses Machine Learning [21]-[23]. As indicated in [24], Machine Learning is a set of algorithms capable of learning to perform certain tasks from the generalization of examples. Machine Learning has been successfully applied to a variety of areas of human endeavor, and has recently been applied to the educational sector, whose purpose is oriented towards the design of algorithms, methods and models, which will allow the exploration of data from teaching-learning environments [25], [26].

Among the multiple algorithms of Machine Learning, there is Ensemble Bagged Trees, which is an algorithm that is used in joint learning [27]. This can combine training and base

classifiers to produce ensemble models or use an algorithm with multiple test data sets as the basis [28]. In this regard, in [29] it is pointed out that the Bagged Trees algorithm forms different trees when there is a change in the starting point of the training data that results in a decrease in stability. This technique or algorithm is also suitable to be used in the search for optimal models for large data, since the classification becomes easier [30], [31].

In this sense, the main objective of this article is to determine the predictive model using Machine Learning through the Ensemble Bagged Trees algorithm, for the predictive analysis of university teaching performance, in order to use it as part of the procedure to improve the quality of the educational process. Initially, the methodology used will be detailed, then the validation of the algorithm will be determined, by means of the accuracy and the confusion matrix, to finally analyze the total performance metrics (Accuracy (A), Precision (P), Sensitivity (S) and Specificity (R)) of the selected algorithm, from obtaining the receiver operating characteristic curve (ROC).

The contribution of the research focuses on applying a novel technique for the higher institution, through machine learning making use of the data and information collected, which allows making preventive and corrective decisions based on reliable results, obtained through a methodology not so complex.

II. Research Methodology

A. Type and Level of Research

The type of research is applied, since it starts from the identification of a problem, related to the improvement of university teaching performance, for which use is made of methods or tools already defined such as predictive models through Machine Learning, which employs the Ensemble Bagged Trees algorithm. Likewise, the research level is descriptive, since it focuses on analyzing and discussing the metrics of the predictive model obtained through the Ensemble Bagged Trees algorithm, applied to the perception data of engineering university students.

This research also seeks to design a predictive multidimensional model that can be used to create and store new data for the higher institution. Based on this technological tool, it determines patterns and calculates association rules, providing support and reliability to the results obtained. Performance metrics such as Accuracy, Precision, Sensitivity and Specificity show improved performance over the manual method of the same procedure commonly performed in research [28].

B. Participants

The participants in this research are made up of students from the sixth to the tenth cycle of professional engineering schools, with a total of 581 students, this selection criterion is part of a regulation established and approved by the higher institution. It should be noted that it was possible to collect data from the entire population, for this reason, it can be noted that the sample coincides with the population.

C. Data Collection Technique and Instrument

The data collection technique is the survey, and the instrument used to collect data regarding university teaching performance is the questionnaire, which was carried out virtually, due to the context of the health emergency declared by the Covid -19. The virtual platform of the higher institution was used, which gave access to the data collection instrument through the code of each student, which guaranteed the security and reliability of the information. The questionnaire consisted of responses on a Likert scale ranging in levels from 1 to 4 (from dissatisfied to very satisfied). These levels of satisfaction in the analysis will be represented as the classes of the predictive model. In Fig. 1, the indicators considered as predictive elements in the perception of university teaching performance are shown.

D. Reliability of the Collected Data

As part of the methodology, the validation of the collected data is carried out, through Cronbach's Alpha coefficient using the SPSS software, this analysis carried out, it is observed in Table I that the consistency coefficient is equal to 0.932. As indicated in [12], values greater than 0.9 indicate great consistency that is, high homogeneity and equivalence of the response of all indicators. Once this result is obtained, the following section shows the results.

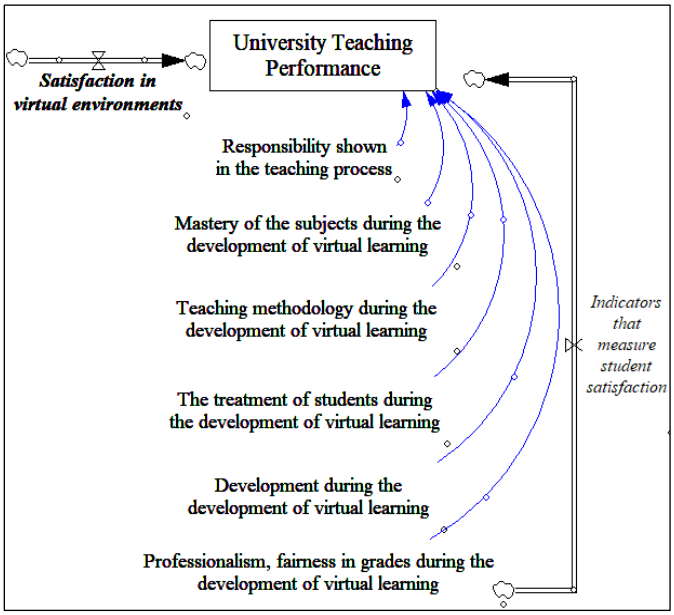


Fig. 1. Indicators that Measure Student Satisfaction with Teaching Performance.

TABLE I. Cronbach's Alpha Test

Reliability statistics	
Cronbach's alpha	***No. of elements***
0.932	6

E. Data Processing Design

The data processing design responded to a non-experimental transactional process, in which data was collected through a virtual questionnaire. In Fig. 2, the methodology of the research process is shown, which begins with the collection of data on the perception of engineering students from a public university in Peru. These data are related to the 6 indicators that are visualized in Fig. 1, whose appreciation regarding teaching performance is of an ordinal qualitative type, thus establishing 4 classes (very satisfied: 4, satisfied: 3, not very satisfied: 2 and dissatisfied :1).

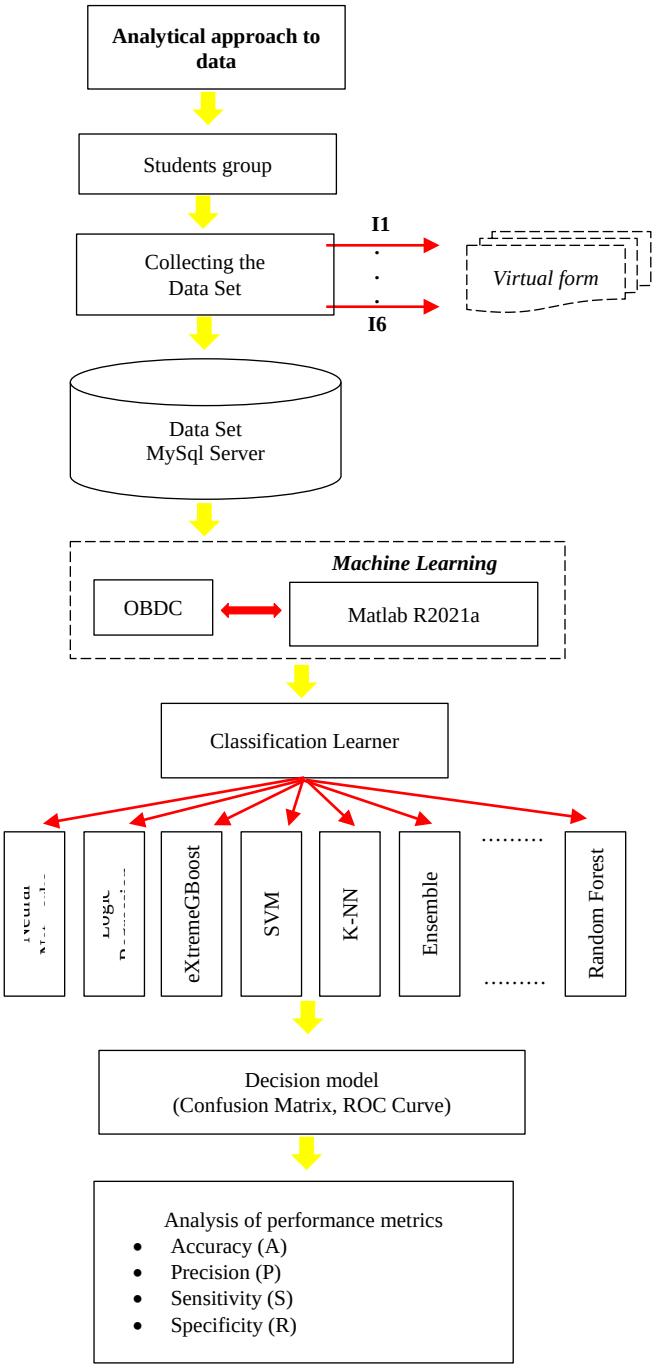


Fig. 2. Methodology of the Research Process through Machine Learning.

Likewise, the information collected was stored in a database in Microsoft SQL Server, associated through the Open Data Base Connectivity (OBDC) driver and the Matlab R2021a software. Using the Matlab software, we proceeded to use the "Classification Learner" tool, in order to identify the best Machine Learning algorithm, through its metrics. This algorithm allows the classification of students from the results obtained from the indicators specified in Fig. 1.

III. Results and Discussion

A. Determination of the Predictive Model

Using the Matlab R2021a software, and using the Classification Learner and Statistics and Machine Learning Toolbox 12.1 application, the best predictive model determined by the validation of the accuracy is identified, in Fig. 3, the results generated by the software are shown. Matlab R2021a.

As shown in Fig. 3, the Machine Learning algorithm that presents the best accuracy, for classifying the level of satisfaction with respect to university teaching performance, is the Ensemble Bagged Trees algorithm with an accuracy of 81.3%.

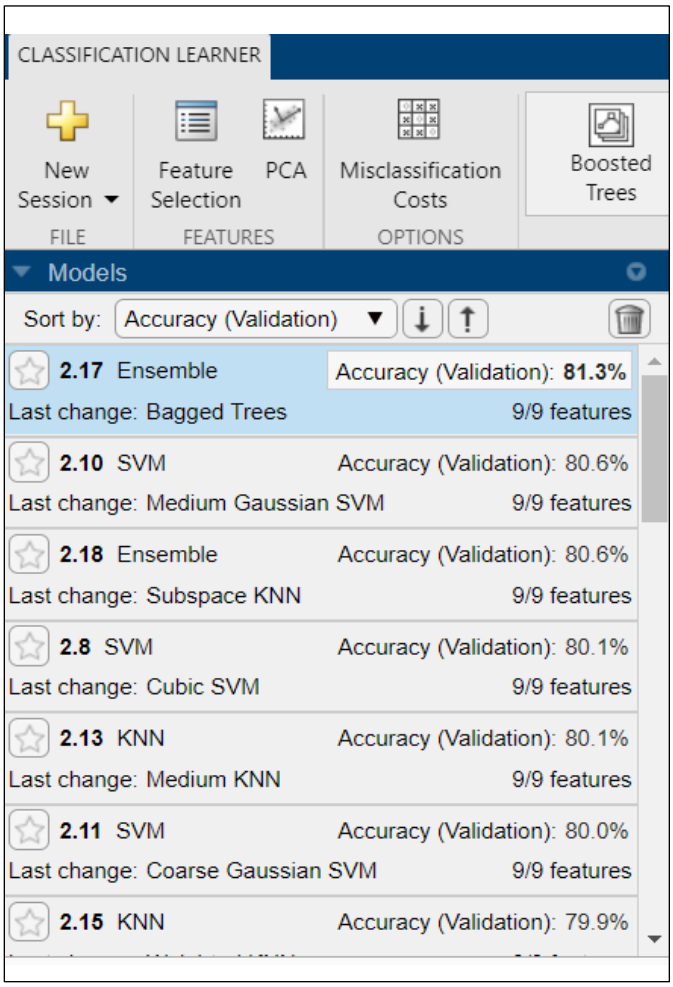


Fig. 3. Validation of the Prediction Algorithms Ordered by their Accuracy.

B. Results of the Predictive Model Metrics

When using the predictive model through Machine Learning through Ensemble Bagged Trees, to determine satisfaction with university teaching performance, confusion matrices are obtained, which represent elements of validation or performance measurement of the predictive model.

In Fig. 4, the confusion matrix is shown, with respect to the sensitivity metric, in it you can visualize the number of observations made by the classification system, and it reports the number of false negatives (FNR), which is the number of positive examples wrongly classified as negative and true positives (TPR) that define the number of positive samples correctly classified as positive, which shows the closeness between the levels of satisfaction predicted (Predicted class) by the model with respect to its true value (True class).

As can be seen in Fig. 4, of the 4 classes on which the predictive model acts through Ensemble Bagged Trees, class 3 shows the highest percentage of sensitivity, this means that the predictive model has the ability to discriminate between a true positive (TP) of a false negative (FN) in this class (satisfied), in this case it is 89.9%, as observed in this class the model was only confused by 10.1%. While the lowest level of sensitivity of the predictive model is shown in class 1 (satisfaction level: dissatisfied), whose value is 63.9%.

In Fig. 5, the confusion matrix is shown with respect to the precision metric, since the values of the main diagonal indicate the precision of the predictive model for each class.

In Fig. 5, the confusion matrix is shown regarding the precision metric of the predictive model for each class, in which it is observed that the predictive model for class 1 (satisfaction level: dissatisfied) shows the highest precision rate, in this case it is 88.5%. This result indicates that the level of dispersion of the data for this class is very low.

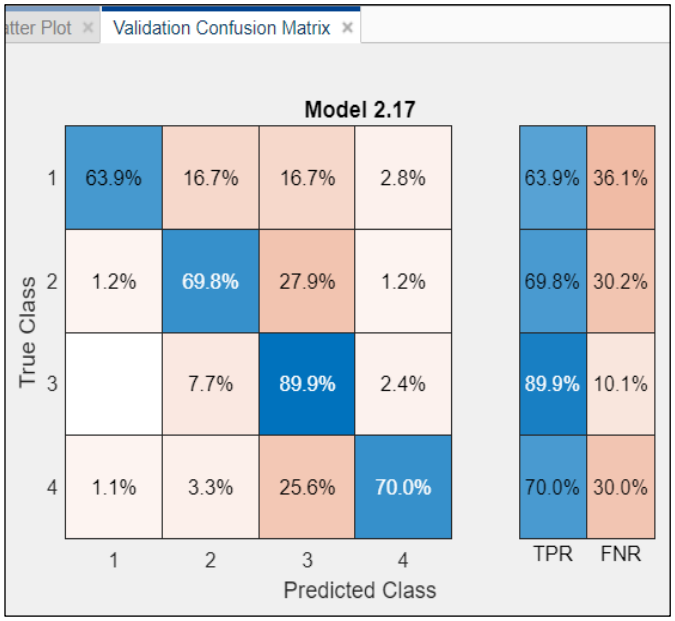


Fig. 4. Confusion Matrix based on TPR and FNR rates.

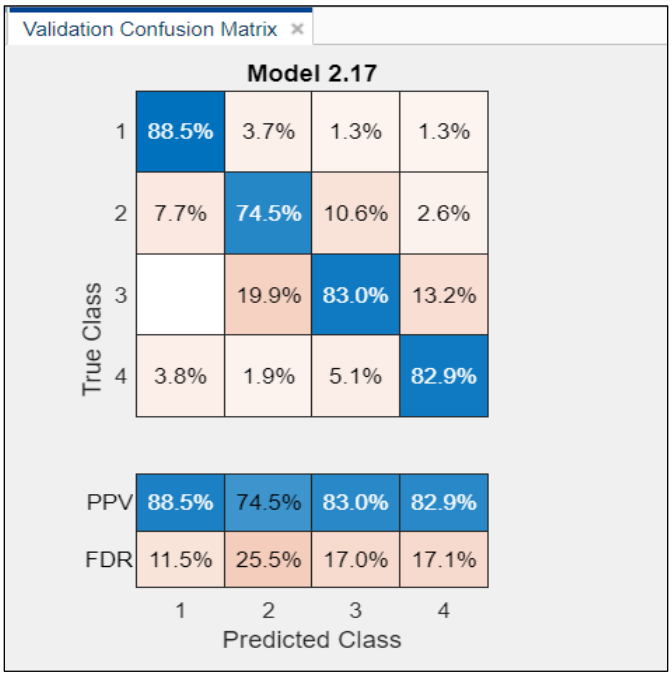


Fig. 5. Confusion Matrix based on PPV and FDR rates.

Table II shows the metrics of the predictive model through Ensemble Bagged Trees, for each class, in which it is evidenced that the total Precision is 82.21%, the total Sensitivity is 73.40% and the total Specificity is 91.02%, and the Accuracy presents a total value of 90.63%.

As part of the predictive model through the Ensemble Bagged Trees algorithm, the response that Matlab provides for each class under study is evidenced, its corresponding Receiver operating characteristic (ROC) graph and considering that the ROC graph describes the Sensitivity and Specificity of the algorithm classifier, the findings in Fig. 6, allow us to establish that for class 1 (dissatisfied), a sensitivity of 93% is shown.

In addition, the discrimination threshold is 0.64 for the rate of true positives and 0.00 for the rate of false positives, showing an area value on the curve (AUC) of 0.93, this value being close to 1, it is noted that the model for class 1 is optimal.

In Fig. 7, the ROC graph for class 2 (not very satisfied) is shown, where a sensitivity of 91% is displayed. In addition, the discrimination threshold is 0.70 for the rate of true positives and 0.08 for the rate of false positives, showing an area value on the curve (AUC) of 0.91, this value being close to 1, it is noted that the model for class 2 is optimal.

TABLE II. CLASSIFICATION PREDICTIVE ALGORITHM METRICS

Class	Metrics			
	Sensitivity	***Specificity***	***Accuracy***	***Precision***
1	63.89%	99.56%	97.76%	88.46%
2	69.77%	92.45%	86.99%	74.53%
3	89.93%	74.16%	83.36%	82.96%
4	70.00%	97.92%	94.41%	82.89%

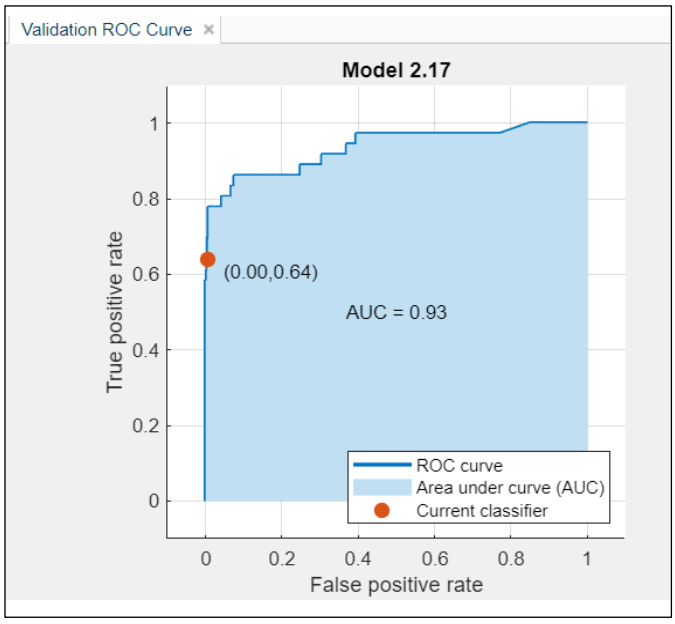


Fig. 6. ROC Charts for Class 1.

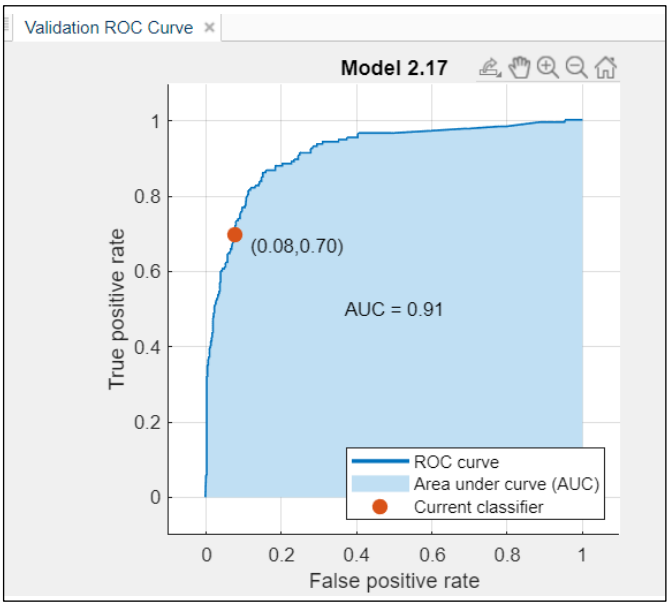


Fig. 7. ROC Charts for Class 2.

In Fig. 8, the ROC plot for class 3 (satisfied) is shown, where a sensitivity of 91% is displayed. In addition, the discrimination threshold is 0.90 for the rate of true positives and 0.26 for the rate of false positives, showing an area value on the curve (AUC) of 0.91, this value being close to 1, it is noted that the model for class 3 is optimal.

Finally, in Fig. 9, the ROC graph for class 4 (very satisfied) is shown, where a sensitivity of 92% is displayed. In addition, the discrimination threshold is 0.70 for the rate of true positives and 0.02 for the rate of false positives, showing an area value on the curve (AUC) of 0.92, this value being close to 1, it is noted that the model for class 4 is optimal.

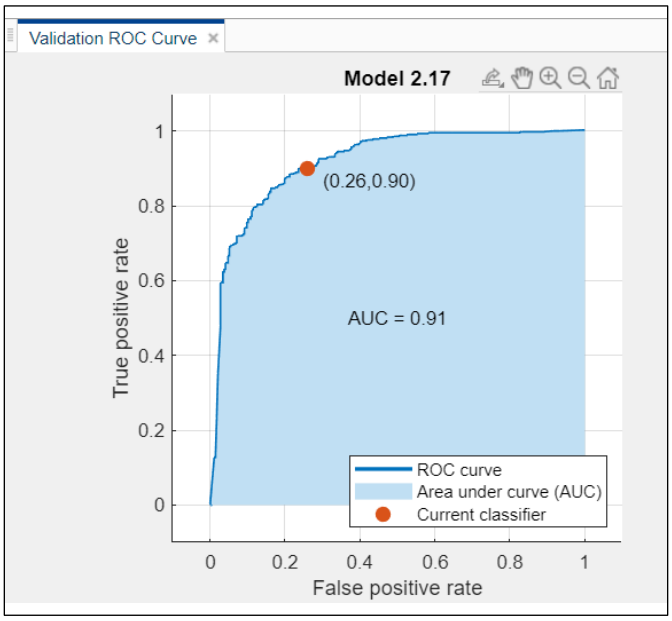


Fig. 8. ROC Charts for Class 3.

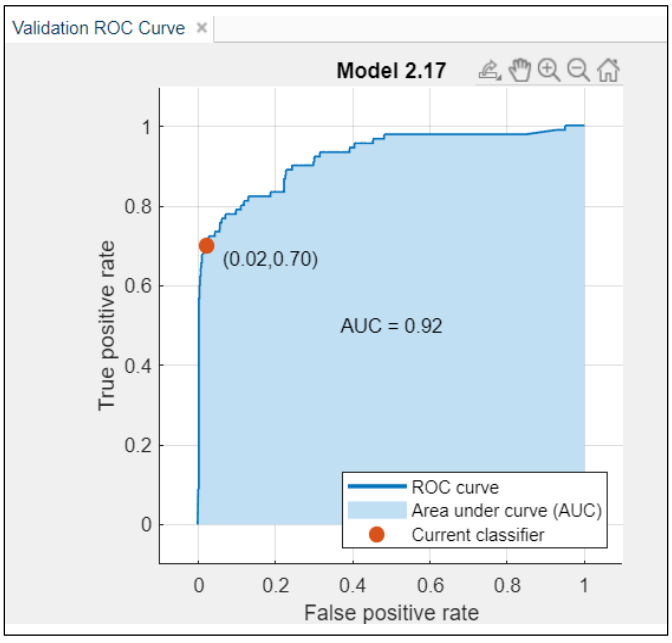


Fig. 9. ROC Charts for Class 4.

C. Discussion

In relation to the results obtained, it is evidenced that the predictive model, based on the Ensemble Bagged Trees algorithm, presents acceptable metrics of precision, sensitivity, specificity and accuracy, in its 4 classes each of its classes, in this way the predictive model obtained provides security and reliability, contributing to decision making to improve the quality of the course content and the pedagogical methodology. In this regard, in [16] it is pointed out that preventive and corrective decision-making in higher education institutions involves building predictive models based on intelligent systems.

As indicated in [6], researchers have been concerned in recent years to work on the development of models that allow understanding aspects of the academic life of the student, teachers and institutions that allow the preparation and making of correct decisions, for the improvement continuity of educational quality. Likewise, in [19] it is indicated that the results obtained and validations show a precision of 82%, therefore, it can be pointed out that the process describes an optimal performance of the algorithms, so its incorporation would be satisfactory to be incorporated to the management of virtual educational knowledge.

In relation to the metrics of the predictive model, the model obtained through Matlab R2021a presents a general precision of 82.21% and an accuracy of 90.63%, being considered an optimal model, in this regard in [20], the author states that his predictive model was good since its general precision was 75.42% and an area under the ROC curve of 0.805. Likewise, in the investigation of [27] it is pointed out that the general result shows that each of the techniques used shows a good result in the classification and prediction performance, obtaining a greater precision of 86.9%.

On the other hand, the results of [26] showed a precision rate of 89.31% and a specificity rate of 91.25%, these measures are substantial to select classifiers since the researcher intends to minimize false negatives.

Regarding the term optimal model, in [4] it is pointed out that the so-called optimal models are combined with the dominant sets, which significantly improve the performance of prediction models and are highly influential in academic performance factors. Likewise, regarding the area on the curve, whose highest value in this research was 0.93 or 93%, in [4] it is indicated that an AUC of 50% of 91% or 99%, which was obtained in the research represents a better Classifier algorithm performance, favorable results for research.

The results of this study, from the perspective of innovation, will make it possible to achieve great changes, delegating functions, promoting competencies and fostering the continuous updating of higher institutions, all from the perspective of visionary leadership. In [10] it is pointed out that the proposed model accurately predicts the completion of the course and the performance of students in the university, thus allowing the organization to provide a better quality of service, since the satisfaction of the student depends on it student.

IV. Conclusion

The use of technological tools such as Machine Learning and its algorithms are supporting and strengthening decision-making from an administrative and academic point of view and in the educational sector. According to the results obtained, it is concluded that the metrics of the Machine Learning model through Ensemble Bagged Trees, applied to the predictive analysis of university teaching performance, present on average optimal values in their validation metrics such in their 4 classes, with a precision of 82.21%, a Sensitivity of 73.40%, a Specificity of 91.02% and an Accuracy of 90.63%. From the validation of the Machine Learning algorithm metrics, its implementation is viable and reliable in improving the performance of university teachers. Finding the 4 classes of the predictive model with relatively high values, the results allow establishing the grouping of engineering students who can achieve a level of satisfaction based on the indicators called predictors (indicators), through which the authorities of the higher institution can make timely decisions to improve the percentage of satisfied students in relation to university teaching performance.

Once the conclusions are presented, it can be noted that the present study achieved its purpose of determining the best performance model for the predictive analysis of university teaching performance, which is why it can be used as part of the procedure to improve the quality of the educational process. Because these results allow to have a relevant, reliable database that is obtained in less time compared to manual processes.

Acknowledgment

Thanks to the researchers who have contributed their knowledge in the development of this paper.

References

[1] E. Frank, M. Hall, L. Trigg, G. Holmes and I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics*, vol. 20, no. 15, pp. 79-81, 2004. DOI: 10.1093/bioinformatics/bth261.

[2] T. J. Fontalvo-Herrera, E. J. Delahoz and A. A. Mendoza-Mendoza, “Data Mining Application for the Classification of High Quality Accredited Industrial Engineering University Programs in Colombia,” *Technological information*, vol. 29, no. 3, pp. 89-96, 2018. DOI: http://dx.doi.org/10.4067/S0718-07642018000300089.

[3] S. Bayne, “Higher education as a visual practice: seeing through the virtual learning environment,” *Teaching in Higher Education*, vol. 13, no. 4, pp. 395-410, 2008. DOI:10.1080/13562510802169665.

[4] P. Sokkhey and T. Okazaki, “Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 8, pp. 492-502, 2020. DOI: 10.14569/IJACSA.2020.0110862.

[5] E. De-La-Hoz and L. Polo, “Application of Cluster Analysis Techniques and Artificial Neural Networks in the Evaluation of the Export Potential of a Company,” *Technological information*, vol. 28, no. 4, pp. 67-74, 2017. DOI: 10.4067/S0718-07642017000400009.

[6] D. Moonsamy, N. Naicker, T. T. Adeliyi and R. E. Ogunsakin, “A Meta-analysis of Educational Data Mining for Predicting Students Performance in Programming,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 2, pp. 97-104, 2021. DOI: 10.14569/IJACSA.2021.0120213.

[7] E. F. Ruiz, E. Moreno, E. A. Carmona and L. I. Garay, “Educational Tool for Generation and Analysis of Multidimensional Modeling on Data Warehouse,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 9, pp. 261-267, 2020. DOI: 10.14569/IJACSA.2020.0110930.

[8] D. Buenaño-Fernández, D. Gil and S. Luján-Mora, “Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study,” *Sustainability*, vol. 11, no. 10, pp. 1-18, 2019. DOI: https://doi.org/10.3390/su11102833.

[9] C. González, E. Elhariri, N. El-Bendary, A. Fernández and R. P. Díaz, “Machine learning based classification approach for predicting students performance in blended learning,” The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt. *Advances in Intelligent Systems and Computing*, vol. 407, pp. 47–56, 2016. DOI: https://doi.org/10.1007/978-3-319-26690-9_5.

[10] K. H. Susheelamma and K. M. Ravikuma, “Student risk identification learning model using machine learning approach,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 3872-3877, 2019. DOI: http://doi.org/10.11591/ijece.v9i5.pp3872-3879.

[11] R. Shadiev and M. Yang, "Review of studies on technology-enhanced language learning and teaching," *Sustainability, MDPI, Open Access Journal*, vol. 12, no. 2, p. 524, 2020. DOI: https://doi.org/10.3390/su12020524.

[12] P. Ramkissoon, L. J. Belle and T. Bhurosy, "Perceptions and experiences of students on the use of interactive online learning technologies in Mauritius," *International Journal of Evaluation and Research in Education*, vol. 9, no. 4, pp. 833–839, 2020. DOI: http://doi.org/10.11591/ijere.v9i4.20692.

[13] L. Medina, "Blended learning: Deficits and prospects in higher education," *Australasian Journal of Educational Technology*, vol. 34, no. 1, pp. 42-56, 2018. DOI: https://doi.org/10.14742/ajet.3100.

[14] L. Soria, W. Ortega and A. Ortega, "Teaching pedagogical performance and learning of university students in the Education career," *Praxis & Know*, vol. 11, no. 27, e. 303, 2020. DOI: https://doi.org/10.19053/22160159.v11.n27.2020.10329.

[15] F. E. Ceballos, J. E. Rojas, L. G. Cuba, L. P. Medina and A. R. Velazco, "Analysis of the quality of services in university centers", *University, Science and Technology*, vol. 25, no. 108, pp. 23-29, 2021. DOI: https://doi.org/10.47460/uct.v25i108.427.

[16] L. Moyan and S. Yawen, "Evaluation of Online Teaching Quality of Basic Education Based on Artificial Intelligence", *International Journal of Emerging Technologies in Learning*, vol. 15, no. 16, pp. 147-161, 2020. DOI: 10.3991/ijet.v15i16.15937.

[17] E. J. De-La-Hoz, E. De-La-Hoz and T. J. Fontalvo, "Methodology of Machine Learning for the classification and Prediction of users in Virtual Education Environments," *Technological information*, vol. 30, no. 1, pp. 247–254, 2019. DOI: http://dx.doi.org/10.4067/S0718-07642019000100247.

[18] F. Makombe and M. Lall, "A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 9, pp. 415-419, 2020. DOI: 10.14569/IJACSA.2020.0110949.

[19] H. Mushtaq, *et al.*, "Educational Data Classification Framework for Community Pedagogical Content Management using Data Mining," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 1, pp. 329-338, 2020. DOI: 10.14569/IJACSA.2019.0100144.

[20] E. Ayala, R. E. Lopéz, and V. H. Menéndez, "Predictive models of academic risk in computing careers with educational data mining," *Distance Education Journal*, vol. 21, no. 66, pp. 1-36, 2020. DOI: https://doi.org/10.6018/red.463561.

[21] V. Pedrero, K. Reynaldos-Grandón, J. Ureta-Achurra and E. Cortez-Pinto, "Overview of machine learning and its application in the management of emergency services," *Medical journal of Chile*, *vol.* 149, pp. 248-254, 2021. DOI: 10.4067/s0034-98872021000200248.

[22] S. Rajagopal, K. Siddaramappa and P. Panduranga. "Performance analysis of binary and multiclass models using azure machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 978-986, 2020. DOI: http://doi.org/10.11591/ijece.v10i1.pp978-986.

[23] A. D. Poernomo and S. Suharjito, "Indonesian online travel agent sentiment analysis using machine learning methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 113-117, 2019. DOI: http://doi.org/10.11591/ijeecs.v14.i1.pp113-117.

[24] R. El-Shawi, S. Sakr, D. Talia, and P. Trunfio, "Big data systems meet machine learning challenges: Towards big data science as a service," *Big data research*, vol. 14, pp.1-11, 2018. DOI: 10.1016/j.bdr.2018.04.004.

[25] D. Pratiba and G. Shobha, "RSECM: Robust Search Engine using Context-based Mining for Educational Big Data," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 12, pp. 39-51, 2016. DOI: 10.14569/IJACSA.2016.071206.

[26] R. Lottering, R. Hans and M. Lall, "A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 10, pp. 417-422, 2020. DOI: 10.14569/IJACSA.2020.0111052.

[27] W. D. Ahmad and A. A. Bakar, "Ensemble Machine Learning Model for Higher Learning Scholarship Award Decisions," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 5, pp. 303-312, 2020. DOI: 10.14569/IJACSA.2020.0110540.

[28] T. C. Smith, and E. Frank, "Introducing machine learning concepts with WEKA," *In Statistical genomics, Humana Press, New York, NY*, vol. 1418, pp. 353-378, 2016. DOI: 10.1007/978-1-4939-3578-9_17.

[29] Z. Ullah, F. Saleem, M. Jamjoom and B. Fakieh, "Reliable Prediction Models Based on Enriched Data for Identifying the Mode of Childbirth by Using Machine Learning Methods: Development Study," *J Med Internet Res.*, vol. 23, no. 6, p. 28856, 2020. DOI: 10.2196/28856.

[30] E. Carpaneto, G. Chicco, R. Napoli and M. Scutariu, "Electricity customer classification using frequency–domain load pattern data," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13-20, 2006. DOI: 10.1016/j.ijepes.2005.08.017.

[31] U. B. Chaudhry and C. I. Phillips, "UAV Aided Data Collection for Wildlife Monitoring using Cache-enabled Mobile Ad-hoc Wireless Sensor Nodes," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 10, pp. 6-17, 2021. DOI: 10.14569/IJACSA.2021.0121002.

Feature Extraction based Breast Cancer Detection using WPSO with CNN

Naga Deepti Ponnaganti[1]
Research Scholar
Department of Computer Science and Engineering
KLEF, Vaddeswaram, India

Dr. Raju Anitha[2]
Associate Professor
Department of Computer Science and Engineering
KLEF, Vaddeswaram, India

***Abstract*—The cancer reports of the past few years in India says that 30% cases have breast cancer and moreover it may increase in near future. It is added that in every two minutes, one woman is diagnosed and one expires in every nine minutes. Early diagnosis of cancer saves the lives of the individuals affected. To detect breast cancer in early stages, micro calcifications is considered as one key symptom. Several scientific investigations were performed to fight against this disease for which machine learning techniques can be extensively used. Particle swarm optimization (PSO) is recognized as one among several efficient and promising approach for diagnosing breast cancer by assisting medical experts for timely and apt treatment. This paper uses weighted particle swarm optimization (WPSO) approach for extracting textural features from the segmented mammogram image for classifying micro calcifications as normal, benign or malignant thereby improving the accuracy. In the breast region, tumour part is extracted using optimization methods. Here, Convolutional Neural Networks (CNNs) is proposed for detecting breast cancer which reduces the manual overheads. CNN framework is constructed for extracting features efficiently. This designed model detects the cancer regions in mammogram (MG) images and rapidly classifies those regions as normal or abnormal. This model uses MG images which were obtained from various local hospitals.**

***Keywords*—Breast cancer; microcalcifications; weighted particle swarm optimization (WPSO); Convolutional Neural Networks (CNNs) mammogram**

I. Introduction

Breast cancer is the most commonly found in women which causes deaths who are aged from 20 to 59. According to the Ministry of Health and Medical Education, it has become the most common disease in recent years in Iran [1]. Today, 88% of women diagnosed with breast cancer have a life expectancy of 10 years. In the United States, it has been reported that about 12% of women were identified during their lifetime, and were referred to as the second cause of women's death [2]. Diagnosing the disease at the earlier stages is important because in the early stages, cancer masses are restricted to the breast and the chance of surgical treatment in a less invasive manner is increased. The mortality rate is also decreased in the early stage [3]. Also, the use of classifiers such as artificial neural networks in various fields of engineering sciences is increasing to analyze the time series and various issues of classification. Due to the invention of techniques in the recent era for early diagnosis of breast cancer, the survival rate of the patients is improved. Now-a-days, X-ray mammography and MRI (Magnetic Resonant Imaging) techniques are widely utilized with few implications and limitations. X-ray is very harm due to the ionizing radiation and thus its contact with patients has to be only for very short duration. Conversely, MRI technique is expensive while mammography is of less cost, but hard to provide consistency and accuracy in analysing breast cancer [4]. Moreover, errors occur while analysis. To increase the rate of accuracy and reduce the occurrence of errors, supervised machine learning approaches like KNN, SVM, LSSVM are developed. These models efficiently classify the features as normal or abnormal classes. These methods are complex and even tedious with low CR. Therefore, to provide a solution for all the drawbacks of breast cancer, an optimal classification model is required for which machine learning approaches based on image processing are developed to classify cancer and non-cancer images which involved mammogram images. As the features are essential to discriminate breast cancer as benign or malignant, feature extraction process is of most important. Once the features are extracted, properties of the image like depth, coarseness, smoothness, and regularity are obtained with the help of segmentation process [5]. Scientifically, with breast cancer, division of tumor cells is uncontrolled and abnormal tumor cells need more nutrients for growing continuously and to reproduce. The cancer cells penetrate into the surrounding for gain in nutrients. There is a heterogeneous variation in the circulation of blood with various tumors and hence lesion morphology characteristics and ambiguity with edges in diagnosing images are significant in dicators for evaluation. The paramagnetic contrast agent spreads in blood which enters into the blood vessel and passes in the intercellular space as well as cells easily via penetrable capillary wall; hence the sputum concentration is high in the tumor rich region. This abnormality can be found using TIC when DCE-MRI is utilized for several imaging of the same tissue in various stages. Thus, edge, shape, etc. which are static characteristics and initial increase and change in signal which are dynamic characteristics of the lesion plays a major role in identifying the tumor as benign or malignant. MRI images are usually clear and complete with multi-angle, multi-faceted imaging. With the breast, surface coil has been used for clinical purpose, and MRI technology is improved to be much clear. However, the true positive rate and the true negative rate obtained while diagnosing breast cancer are also improved simultaneously [6].

Here this paper contributes weighted particle swarm optimization (WPSO) approach for extracting textural features from the segmented mammogram image for classifying microcalcifications as normal, benign or malignant thereby improving the accuracy. In the breast region, tumor part is extracted using optimization methods. Here, Convolutional Neural Networks (CNNs) is proposed for detecting breast cancer which reduces the manual overheads. CNN framework is constructed for extracting features efficiently. This designed model detects the cancer regions in mammogram (MG) images and rapidly classifies those regions as normal or abnormal.

The remaining part of this work is presented as follows. An outline of relating works is discussed in Section 2; Section 3 elaborates the proposed methodology while Section 4describes the experiments and discusses the obtained results. Finally, Section 5 concludes the work with future improvements.

II. Related Work

This section discusses few related works carried out for diagnosing breast cancer which involved various optimization techniques. It is well known that breast cancer is one serious and dangerous cancers among women and hence diagnosing at the earlier stages is more effective to provide treatment and protect the lives of patients. Till now, several approaches are coined for detecting breast cancer which addresses different sorts of challenges and few of them are reviewed here. Asri et al. [7], in 2016, employed machine learning methods for the prediction and classification of WBC actual dataset. Various classifiers were used which includes SVM, Naive Bayes, KNN and decision tree C4. SVM produced an accuracy with Weka tool. In [8], Chowdhary et al. utilized mammography images for the detection of breast cancer using intuitive fuzzy histogram magnification approach there by data was processed and image quality was improved. Then, probabilistic Fuzzy Clustering approach was employed for segmenting and separating the cancer tissues. Hence, this model was suitable for processing larger cancer datasets with the objective to offer better accuracy. Next, with the methods like grey area coefficient and linear binary pattern, textural properties were extracted. The accuracy obtained was 94% but hard while dealing with larger datasets and extends the processing time. Aalaei et al. [9] employed genetic meta-specificity reduction for classifying breast cancer. Three datasets namely WBC, WDBC and WPBC were used for evaluating which used Artificial Neural Network (ANN) cluster. The accuracy estimated for the method used with WBC, WDBC and WPBC datasets were 96, 96.1 and 76.3, respectively. Even though feature set was reduced, accuracy could be improved. Nilashi et al. [10], in 2017, designed a knowledge-based system which involved fuzzy logic. The process was carried out in three steps: initially Wisconsin Breast Cancer data was processed. Then, data with similar groups was clustered by the use of Expectation Maximization (EM) clustering technique. Finally, once the features were reduced by PCA, fuzzy rule set was categorized as data by means of regression tree. The accuracy obtained was 93.2%. Sometimes when learning rules are applied on datasets, the classification task is complicated. In [11], the use of Bat algorithm selected optimal features for diagnosing breast cancer. 286 samples were selected from WDBC dataset for which simple random sampling approach was involved for feature selection. After selecting the features, according to the classification similarity which involved Random Forest (RF), overall ranking was performed and obtained an accuracy. As samples are selected at random, selection of features was sometimes difficult. In [12], Dore swamy et al. improved Bat algorithm to classify breast cancer images. 569 samples of UCI data were involved in experimenting this method. The accuracy for training set was 92.61 while that of the testing was 89.95. In [13], an approach using PSO was utilized for reducing the specificity in diagnosing breast cancer. The objective was to estimate the level of breast cancer. 699 pre-processed samples of UCI data after reducing the specificity were used by PSO algorithm along with decision tree C4.5 to classify the samples into two classes namely malignant and benign. The accuracy achieved was 95.61%. Sahu et al. [14] used a hybrid approach for classifying and diagnosing breast cancer. With PCA feature reduction and various clusters, it was found that ANN classification produced 97% of performance than other clusters. 699 samples with 9 features were used in the experiment to label them as benign and malignant. Even though results achieved are better, every method has few weaknesses and limitations. In [15], Gao et al. integrated shallow CNN with deep CNN and formulated SD-CNN. Shallow CNN was used with the intention to extract “virtual” recombination of images which has lower energy, while deep CNN extracted the novel features related to LE. Additionally, knowledge of nonlinear mapping was gathered from LE for recombining images; shallow CNN and deep-CNN was created with 49 CEDM each. The performance was enhanced in terms of AUC and was accurate than the methods existing. In [16], Ting et al. developed CNNI-BCC which helped medical experts in diagnosing breast cancer at the earlier stage. This model enhanced the classification using CNN and the breast cancer images were classified as various types like benign, malignant, and healthy. From the numerous experiments conducted, this model enhances accuracy, sensitivity as well as AUC.

Diagnosing and classifying approaches involved for breast cancer were not tested and evaluated with three various data sets of breast cancer. In [17], Araujo et al. proposed a method for the classification of hematoxylin and eosin stained breast biopsy images using CNNs. High Sensitivity for Carcinoma cases are obtained from this classification scheme. In [18], Belsare et al. Classification of image is performed on different classes using LDA. This proposed approach is helpful for Pathologist. In [19], Tan et al. Developed a new version of breast cancer detection by using CNN, tested on mammogram images. This shows the improved accuracy results. In [20], Henry et al. Effectiveness of CNN was analyzed based on the existence of breast abnormalities in mammograms. In [21], Khan et al. developed a method of cancers classification for relevant genes using ANNs. In [22], Dhungel et al. introduce a novel automated CAD system with minimal user. Deep learning and structured output models are explored, proposed CAD system results for INbreast data set. In [23], Al-antari et al. proposed CAD system based deep learning, deep convolutional network (CNN) is used to recognize the mass

and classify it as either benign or malignant. In [24], Al-antari et al. proposed an integrated CAD system of deep learning detection and classification, develop a Cad system for practical breast cancer diagnosis. In [25], Fang et al. constructed CNN to utilize three dimensional spatial correlations information of breast tumors very effectively. In [26], Aya et al. introduce a new automatic segmentation Method (SM) for identifying the ROI from breast thermograms. In [27], Hu et al. provide an overview on deep learning and hope on the survey for cancer detection and diagnosis. The specialties of the present investigation are reduction is detecting costs, using better classifier with no adverse effects of aggressive approaches, higher accuracy of detection than the paper cited, choosing titles appropriate with the data available and comprehensive comparison with the researches made so far.

III. Proposed Methodology

The work flow of the methodology developed is illustrated in Fig. 1. The phases like pre-processing, segmentation and feature extraction are discussed below. CNN classifier is involved to obtain the accuracy of classification.

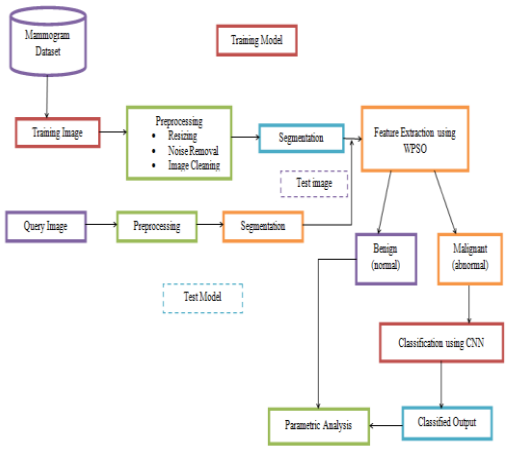


Fig. 1. Architecture of the Proposed Technique.

A. Pre-processing Steps

Step 1: Looking for an input Breast Image.

Step 2: The raw image provided as input raw undergoes resizing to 256 x 256.

Step 3: When 3-dimensional (3D) images are provided as input, they are converted to 2D, since mostly image processing is carried out only with 2D images i.e, RGB image is converted into gray scale image.

Step 4: Two filtering techniques are applied for de-noising as described below:

Step 4.1: Out_1 = Laplacian filter is applied on the gray scale image .

Step 4.2: Out_2 = Then mean filter is applied on the gray scale image.

Step 4.3: Out_3 = Out_1 – Out_3.

Step 4.4: The final output of the pre-processing stage is the pre-processed breast imageOut_3.

Step 5: Out_3.

B. Segmentation Steps

Input: Out_3 Pre-processed Image.

Step 1: Gradient is obtained along X and Y axis in variables Out X and Out Y.

Step 2: Gradient values are combined to obtain gradient vector G val which is given by.

G val=[1/(1+(Out X+ Out Y))]

Step 3: G val obtained in radians is converted to degrees so that orientation information of image pixels can be attained.

Step 4: Out_3 image is partitioned to grids GRi.

Step 5: Threshold values are defined for intensity Ti and orientation To.

Step 6: for every grid GRido.

1) Histogram Hi for every pixel Pjis computed over grid GRi.

2) 6.2. Most frequent histogram of grid GRiis found which is represented by Freq H.

3) 6.3. Any arbitrary pixel Pjis selected which is related to Freq H which is then assigned to pixel information seed point (SP) with Intensity Ip and Orientation Op.

4) 6.4. Intensity along with orientation constraints for adjacent pixel is verified.

5) 6.5. When both constraints are fulfilled, then decided that region is grown, or else.

Next grid is considered for further process.

Step 7: Output: Segmented Image

C. Weighted PSO based Feature Extraction

A heuristic global optimization technique namedWeighted Particle Swarm Optimization (WPSO) algorithm simulates the social behaviour of flock in g bird towards a position for attaining the exact objective in a multidimensional space. This approach involves a population of particles (called swarm) in the search space. For every particle, the status is categorized based on its location $\vec{x}_i = \{x_{i1}, x_{i2}, \ldots . x_{id}\}$and the velocity of particle i is given by $\vec{v}_i = \{v_{i1}, v_{i2}, \ldots . v_{id}\}$.To find the optimal solution, every particle deviates from its actual searching direction to a new direction based on two concepts namely the best location of the given particle (pbest) and the one obtained so far by swarm (gbest). WPSO identifies the optimal solution after velocity and position of every particle are updated in relation with the equations,

$$v_{id}^{(t+1)} = wv_{id}^{t} + c_1 r_1 (p_{id} - x_{id}^{t}) + c_2 r_2 (p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id}^{(t+1)} = x_{id}^{t} + v_{id}^{(t+1)} \quad (2)$$

where t and d represent the iteration in the evolutionary space and dimension in search space respectively. W denotes the weight of inertia. c1 and c2 represent the personal and

social learning factors. r1 and r2 are uniformly distributed random values ranging between 0 to 1. pid and pgd denotes pbest and gbest in the dimension d.

The basic steps performed in WPSO algorithm are as described below:

- Initialization: Random positions and velocities are used to initialize the particles.
- Evaluation: For every particle, value of the objective function is estimated.
- Finding pbest: When the value obtained with the objective function is better than the p best for particle i, then the current value is assigned as the new p best.
- Finding gbest: When p best is better than gbest, then gbest is assigned to the current value.
- Updating the position and velocity: For every particle, velocity is updated using Equation 1, and the particle is moved to the next position based on Equation 2.
- Terminating Criteria: When the required number of iterations are reached, the process ends or else repeated from step 2.

For the search space, exploration and exploitation are controlled by weight as velocity is adjusted dynamically. Moreover, weight controls the impact of the previous velocities on the current one. Thus, the exploration capabilities are compromised between global and local swarm. Larger weight simplifies global search for new areas while the smaller weight simplifies local search. When the weight is chosen properly, global and local exploration of swarm is balanced providing better solution. Hence, weight can basically set to a larger value for better global exploration of the search space and then decrease it gradually to obtain refined solution. When the weight decreases linearly, exploration from global to local change linearly. Search algorithm are required to have non-linear searching ability. With few statistical features obtained, PSO search is easily understood and the suitable weight is calculated for the next iteration. Here, when there is an increase in total generation, there is a linear decrease in weight w while optimization in relation to

$$W = w_{max} - \left(\frac{w_{max} - w_{min}}{iter_{max}}\right) * iter \quad (3)$$

Where w_{max} andw_{min}denote the maximum and minimum inertia weight respectively, $iter$and $iter_{max}$are the current iteration and maximum number of iterations respectively. For particle i, the best position is position that the particle visited (past value of Xi), which provides highest fitness value. For minimization, a position with small function value is considered to have fitness. f(X) denotes the minimized objective function for which the updated equation is

$$P_{bestid}^{(t+1)} = \{x_{id} t \; if \; f(x_{id}(t+1) \geq P_{bestid}^{t} \text{or} P_{bestid}^{(t+1)} = \{x_{id}(t+1) \; if \; f(x_{id}(t+1) \geq P_{bestid}^{t} \quad (4)$$

A faster rate convergence is provided by the g best with the expense of robustness and only a single best solution is maintained termed as global best particle. The role of this particle is to act as an attractor and hence pulls every particle towards it. Ultimately, every particle converges at this position and thus has to be regularly updated if not swarm converges prematurely. For every particle in the swarm, fitness value is computed using the objective function. Then, Pid and Pgd values are evaluated and updated with the global best position or better particle best position if obtained.

Steps for WPSO:

Initialize the function.

Create objective function.

Objective function is based on intensity of pixel.

Set iteration count = 1000.

Calculate pixel intensity of images.

Optimize the cancer image pixel intensity.

Calculate optimal value for input image pixel.

Extract tumor part with maximum pixel intensity.

Calculate accuracy.

D. Classification using Convolution Neural Networks

CNN takes the breast cancer image dataset an input for classification. Then, deep convolutional kernels are trained using the introduced CNN architecture. RELU nonlinearity is used in convolution layers and are defined as:

$$f(x) = \begin{cases} x, if \; x > 0 \\ ax, otherwise \end{cases} \quad (5)$$

Generally, convolution layer is stated as:

$$y^j = fb^j + \sum_i k^{ij} + x^i \quad (6)$$

Here, xi represents the ith input map and yi denotes the j^{t} ^{h}output map. bjis the bias parameter of jth map, convolution process between two functions is given by *, and convolutional kernel involved between i and j maps is bij. Max-pooling layer was the next layer following the convolutional layer. In max-pooling layer, every neuron provides yi pools in the output map yi pools against s * s non-overlapping areas of xi. In general, max-pooling layer is defined as:

$$y_j^i = max_{0 \leq m \leq s}\{x_{j.s+m}^i\} \quad (7)$$

Convolutional as well as max-pooling layers are fully connected which is followed by Softmax classifier containing output classes which equals the number of outputs. In the architecture introduced, tan h is used as a non-linear protocol in connecting one layer with another. The function of Softmax function equals squashing, and dataset with k-dimension is re-normalized producing real values ranging from 1 to 2. This is represented mathematically as:

$$\sigma(z)_j = \frac{e^{zj}}{\sum_{k=1}^{k} e^{zk}}, for \; j = 1, \ldots k \quad (8)$$

Error obtained while developing ML approaches are training and generalization errors. The former is observed

while training the neural network, where the latter is produced while testing the proposed classifier. In the process of deep learning, training is frequently affected with the process of overfit and under-fit. To surpass these issues, after every layer, batch normalization is applied in the proposed architecture for BCC. Dropout layer was added next to the first fully connected layer. The entire architecture developed for breast cancer classification is illustrated in Fig. 2.

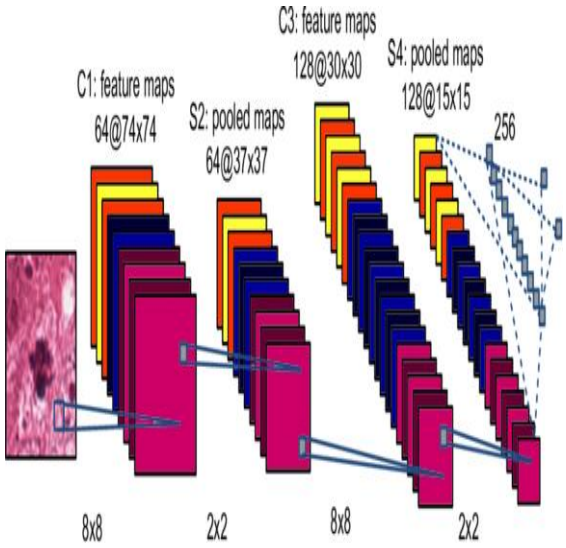


Fig. 2. Proposed CNN Architecture for Breast Cancer Classification.

E. Training of CNN

The proposed CNN architecture has two classes namely benign and malignant. Weighted loss function was employed for training the proposed CNN classifier.

$$\xi(w, x_n, y_n) = -\frac{1}{N}\sum_{n=1}^{N} \alpha_n \sum_{k=1}^{k} t_{kn} ln y_{kn} \quad (9)$$

Here xn represents input vector, yn the prediction obtained from classifier for n^{th} clinical input, and tn its actual response. K and N are the number of classes total clinical samples.

For recognizing, patch results of the entire image are combined. As the model is trained with image patches, strategy is necessary for partitioning the actual testing images into patches, then executing and combining the results obtained to get optimal result but is computationally too complex. Rather, grid patches are obtained from the images which provide the set of non-overlapping patches, and this was reasonable and balanced the performance of classification as well as computational cost. By implementing this model, every patch produced the probability of every possible class for the given patch of the image. For combining the results produced by the patches for the test image, three various fusion rules were involved and found that Sum rule produced better results.

IV. Performance Analysis

Detection of breast cancer in the earlier stages is critical for treatment and managing its condition. This study presented a detailed derivation methods and processes along with way it is applied in detecting tumor. According to the tissue segmenting method, the effects of the obtained number of glandular tissues is analysed. It is observed that presence of numerous glandular tissues worsen the imaging effect later. Simultaneously, a progressive approach to detect multiple tumors is also introduced. Imaging is done in three steps: preliminary examination, refocusing, and image optimization by which every tumor is detected successfully. Here, WPSO-CNN is used for extraction of features and classification of tumor, and this has obtained enhanced accuracy. Features were obtained and classified the image of histopathology. The below Fig. 3 shows extracted feature of histopathology image using WPSO-CNN. The Fig. 4 shows classified image with malignancy.

The accuracy, precision, recall and F-1 score graph has been shown below in Fig. 5,6,7,8. The below graphs show comparison of parameters between existing and proposed techniques.

From analysis of results obtained for proposed technique, it is observed that performance is significantly improved. The analysis of results expressed that proposed technique exhibits its significant performance in Classification.

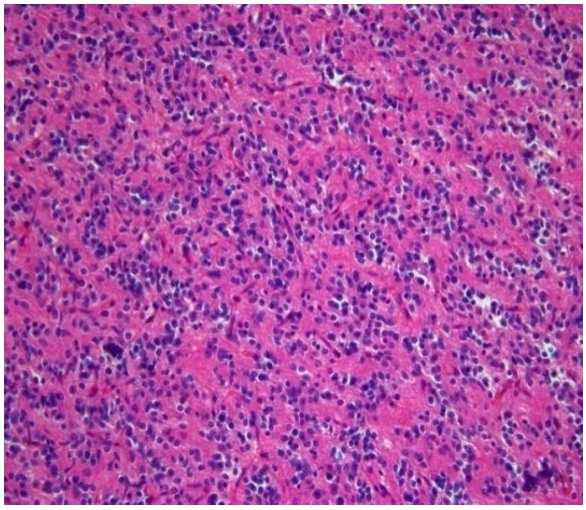


Fig. 3. Feature Extracted Tumor Image of Histopathological Image.

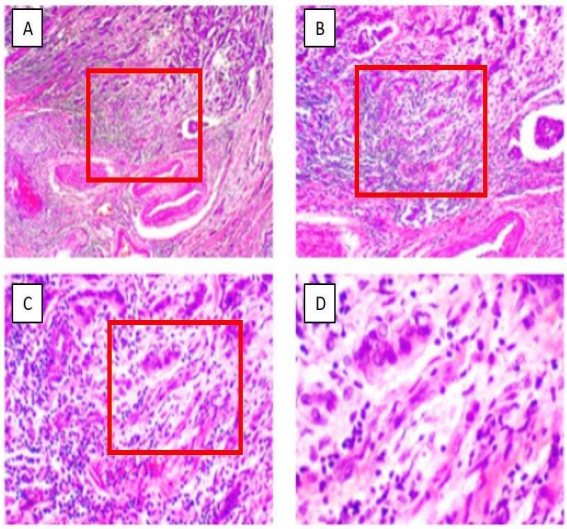


Fig. 4. Classified Image of Histopathology Detecting Malignancy.

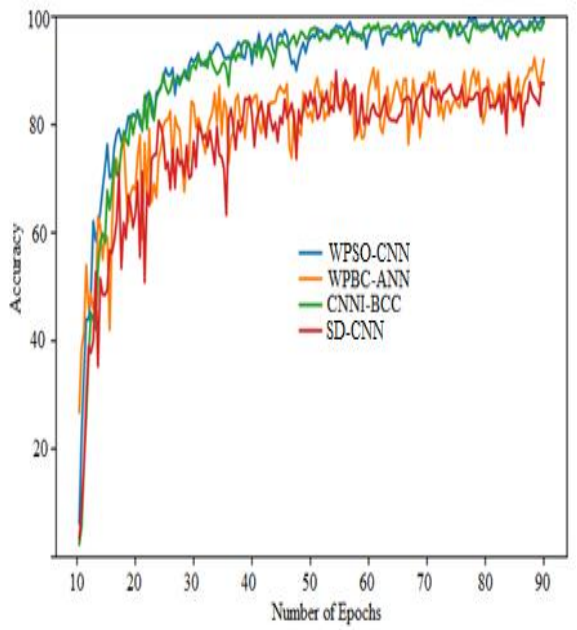


Fig. 5. Comparison of Accuracy.

In Fig. 5, shows the classification of accuracy analysis for proposed technique. The accuracy measured for proposed technique is measured significantly higher than the existing techniques.

In Fig. 6, shows the classification of precision analysis for proposed technique. From overall comparison of proposed technique with existing classifiers is presented. The measurement of precision provides the analysis is expressed that the proposed technique provides improved performance rather than existing techniques.

In Fig. 7, shows the classification of Recall analysis for proposed technique. The Recall measurement provides the analysis and the comparative analysis expressed that the proposed technique provides improved performance than the existing classification techniques.

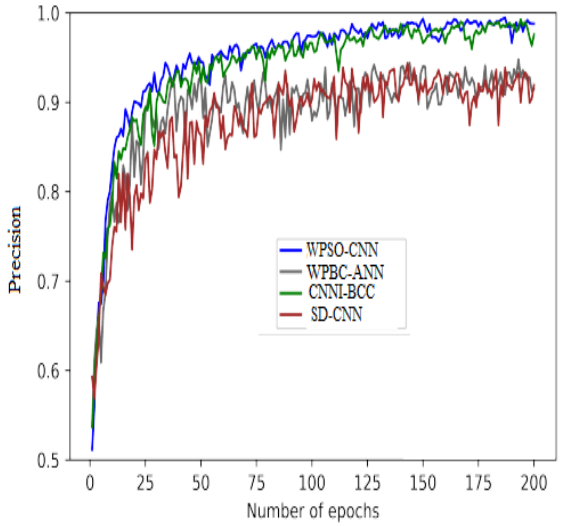


Fig. 6. Comparison of Precision.

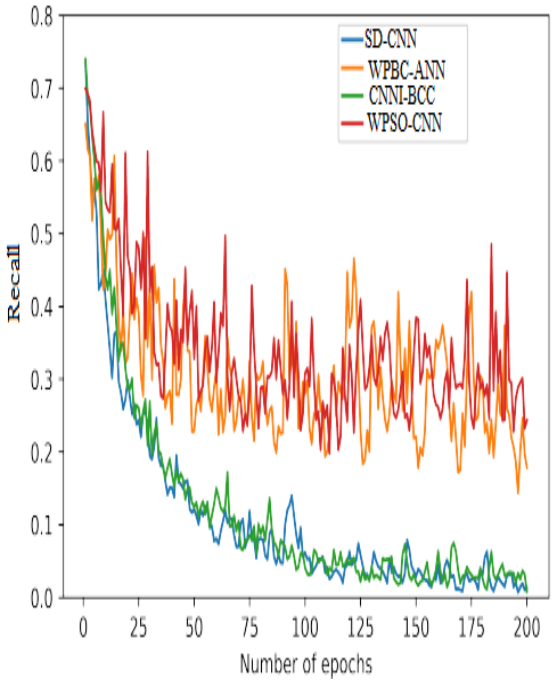


Fig. 7. Comparison of Recall.

In Fig. 8, shows the classification of F1-Score analysis for proposed technique. The F1-Score is measured and provides the analysis and through analysis, it is concluded that the proposed technique exhibits improved performance rather than the existing classification technique.

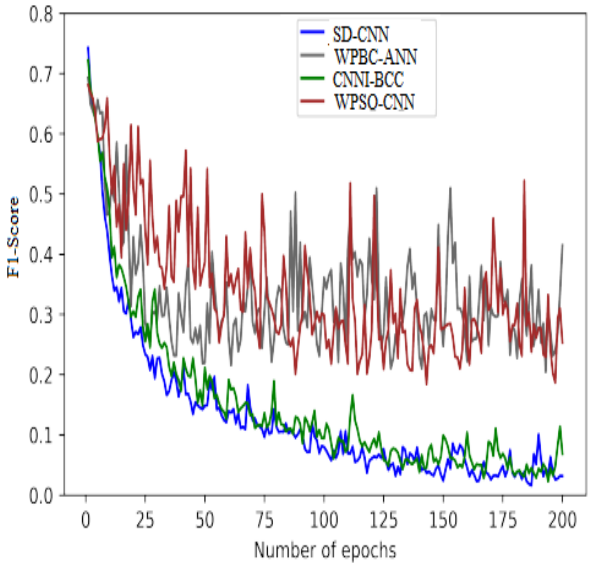


Fig. 8. Comparison of F-1 Score.

Limitations of the proposed WPSO-CNN:

- WPSO-CNN is best suitable for image-classification, dealing with high-dimensional Data (images).
- WPSO-CNN is not best suitable for smaller data sets.
- WPSO-CNN becomes slower, if there are more layers.

V. CONCLUSION

The objective to carry out this research is to improve accuracy of detection using CAD technique for detecting breast cancer. With this objective, a framework was contributed along with its flow and parameters used for simulation. Publicly available dataset is involved for analysing the effectiveness of the method for classifying normal and abnormal breast images of several individuals. Here, weighted particle swarm optimization (WPSO) with CNN (Convolutional Neural Networks) is employed named as WPSO-CNN with the objective to extract the features and estimate the error between the estimated and true density using kernel density estimation based classifier for diagnosing breast cancer. From the results it is observed that the performance of WPSO-CNN is remarkable than existing approaches. The future work is to possibly develop an online breast cancer detection system since the detecting systems used currently are offline.

REFERENCES

[1] Ganggayah, Mogana Darshini, "Predicting factors for survival of breast cancer patients using machine learning techniques", BMC medical informatics and decision making, vol.19, no.1, pp.1-17, 2019.

[2] Houssein, Essam H, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review", Expert Systems with Applications, 2020.

[3] Beura S, Majhi B, Dash R, Roy S,"Classification of mammogram using two-dimensional discrete orthonormal S-transform for breast cancer detection",HealthcTechnolLett.,vol.2, no.3, pp.46-51, 2015.

[4] Hamian M, Darvishan A, Hosseinzadeh M, Lariche MJ, Ghadimi N, Nouri A,"A framework to expedite joint energy reserve payment cost minimization using a custom-designed method based on mixed integer genetic algorithm" ,Engineering Applications in Artificial Intelligence,pp.203-212, 2018.

[5] Leng H, Li X, Zhu J, Tang H, Zhang Z, Ghadimi N,"A new wind power prediction method based on ridgelet transforms, hybrid feature selection and closed-loop forecasting", Advances in Engineering Information,pp.20-30,, 2018.

[6] Qi X, Zhang L, Chen Y, "Automated diagnosis of breast ultrasonography images using deep neural networks",Medical Image Analysis,pp.185-198, 2019.

[7] Asri H, Mousannif H, AlMoatassime H, Nodel H,"Using machine learning algorithms for breast cancer risk prediction and diagnosis",Procedia Computer Science, pp.1064-1069, 2016.

[8] Chowdhary C.L, Acharjya D.P,"Breast cancer detection using intuitionistic fuzzy histogram hyperbolization and possibilitic fuzzy c-mean clustering algorithms with texture feature based classification on mammography images,Proceedings of the International Conference on Advances in Information Communication Technology & Computing, pp. 1-6, 2016.

[9] Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S,"Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets",Iran Journal of Basic Medical Sciences,vol.19, no.5, pp.476-482, 2016.

[10] Nilashia M, Ibrahima O, Ahmadic H, Shahmoradi L,"A knowledge-based system for breast cancer classification using fuzzy logic method",Telematics and Informatics, vol.34, no.4, pp.133-144, 2017.

[11] Jeyasingh S, Veluchamy M,"Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset",Asian Pacific Journal of Cancer Prevention, vol.18, no.5, pp.1257-1264, 2017.

[12] Doreswamy H, Salma M.U,"A binary bat inspired algorithm for the classification of breast cancer data",International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), pp.1-21, 2016.

[13] Muslim M, Hardiyanti S, Sugiharti E, Prasetiyo B, Alimah S,"Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis",International Conference on Mathematics, Science and Education 2017 (ICMSE2017), pp.1-7, 2016.

[14] Sahu B, Nandan S, Mohanty Kumar Rout S,"A hybrid approach for breast cancer classification and diagnosis",EAI Endorsed Transactions on Scalable Information Systems, pp.1-8, 2019.

[15] Gao F, Wu T, Li J, Zheng B, Patel B,"SD-CNN: a shallow-deep CNN for improved breast cancer diagnosis",Computing Medical Imaging Graph,pp.53-62, 2018.

[16] Ting FF, Tan YJ, Sim KS,"Convolutional neural network improvement for breast cancer classification",Expert Systems and Applications,pp.103-115, 2019.

[17] Arau´jo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. (2017) Classification of breast cancer histology images using Convolutional Neural Networks. PLoS ONE 12(6): e0177544. https://doi.org/10.1371/journal. pone.0177544.

[18] Belsare, A. D.; Mushrif, M. M.; Pangarkar, M. A.; Meshram, N. (2015). [IEEE TENCON 2015 - 2015 IEEE Region 10 Conference - Macao (2015.11.1-2015.11.4)] TENCON 2015 - 2015 IEEE Region 10 Conference - Classification of breast cancer histopathology images using texture feature analysis. , (), 1–5. doi:10.1109/TENCON.2015.7372809.

[19] Tan, Y. J.; Sim, K. S.; Ting, F. F. (2017). [IEEE 2017 International Conference on Robotics, Automation and Sciences (ICORAS) - Melaka, Malaysia (2017.11.27-2017.11.29)] 2017 International Conference on Robotics, Automation and Sciences (ICORAS) - Breast cancer detection using convolutional neural networks for mammogram imaging system. , (), 1–5. doi:10.1109/ICORAS.2017.8308076.

[20] H. Zhou, Y. Zaninovich, and C. Gregory, "Mammogram Classification Using Convolutional Neural Networks", Vaialable:http://ehnree.github.io/documents/papers/mammogram_conv_net.pdfLast Accessed: 02 October 2017.

[21] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, and P.S. Meltzer, " Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature medicine,vol. 7(6), pp. 673, 2001.

[22] Dhungel, N., Carneiro, G., Bradley, A.P., 2017. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med. Image Anal. 37,114–128. https://doi.org/10.1016/j.media.2017.01.009.

[23] Al-antari, Mugahed A.; Al-masni, Mohammed A.; Choi, Mun-Taek; Han, Seung-Moo; Kim, Tae-Seong (2018). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. International Journal of Medical Informatics, 117(), 44–54. doi:10.1016/j.ijmedinf.2018.06.003.

[24] Al-antari, Mugahed A.; Han, Seung-Moo; Kim, Tae-Seong (2020). Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Computer Methods and Programs in Biomedicine, 196(), 105584–. doi:10.1016/j.cmpb.2020.105584.

[25] Fang, Yan; Zhao, Jing; Hu, Lingzhi; Ying, Xiaoping; Pan, Yanfang; Wang, Xiaoping (2019). Image Classification toward Breast Cancer using Deeply-Learned Quality features. Journal of Visual Communication and Image Representation, (), 102609– doi:10.1016/j.jvcir.2019.102609.

[26] Hossam, A., Harb, H. M., & Abd El Kader, H. M. (2018). Automatic image segmentation method for breast cancer analysis using thermography. Journal of Engineering Sciences, 46 , 12–32.

[27] Hu, Zilong; Tang, Jinshan; Wang, Ziming; Zhang, Kai; Zhang, Lin; Sun, Qingling (2018). Deep learning for image-based cancer detection and diagnosis − A survey. Pattern Recognition,83(),134–149. doi:10.1016/j.patcog.2018.05.014.

Real-Time Emotional Expression Generation by Humanoid Robot

Master Prince
Assistant Professor, Department of Computer Science
College of Computer, Qassim University, Buraidah, 51452, Al Qasim, Saudi Arabia

***Abstract*—Emotion integrates different aspects of a person, including mood (current emotional state), personality, voice or speech, color around the eyes, and facial organs' movement. We are considering the mood because a person's current emotional state must always affect upcoming emotions. So behind an emotion, all these parameters are involved, and a human being can easily recognize it by seeing that face even if more than one person is there, so for the robot to make human-like emotion, all these parameters have to be considered to imitate artificial facial expression against that emotion. Most researchers working in this area still find difficulties in determining exact emotion by the robot because facial information is not always available, especially when interacting with a group of people and mimicking exact emotion that the user can effectively recognise. In our study, the loud most speeches among the people sensed by the robot and color around eyes are considered to cope with these issues. Another issue is the rise time and fall time of emotional intensity. In other words, how long should the robot keep an emotion here? An experimental approach is applied to get these values. The proposed method used an emotional speech database to recognize the human emotion using convunational neural network (CNN) and RGB patterns to mimic the emotion, which simulates an improved humanoid robot that can express emotion like human beings and give real-time responses to the user or group of users that can make more effective Human-Robot Interaction (HRI).**

Keywords*—*Artificial facial expression; emotional speech database; convunational neural network; RGB pattern; humanoid robot; human-robot interaction

I. Introduction

The human face is very special in different aspects; one of those aspects is expressing emotion. By expressing emotion, human beings express their feelings, and others can easily understand the feeling and respond as well [Y. Yang et al., (2007)]. However, when we talk about HRI, it becomes challenging for the humanoid robot to determine the exact emotion expressed by the person (human) who is interacting at a particular moment, especially when interacting with a group of people and when emotion is not evident with the face because according to psychologist numerous types of expression can be produced by a human. Moreover, we have only seven recognized expressions: Natural, Happy, Sad, Anger, Surprise, Fear and Disgust.

Robotics has become a very emerging area in today's world; it plays a significant role in various fields like medical science, military applications, home appliances, education, and many more. In recent years, a popular research area in robotics has been developing intelligent robots that can interact with people as companions rather than machines. To interact with a humanoid robot, HRI is very important; studies of human-robot interaction will be improved by automated emotion interpretation. A humanoid robot must be able to understand the person's actual emotional state at a particular instance.

Here the proposed method used our previous emotion recognition method that represents the intensities of the emotions instead of emotion. Once the intensity of the emotion was known, the main goal was to determine fusion weight for each primary emotion based on all those parameters, which include mood, personality, and intensities of the recognized emotional states employing fuzzy Kohonen clustering network (FKCN), which give us a smooth variation of facial expressions. Finally, the control point's vector is used to mimic the artificial face simulator [Prince. M. (2017)].

The parameter is determined as; for the mood previous emotional state has been buffered, for personality Big Five model of personality has been considered [Power R. A. et al. (2015)], for Euclidean intensity distances between the feature vector of standard and user emotion are proposed being used, for speech the training data set of data statistic and machine learning based on Ultra-large-scale database of natural language are used [McGilloway et al. (2000); Greasley (2000); Mohamad Nezami et al. (2019)], and for the colour around eyes RGB color patterns for different emotions have been used [Johnson et al. (2013)].

One of the objectives of most of the research is to improve the life of human beings. So in this regard studying human-machine behaviour becomes essential. Moreover, Human-Computer Interaction (HCI) is one of the areas under which humans and machines should have better communication skills. Ultimately, when we communicate with the humanoid robot, we want to communicate as a companion rather than as a machine closer to human nature.

Here, we are proposing a humanoid robot model that can emotionally respond and a human being. The robot can recognize the user emotional state and would respond accordingly. That could mean that if the user is happy, then the robot should behave like if it is also happy, which would improve the interaction between a human and a machine. The problem arises for the robot to communicate with a user whose expression is not clear on his face, especially when people are involved.

The rest of the paper organized as the related work section is following this section then the complete methodology is presented, the result and discussion section is presented following the methodology and finally the work was concluded.

II. Literature Review

Most research on robotics heads mimicking human facial expression is done in some universities and research institutions of the United States, Japan, and the European Union. A robot called Kismat is one of the examples of it, developed by Cynthia Breazeal at the Massachusetts Institute of Technology. Waseda University, Japan, has developed a series of robots named WE-R since 1996 [Hiroyaus Miwa, (2001)].

In the recent decade, many researchers have been trying to recognise human beings' facial expressions automatically. Various pattern recognition methods have been used in order to recognize facial expressions. As in our previous works [Master Prince, (2013a); Master Prince (2013b)], the novel approach has been introduced to recognize facial expressions. As discussed in [Young. A. W. et al., (1989); Padgett. C. et al. (1997)] reported that approaches to emotional robot design often adopted results from psychology to design robot behaviours to mimic human beings. Miwa et al. proposed a mental model to build the robotic emotional state from external sensory inputs [Miwa. H.et al., (2003), Miwa. H.et al. (2004)]. Duhaut presented a computational model which includes emotion and personality in robotic behaviours [Duhaut. D. (2008)]. Moshkina et al. give a model of time-varying effective response for humanoid robots based on the Traits, Attributes, Moods and Emotions [Moshkina et al. (2011)]. One of the most important aspects is the robot mood transition from the current to the next mood state, which influences the robot's interaction behaviour and a user's feeling. Meng-Ju Han et al. introduced an effective model to make transition among mood states would become smoother and thus might enable a robot to respond with more natural emotional expressions [Meng-Ju Han et al. (2013)]. Imitating emotions with RGB patterns has previously been proposed with other types of robots. For example, Kanoh et al. asked 50 people to identify which of Ekman and Friesen's six basic emotions the Ifbot was imitating with its 29 pre-programmed facial expressions [Ekman P. et al. (1969), Kanoh. M. et al. (2005)]. Angelica Lim and Hiroshi G. Okuno proposed speech and gait analysis to recognize human emotion [Angelica Lim et al. (2012)].

Previous researches have shown powerful tools for designing emotional robots. It is observed that exact emotion recognition and exact mimicking the artificial emotion plays a vital role in effective HRI. These representations lack a theoretical basis to support the assumptions in their emotion recognition design and simulation of the artificial face. That motivates me to investigate an effective emotion recognition system and an excellent artificial face simulator. Emotion recognition system by recognizing the pattern of the facial muscles is not enough. The speech recognition and speech synthesis function modules are embedded as emotion recognition procedures [Jianfeng *et al*. (2019); Mehmet *et al.* (2020)]. The combination of control point vectors and RGB patterns is used to imitate the artificial emotion with a smooth mood transition and get back to its normal intensity state of the current emotion after showing its actual intensity. Questionnaire surveys were conducted to evaluate the effectiveness of the proposed method.

III. Proposed Work

Our previous work proposed a model artificial brain emotion recognition and generation system (ABERGS) [Prince. M. (2017)]. This work is an extension of ABERGS, where facial expression is fused with the voice to simulate artificial expression. In our proposed work, the loud voice recognized by the robot were classified as recognized emotion using the trained CNN model [Jianfeng et al. (2019)] and RGB patterns were used to enhance artificial expression [Johnson et al. (2013)]. The purpose of the RGB pattern is to use it around the robot's eyes to express exact emotion.

In order to find out RGB color patterns against each specific emotion, an experiment was set up.

A. Speech Emotion Recognition

Fig. 1 illustrates the layers which are substituted to the CNN (feature learning block (FLB) and LSTM (Long short-term memory). Four FLBs extract the low-level features of speech, such as emotional features, and LSTM can learn the high-level features, which contain both the local information and the long-term contextual dependencies.

The pretrained model [Jianfeng *et al*. (2019)] was used as transfer learning to get trained with emotional speech dataset [McGilloway *et al.* (2000)], and the accuracy was outstanding, as shown through the confusion matrix in Fig. 2. The idea was to verify the accuracy of the model.

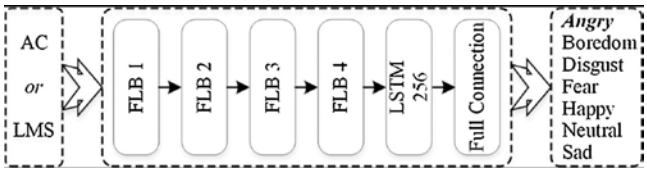


Fig. 1. Block Diagram of the Overall Architecture of the CNN [Jianfeng et al. (2019)].

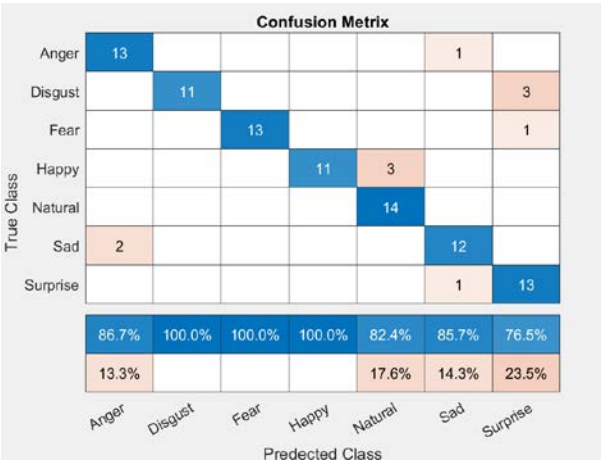


Fig. 2. Confusion Matrix Post-Training Model [Jianfeng *et al*. (2019)] with 98 Samples of [McGilloway *et al.* (2000)] Emotional Speech Dataset.

B. RGB Pattern Recognition

A movie clip against each major emotion (Natural, Happy, Sad, Anger, Surprise, Fear, and Disgust) has been selected. In a small theatre room, 20 participants are invited to watch the movies. A video (pointing to eyes) of each participant has been made against each movie clip. Now we have 7*20 videos (20 videos for each emotion). Now, each video is analyzed as below:

Measurements of the following attributes of each color (R, G and B) for each emotion.

- Intensities: An intensity of emotion.
- Duration: How long does emotion exist?

Intensity calculation: Here, the intensity is measured in terms of the RGB scheme. Each frame of the video is converted into a grayscale color scheme. Each color's intensities (R, G and B) can range between 0-255.

$$^{H}I_r = \sum_{f=1}^{20} r/20 \quad (1)$$

$$^{H}I_g = \sum_{f=1}^{20} g/20 \quad (2)$$

$$^{H}I_b = \sum_{f=1}^{20} b/20 \quad (3)$$

The above equations show the average intensity of each color pattern (R, G and B) against happy emotion. Here R, G and B value ranges from 0 to 255. The same intensities of each color pattern (R, G and B) for all other emotions can be calculated.

Duration calculation: The time duration between emotions initiated and coming back to the normal state. The RGB value for normal emotion has got 150, 150, and 0, respectively. It consists of two things;

- Rise time: Time is taken to rise from normal to peak value.
- Fall time: Time is taken to return to the normal from the peak value.

$$D_H = D_r + D_f \quad (4)$$

where D_H is the total duration of an emotion existing on the face, D_r is the time taken to rise from normal to peak, and D_f is the time taken to return to the normal from the peak value. Same as intensity, duration is also calculated on average.

After analysis, all the videos, follow Table I is obtained. Which shows duration, color (RGB), rise and fall time for all recognized emotions. Table I: RGB color pattern's Intensities, Duration, Rise time and Fall time.

C. RGB Pattern Evaluation

Table I from the previous experiment determined which pattern of color intensity and duration of the RGBs that humans associate with specific emotions. The purpose of this experiment is to examine whether humans recognize these RGB patterns as emotions or not. Two RGB patterns for each emotion are based on Table I and glow through the computer screen, as shown in Fig. 3.

TABLE I. RGB COLOR PATTERN'S INTENSITIES, DURATION, RISE TIME AND FALL TIME

Emotion	Duration (Sec)	Color(RGB)	Rise time (%)	Fall time (%)
Normal 0 150, 150, 0 0 0 Anger 1.7 255, 0, 0 83 17 Surprise 4.0 255, 255, 0 83 17 Disgust 2.3 0, 75, 0 83 17 Sadness 4.3 0, 100, 225 20 80 Happiness 2.1 200, 120, 0 80 20 Fear 4.0 0, 20, 80 20 80				

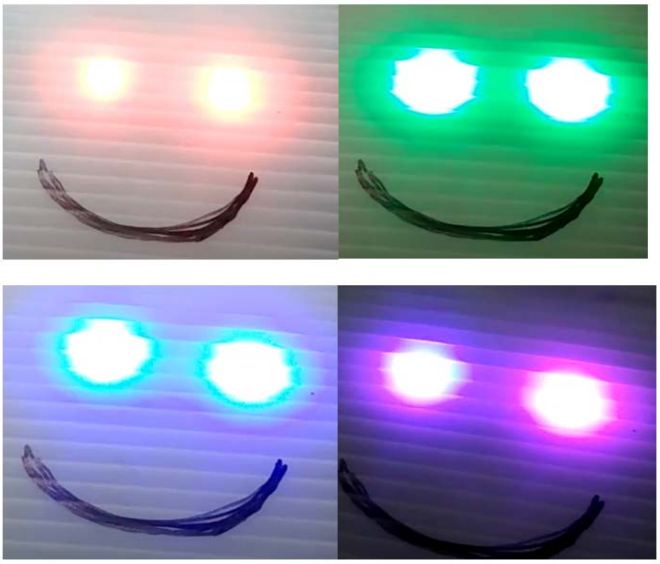
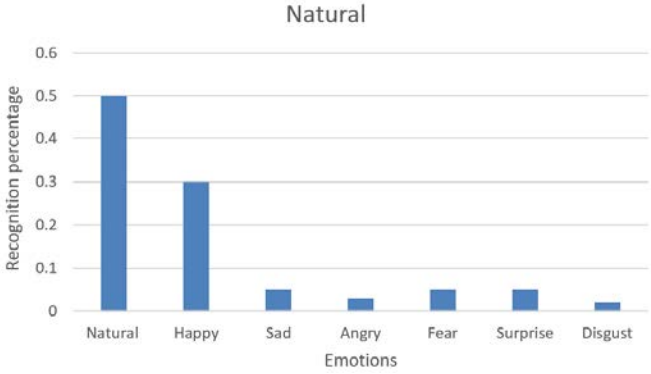


Fig. 3. RGB Patterns shown on the Computer Screen to Recognize Emotion for RGB Pattern Evaluation.

For this, fifty participants were invited, and a survey was conducted. They were asked to check on the recognized emotion after seeing the RGB pattern on the computer screen into robot eyes, and the patterns were shown in random order to avoid ambiguity.

The responses of the participants are interpreted into a bar chart, as shown in Fig. 4. The result was significant as all the emotions were recognized. Moreover, some ambiguity was witnessed between Natural and Happy, Sad and Angry, and Fear and Surprise.

Finally, the data has been collected and analyzed.



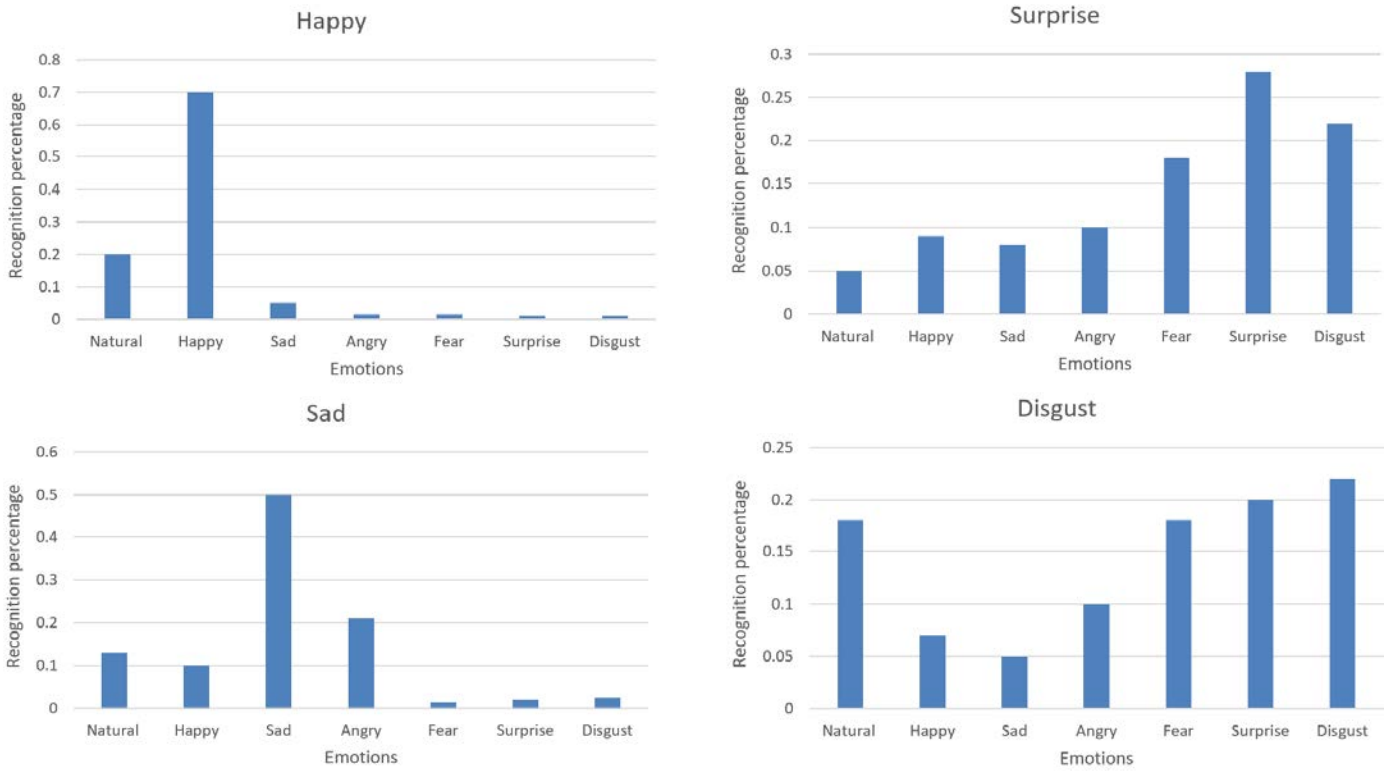


Fig. 4. Average Degree of Recognition for RGB Patterns against each Emotion.

IV. Result and Discussion

The result shows a moderate degree of recognition against each emotion. As a result, it was a bit ambiguous, but when it combined with ABERGS [Prince. M. (2017)], the result was very effective. Fig. 5 illustrates the simulated image.

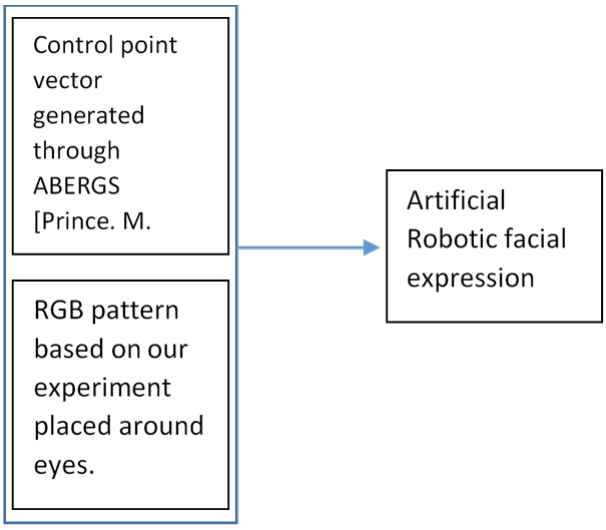


Fig. 5. ABERGS [Prince. M. (2017)] and the RGB Pattern used simultaneously to Mimic Robotic Facial Expression.

TABLE II. COMPARATIVE SATISFACTION LEVEL OF ABERGS [PRINCE. M. (2017)] VS ABERGS [PRINCE. M. (2017)] WITH RGB

Emotions	Satisfaction Level (without voice)	Satisfaction Level (With voice fusion)	Satisfaction Level (With RGB Pattern)
Natural	96.5%	98%	98.7%
Happiness	97%	98%	99%
Surprise	96.5%	97.5%	98%
Fear	97.5%	99%	99.2%
Sadness	96%	98%	98.3%
Disgust	97.5%	98.5%	99%
Anger	95.0%	98%	99.8%

Table II illustrates the effectiveness of the RGB pattern with ABERGS [15].

V. CONCLUSION AND FUTURE WORK

The proposed model uses facial image, voice and simulate artificial emotion on a robotic face with an RGB pattern in the eyes. With the first experiment and RGB patterns were determined, and with the second experiment, the model's effectiveness was tested. The results were very effective. For future work, human gait can also be considered in order to recognize the emotion of the human being. Second, we will focus more on processing speed by using GPU with an efficient algorithm.

ACKNOWLEDGMENT

The author would like to gratefully acknowledge Qassim University, represented by the Deanship of Scientific Research, for the support for this research under the number 1319-coc-2016-1-12-S.

REFERENCES

[1] Angelica Lim and Hiroshi G. Okuno, (2012): Using speech data to recognize emotion in human gait", A. A. Salah et al. (Eds.): HBU, LNCS 7559, pp. 52-64, Springer-Verlag Berlin Heidelberg.

[2] Duhaut D. (2008): A generic architecture of emotion and personality, in Proc. IEEE INT. Conf. Adv. Intell. Machatron., Xian, China, pp. 188-193.

[3] Ekman P. Friesen WV (1969) The repertoire of nonverbal behaviour: categories, origins, usage, and coding. Semiotics 1:49-98.

[4] Greasley P, Sherrard C, Waterman M (2000) Emotion in language and speech: Methodological issues in naturalistic approaches. Lang Speech 43:355–375.

[5] Hiroyaus Miwa, Tomohiko Umetsu, Atsuo Takanishi, et al. (2001): Human2 like robot head that has olfactory sensation and facial color expression" Proceeding of the 2001 IEEE, ICRA, and @001: 459-464.

[6] Jianfeng Zhao, Xia Mao, Lijiang Chen (2019): Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Volume 47, Pages 312-323, ISSN 1746-8094.

[7] Johnson, D.O., Cuijpers, R.H. & van der Pol, D. (2013). Imitating Human Emotions with Artificial Facial Expressions. Int J of Soc Robotics 5, 503–513.

[8] Kanoh M, Iwata S, Kato S., Itoh H (2005) Emotive facial expressions of sensitivity communication robot "Ifbot". Kansei Eng Int 5(3):35-42.

[9] Master Prince, (2013a): A Simple Method for Face Normalization Based on Novel Normal Facial Diagram", International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS Vol:13 No:03.

[10] Master Prince, (2013b): Enhancing face normalization based on novel normal facial diagram, The 26th International Conf. On Computer Applications in Industry and Engineering (CAINE-2013), Los Angeles, California, USA, pp. 97-100, 25-27 September 2013.

[11] McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, and Stroeve S (2000) Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proceedings of ISCA workshop on speech and emotion, Belfast, United Kingdom.

[12] Mehmet Berkehan Akçay, Kaya Oğuz, Speech emotion recognition (2020): Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, Volume 116, Pages 56-76, ISSN 0167-6393.

[13] Meng-Ju Han, Chia-How Lin, and Kai-Tai Song, (2013): Robotics emotional expression generation based on mood transition and personality model, IEEE Transactions on Cybernetics, VOL. 43, NO. 4, August.

[14] Miwa. H., Itoh. K., Matsumoto. M., Zecca M., H. takanobu, S. Rocella, M. C. Carrozza, P. Dario, and A. Takanishi, (2004): Effective emotional expressions with expression humanoid robot WE-4RII: Integration of Humanoid Robor hand RCH-!," in Proc. IEEE/RSJ Int. Conf. Intell. Robots. Syst., Sendai, Japan, pp. 2203-2208.

[15] Miwa. H., Okuchi. T., Itoh. K., Takanobu. H., and Takanishi. A. (2003): A new mantal model for humanoid robots for human friendly communication introduction of learning system, mood vector and second order equations of emotion, in Proc. IEEE Int. Conf. Robots Autom., Taipei, Taiwan, pp3588-3593.

[16] Mohamad Nezami, O., Jamshid Lou, P. & Karami, M. ShEMO, (2019).: a large-scale validated database for Persian speech emotion detection. Lang Resources & Evaluation 53, 1–16.

[17] Moshkina. L., Park. S. , Arkin. R. C., Lee J. K., and JungH. (2011): TAME: Time varying affective response for humanoid robots, Int. J. Social Robot., vol. 3 no. 3. 207-221.

[18] Padgett. C. and Cottrell. G. (1997): Representing face images for classification emotions, vol-9. Cambridge, MA: MIT press.

[19] Power RA, Pluess M. (2015): Heritability estimates of the Big Five personality traits based on common genetic variants. Transl Psychiatry.;5:e604.

[20] Prince. M. (2017): Adaptive artificial brain for humanoid robot using Pattern Recognition and Machine learning, IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, May.

[21] Yang, Y., Ge, S.S., Lee, T.H. et al. (2008), Facial expression recognition and tracking for intelligent human-robot interaction. Intel Serv Robotics 1, 143–157.

[22] Young. A. W and Ellis. H. D. (Editors) (1989): Handbook of Research on Face Processing. North-Holland, Amsterdam: Elsevier Science publishers B. V., ISBN 0 444 87143 8.

AUTHORS' PROFILE

Master Prince, received the B.S degree in computer science from Patna University, India, in 1996, the M.S degree in computer science from Indira Gandhi National Open University, New Delhi, India, in 2004, and the Ph.D. degree in computer science from Pune University, India, in 2008.

Since 2009, he has been working as an Assistant Professor with the Department of Computer Science, Qassim University, Saudi Arabia. His research interests include computer vision and machine learning.

Dr. Prince received the Best Ph.D. Thesis Dissertation of the Year 2009 Award of the Pune University, India.

Deep Learning-enabled Detection of Acute Ischemic Stroke using Brain Computed Tomography Images

Khalid Babutain[1], Muhammad Hussain[2], Hatim Aboalsamh[3], Majed Al-Hameed[4]
Department of Computer Science, College of Computer and Information Sciences, King Saud University, Saudi Arabia[1,2,3]
Department of Neurology, National Institute of Neuroscience, King Fahad Medical City, Saudi Arabia[4]

Abstract*—Stroke is the second leading cause of death globally. Computed Tomography plays a significant role in the initial diagnosis of suspected stroke patients. Currently, stroke is subjectively interpreted on CT scans by domain experts, and significant inter- and intra-observer variation has been documented. Several methods have been proposed to detect ischemic brain stroke automatically on CT scans using machine learning and deep learning, but they are not robust and their performance is not ready for clinical practice. We propose a fully automatic method for acute ischemic stroke detection on brain CT scans. The system's first component is a brain slice classification module that eliminates the CT scan's upper and lower slices, which do not usually include brain tissue. In turn, a brain tissue segmentation module segments brain tissue from CT slices, followed by tissue contrast enhancement using the Extreme-Level Eliminating Histogram Equalization technique. Finally, the processed brain tissue is classified as either normal or ischemic stroke using a classification module, to determine whether the patient is suffering from an ischemic stroke. We leveraged the use of the pre-trained ResNet50 model for slice classification and tissue segmentation, while we propose an efficient lightweight multi-scale CNN model (5S-CNN), which outperformed state-of-the-art models for brain tissue classification. Evaluation included the use of more than 130 patient brain CT scans curated from King Fahad Medical City (KFMC). The proposed method, using 5-fold cross-validation to validate generalization and susceptibility to overfitting, achieved accuracies of 99.21% in brain slice classification, 99.70% in brain tissue segmentation, 87.20% in patient-wise brain tissue classification, and 90.51% in slice-wise brain tissue classification. The system can assist both expert and non-expert radiologists in the early identification of ischemic stroke on brain CT scans.

Keywords*—Acute ischemic brain stroke; deep learning; convolutional neural network; CT brain slice classification; brain tissue segmentation; brain tissue contrast enhancement; brain tissue classification

I. Introduction

Globally, stroke is a leading cause of death, accounting for around 15 million deaths annually [1], [2]. Even in low-income and middle-income countries, stroke is a major cause of mortality, and in the Kingdom of Saudi Arabia (KSA), annual stroke incidence has increased to 29.8 per 100,000 [3]. Notably, 87% of all stroke incidents result from ischemic stroke, whereas the remaining 13% are hemorrhagic [4]. Andersen et al. [5] investigated 39,484 stroke patients, reporting that 35,491 (89.9%) suffered from ischemic stroke and 3,993 (10.1%) experienced hemorrhage stroke. Generally, stroke arises from the sudden interruption of blood flow to neuronal tissue; a blockage within blood vessels leads to ischemic stroke, while blood vessel rapture causes hemorrhagic stroke [2]. To manage ischemic stroke, anti-thrombolytic therapy (removing the blockage by clot breaking) and thrombectomy (removing the clot mechanically) are used, while decompression and blood pressure reduction are used for hemorrhagic stroke [2].

Diagnostic imaging is essential in routine clinical practice to confirm early-stage ischemic stroke. Computed Tomography (CT) is regarded as the front-line modality to evaluate patients with suspected stroke due to its accessibility and cost-effectiveness, which is not the case with Magnetic Resonance Imaging (MRI) [6], [7]. Typically, suspected stroke patients are handled by emergency room physicians, and the condition is often misdiagnosed or diagnosed late due to difficulties in arranging urgent assessments with experienced neuro-radiologists. This often negatively influences stroke management [8].

State-of-the-art methods in computer science are assisting clinicians and neurologists, including for the application of image processing techniques to digital medical images [9], [10], [11]. For example, Chung et al. [12] developed a system to detect hyperacute ischemic stroke in CT images, achieving an accuracy of 81% in classifying stroke and non-stroke images. Methodologically, the authors extracted ranklet features from pre-processed CT images and identified 23 important features for stroke detection, 8 of which were used to establish the prediction model. Guoqing et al. [13] developed a system based on asymmetric image patch classification to detect ischemic stroke signs on non-contrast CT images, achieving an accuracy of 76.84% on 108 stroke cases that trained radiologists did not detect.

Chin et al. [14] developed a CNN model for automatic ischemic stroke diagnosis. Model training and testing involved 256 patch images of size 32×32 extracted manually from CT images. Their system achieved testing and training accuracies of 92% and 97%, respectively. Pereira et al. [15] used two CNNs, one with a 50/50 protocol and another with 75/25, for training and testing with 300 CT images (100 healthy, 100 ischemic, 100 hemorrhagic). Contrasting architectures were used, with the most effective results being 97.5%, 100%, and 99.1% classification accuracies for hemorrhagic, ischemic, and healthy images, respectively, using the 75/25 protocol on their second model. Anis et al. [16] applied deep transfer learning for ischemic stroke detection on CT images, using 400 images with data augmentation (specifically, horizontal flipping) to compare the results to ResNet50, GoogleNet, and VGG-16 pre-

trained CNNs. Using 5-fold cross-validation, they reported that ResNet50, GoogleNet, and VGG-16 achieved 100%, 99.4%, and 92.2% accuracies, respectively, on their training set, while accuracies of 100%, 98.8%, and 90% were reported on the validation set.

Gautam and Raman [17] developed a 13-layer CNN model to classify ischemic and hemorrhagic stroke. Their CT image dataset (300 healthy, 300 ischemic, 300 hemorrhagic) was pre-processed using quadtree-based multi-focus image fusion [18]. Two copies were created for each image, after which contrast adjustment was applied to the first and filtering to the second using a 3×3 averaging filter. The copies were fused and passed to the CNN model, which consisted of an input layer (512×512×1) and two convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation and max-pooling layer, two fully connected layers with a ReLU activation and dropout layer after the first fully connected layer, and a softmax classification layer. Two datasets were established, one containing stroke images only and another containing both stroke and healthy images, and an 80/20 data split protocol with 10-fold cross-validation was applied in each case. They achieved 98.33% and 98.77% classification accuracies on the first dataset, respectively, whereas 92.22% and 93.33% were achieved for the second dataset. The model significantly outperformed fine-tuned AlexNet and ResNet50.

These results highlight the clinical value of computational techniques in stroke detection [19], [20]. However, there is room for improvement, which has caused researchers to leverage Deep Learning (DL) techniques. DL is a subfield of artificial intelligence wherein algorithms learn to make accurate predictions without explicit programming [9], [11], [19]. Convolutional Neural Network (CNN) is a biologically inspired DL paradigm that holds promise in diagnostic medicine due to its ability to outperform humans in image and speech recognition/translation tasks [21].

This research proposes a DL system for acute ischemic stroke classification on brain CT scans. The system determines whether a brain CT slice contains brain tissue, segments the brain tissue, enhances the contrast of the segmented brain tissue, and identifies signs of acute ischemic to determine stroke incidence. CNN-based techniques are adopted in the first two tasks, whereas an efficient multi-scale CNN model – the 5-Scale CNN model (5S-CNN) – is proposed to resolve difficulties associated with distinguishing between normal/abnormal brain regions. A dataset from King Fahad Medical City (KFMC) containing over 130 annotated patient records is used to design, develop, and validate the system, and cross-validation is performed to evaluate the CNN models.

II. Proposed Method

The main technical objective of this research is to develop a robust and intelligent method for the diagnosis of ischemic stroke on brain CT scans, which will assist the clinical decision-making of neurologists. In routine clinical practice, brain CT scans are manually interpreted by professionals, expert operators, or both. This process involves the manual scanning of each slice of the patient's brain CT scan for the presence of stroke. Each patient's CT scan contains 35 to 45 CT slices on average (based on the collected dataset). Manual scanning also includes manual adjustment and enhancement of the contrast of the scanned slice for better visualization.

To formulate such a process, let $C = \{S_i\}_{i=1}^{n}$ represent a patient's CT scan, which consists of n slices S_i of size 512×512×1, as shown in Fig. 1.

Ischemic stroke can appear in any slice within a patient's CT scan; it can also appear at any location in the brain tissue within a slice. Ischemic stroke changes the texture of the affected region of brain tissue, as indicated by comparing the left panel of Fig. 2, which shows a normal brain CT slice, to the right panel, which shows an example of acute ischemic stroke.

Fig. 2 reflects the fact that ischemic stroke can affect any region of the brain tissue. Additionally, the affected area can have any regional size. The affected area becomes darker in texture as the time from its occurrence increases [22].

To determine whether a patient is experiencing an ischemic stroke, it is necessary for a sign of the ischemic stroke to appear on at least one slice of the patient's CT scan. Therefore, each slice must be processed individually, which can be categorized as a classification problem. In any given brain CT slice, there are usually parts that are irrelevant (e.g., scalp, skull, and unrelated background objects) because they do not contribute to the diagnosis. Such parts must be removed, enabling only the brain tissue region to be segmented and separated, which can be categorized as a segmentation problem. Additionally, certain slices in the upper and lower parts of the CT scan do not include brain tissue, and so these slices must be excluded before any processing occurs. This also is categorized as a classification problem, the objective of which is to identify whether a given CT slice contains brain tissue.

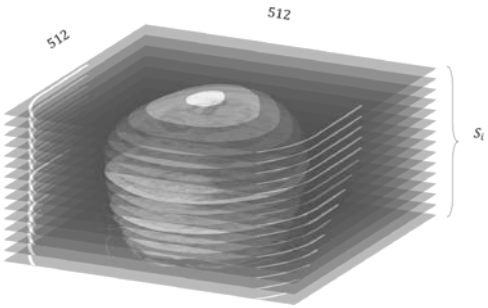


Fig. 1. Brain CT Scan of a Patient.

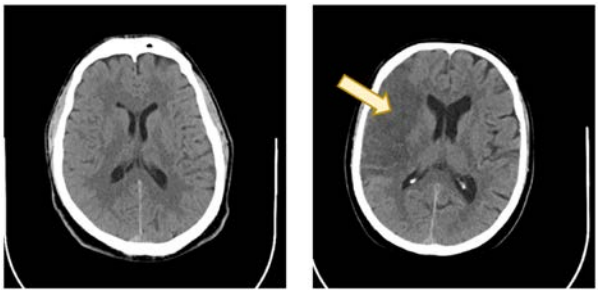


Fig. 2. Normal Brain CT Slice (Left) and Acute Tschemic Stroke (Right).

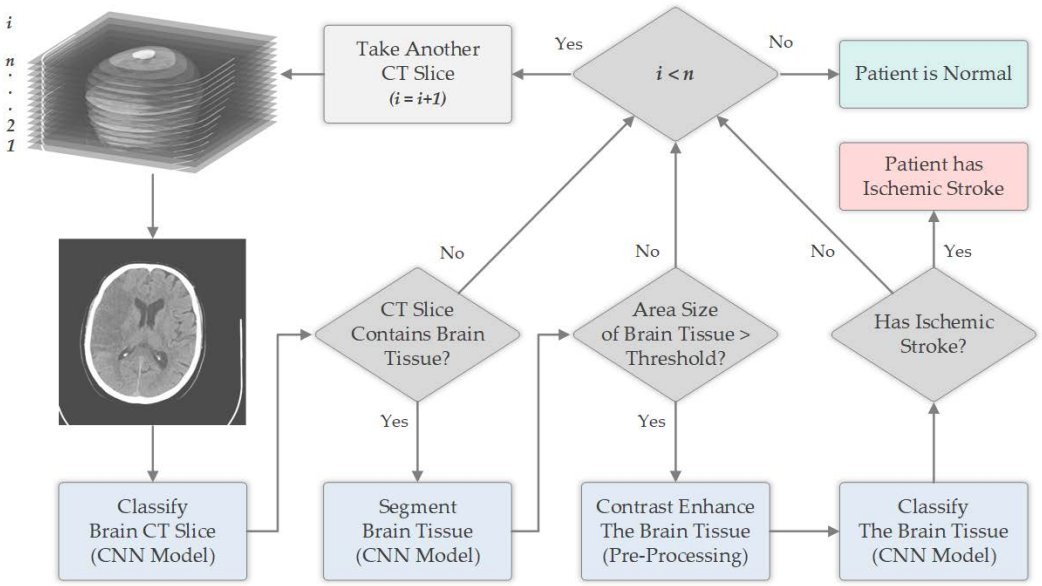


Fig. 3. The Architecture of the Proposed Method.

Given the above discussion, the development of an intelligent method for the automated diagnosis of a patient's brain CT scan involves four main components, which collectively process each slice of a given brain CT scan and systematically determine whether the patient is experiencing an ischemic stroke. As shown in Fig. 3, these components are the Brain CT Slice Classification Model, Brain Tissue Segmentation Model, Brain Tissue Contrast Enhancement Component, and Brain Tissue Classification Model.

The proposed method begins by reading an input CT slice to determine whether the slice should be considered for processing. In any brain CT scan, it is common for the first and last few slices not to contain brain tissue; these slices must be removed prior to any further processing. Also, since some of these CT slices do not contain brain tissue, despite having textural similarities to slices with brain tissue (see Fig. 4), it is necessary to leverage a classification model to determine whether a given slice contains brain tissue.

Before developing this model, a dataset was prepared containing labeled images of CT scan slices with and without brain tissue. Fine-tuning, training, and evaluation of the pre-trained networks were performed. Fine-tuning involves updating each model's input layer to match the size of brain CT slices, which are fixed at 512×512×1. Also, the first convolutional layer's kernels, which are 3-dimensional (spatial × depth) kernels $k(x, y, 3)$, are updated to 2-dimensional kernels $k'(x, y)$ using the mapping $K: k(x, y, 3) \rightarrow k'(x, y)$ such that

$$k'(x, y) = K(k(x, y, 3)) = \frac{1}{3}\sum_{d=1}^{3} k(x, y, d) \quad (1)$$

The final classification layer was also replaced with a new classification layer consisting of two neurons. This is because CT slice classification is a two-class problem (i.e., slice with/without brain tissue). Fig. 5 shows the concept of updating a pre-trained network to be re-trained for brain CT image classification. Training and evaluation included the use of K-fold cross-validation, which increased reliability and generalization and enabled the selection of the model with the best results.

After confirming the existence of brain tissue within the input CT slice, the brain tissue segmentation component detects and segments the brain tissue. In this process, irrelevant parts are removed and only the brain tissue is retained. It begins with an original input slice, as shown in Fig. 6(a). A trained semantic segmentation model is then applied, resulting in an image segmented by class, as in Fig. 6(b). Finally, morphological image analysis is implemented by maintaining the largest segmented group of pixels and filling its holes to obtain the segmented brain tissue [25], as shown in Fig. 6(c).

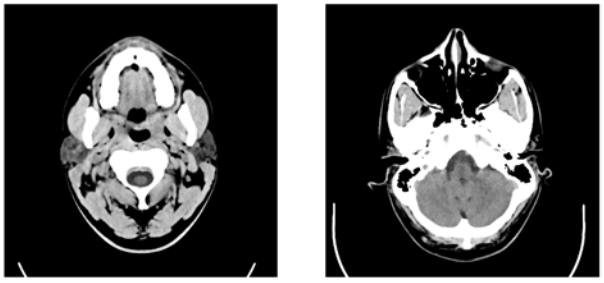


Fig. 4. CT Image with No Brain Tissue (Left) and CT Image with Brain Tissue (Right).

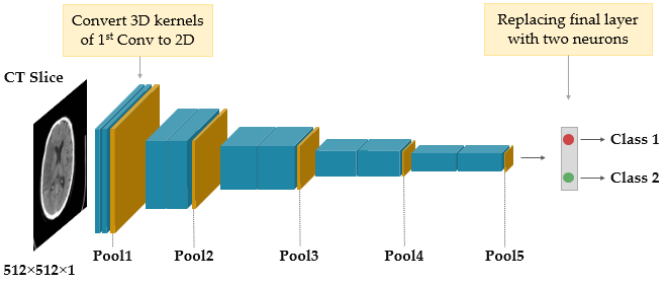


Fig. 5. Adopting a Pre-trained CNN Model for CT Slice Classification.

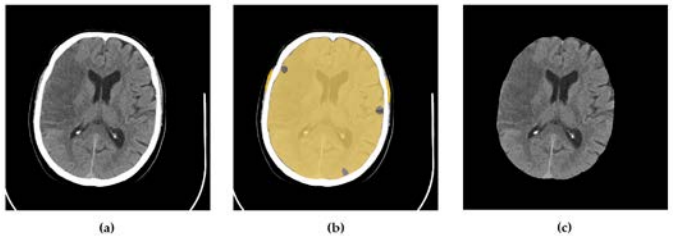


Fig. 6. Brain Tissue Segmentation: (a) Original Image, (b) Applying Semantic Segmentation Model, and (c) Maintaining the Largest Segmented Group of Pixels and Filling its Holes.

DL-based segmentation methods have shown promising results for the segmentation of medical images [26], [27], [28]. The state-of-the-art DL-based framework, namely Fully Convolutional Networks (FCNs) [29], has been used for brain tissue segmentation. FCNs implement a pixel-wise classification process using a CNN model as a classification backbone, which classifies each pixel within an image and assigns it to a particular class, thereby resulting in an image segmented by class. We employed the FCN-8 architecture and tested the four widely-used state-of-the-art pre-trained CNN models as a classification backbone: AlexNet, GoogleNet, ResNet18, and ResNet50. As in brain CT slice classification, ResNet50 yielded the best segmentation results compared to the other models.

Before developing this model, an annotated dataset was prepared by manually annotating the pixels of each CT slice. Each pixel value in every image of this dataset represents a categorical label (either a brain tissue pixel or not). Fine-tuning involved updating the backbone model's input layer to take a CT slice of size 512×512×1 as input and, in turn, updating the first layer's kernels from 3D to 2D, as implemented for slice classification. In addition to this update, all pre-trained models were modified according to the FCN-8 up-sampling structure of the FCNs approach [29]. For training and evaluation, K-fold cross-validation was also used to select the most reliable and generalized segmentation model. Fig. 7 illustrates the concept of updating a pre-trained network according to the FCN-8 architecture to be re-trained for brain tissue segmentation.

Although the optimal segmentation model may lead to high segmentation accuracy, it is common for certain pixels or groups of pixels to be misclassified. This is shown in Fig. 6(b), where some pixels are classified as brain tissue while they are not, and vice versa. Therefore, post-processing is needed, which involves removing small connected components of pixels, and retaining the largest connected group of pixels (which usually represents brain tissue). This is followed by filling in the holes [30], [25] if there is a sufficient ischemic stroke identification area size. Based on recommendations from a medical team, the final segmented brain tissue must be sufficiently large in terms of its area to contribute to the identification of an ischemic stroke. Thus, to make decisions about whether to include or exclude the segmented brain tissue for further processing, the area of the segmented brain tissue (in pixels) is compared to a fixed threshold recommended by the medical team. This comparison excludes any insufficient and small, segmented brain tissue that has an area smaller than the fixed threshold.

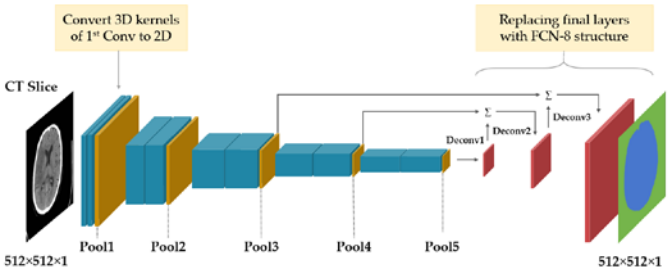


Fig. 7. Adjusting a Pre-Trained Network for Brain CT Image Ssegmentation using FCN-8 Structure.

Due to the low contrast of brain CT slices, multiple image contrast enhancement techniques have been proposed, particularly for medical images [31], [32], [33], [34]. These techniques are used to boost the interior details within brain tissues for visibility, classification, and ultimately interpretation. The Extreme-Level-Eliminating Histogram Equalization (ELEHE) method, proposed by Tan et al. [35], was developed mainly to improve ischemic stroke detection on brain CT images. ELEHE ensures that substantial differences in the distribution of the input CT image histogram are eliminated, resulting in a stretched histogram containing every intensity level value other than the unnecessary two extreme levels that result from regular Histogram Equalization (HE) [30].

Before enhancement, each slice is normalized by stretching the grey-level values within the range of 0 to 216. To enhance a CT slice using ELEHE, the first step calculates the Probability Density Function (PDF) of the slice's grey-level values. The next step involves eliminating the two extreme levels of the resulting PDF, thereby ensuring the maintainability of those levels while stretching the remaining grey levels. Following this, the Cumulative Density Function (CDF) is computed, after which a Transfer Function (TF) is applied to the values of the resulting CDF. Fig. 8 shows a brain CT image enhanced with ELEHE contrast enhancement.

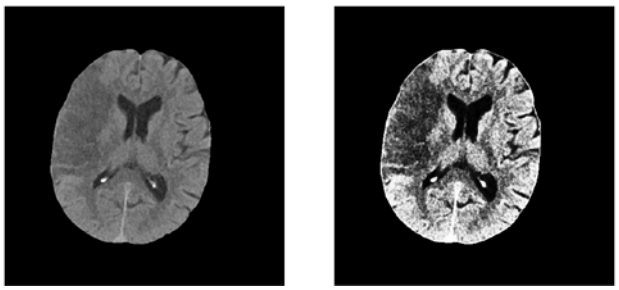


Fig. 8. Brain CT Image (Left) and Brain CT Image after applying ELEHE (Right).

After segmenting the brain region and applying contrast enhancement, the resulting brain tissue is classified to determine whether the CT slice is a case of normal or ischemic stroke, the latter of which indicates that the patient is suffering from ischemic stroke. If the classification of every processed slice of the CT scan is normal, then the patient is regarded as not suffering from ischemic stroke. Since there are two classes to consider (i.e., ischemic stroke or normal), the dataset preparation for this classification included labeled CT slices

indicating an incidence of ischemic stroke or not. All slices with ischemic stroke incidence were labeled as acute, while slices with no stroke incidence were labeled as normal.

Ischemic stroke can manifest with any regional size and it can appear at any location within the brain tissue visualized by CT. Given this, a multi-scale analysis of brain tissue is needed to determine whether there is an instance of ischemic stroke. State-of-the-art pre-trained CNN models such as AlexNet, GoogleNet, ResNet18, and ResNet50 deal with only one scale, therefore the performance of these models is unsatisfactory. To overcome this issue, a lightweight multi-scale CNN model named the 5-Scale CNN model (5S-CNN) is proposed. After segmentation and contrast enhancement of the input CT brain tissue image, 5S-CNN uses a 5-branch architecture wherein each branch applies a different filter size to learn features at different scales.

5S-CNN consists of 77 layers, as shown in Fig. 9. Within this model, a total of 16 convolutional (Conv) blocks are used, where each block starts with a Conv layer followed by Batch Normalization (BN) and Rectified Linear Unit (ReLU) layers. The ReLU layers introduce non-linearity into the model in a very simple way of applying a thresholding operation to the pixels resulting from the BN layers, in which positive pixels are retained, and negative ones are assigned to zero. A max-pooling layer is positioned after each Conv block, resulting in a total of 16 pooling layers in the proposed model. The use of max-pooling layers enables the selection of only one pixel whose value is the highest compared to the other pixels within the pooling receptive field. This leads to the extraction of relevant features as well as a reduction in image size.

Within the first Conv layer of the proposed model, 16 filters with a size of 4×4 and stride of 2×2 are used, which down-samples the CT input image from 512×512×1 to 255×255×16. Thereafter, max-pooling with a size of 3×3 and stride of 2×2 is applied, which reduces the first Conv output size to 127×127×16. It is then passed to 5 branches having three consecutive Conv blocks in each branch. The first Conv blocks in these branches contain 32 filters in each branch with sizes of 11×11, 9×9, 7×7, 5×5, and 3×3 in order to extract and learn features at different scales. After the first Conv block of each branch, two Conv blocks are used with 64 and 128 filters, respectively, of size 3×3 and stride 1×1; in these Conv blocks, a 3×3 filter size is used to reduce the number of learnable parameters as well as the model's computational cost. Additionally, each Conv block within each branch is followed by a max-pooling layer with a receptive field size of either 3×3 or 2×2 depends on whether the previous output feature map is odd or even in terms of its width and height.

At the end of each branch, a Global Average Pooling (GAP) layer is used to generate a channel descriptor for that branch. The resulting 5 GAP descriptors are combined using a concatenation layer and, in turn, for fusion, the output is used as an input to a Fully Connected (FC) block that has an FC layer with 256 filters, followed by a BN layer and ReLU layer, respectively. In the end, a single FC layer with two neurons and a softmax are used as a classifier. The number of neurons in this FC layer is targeted at the number of classes of interest (i.e., ischemic stroke or normal).

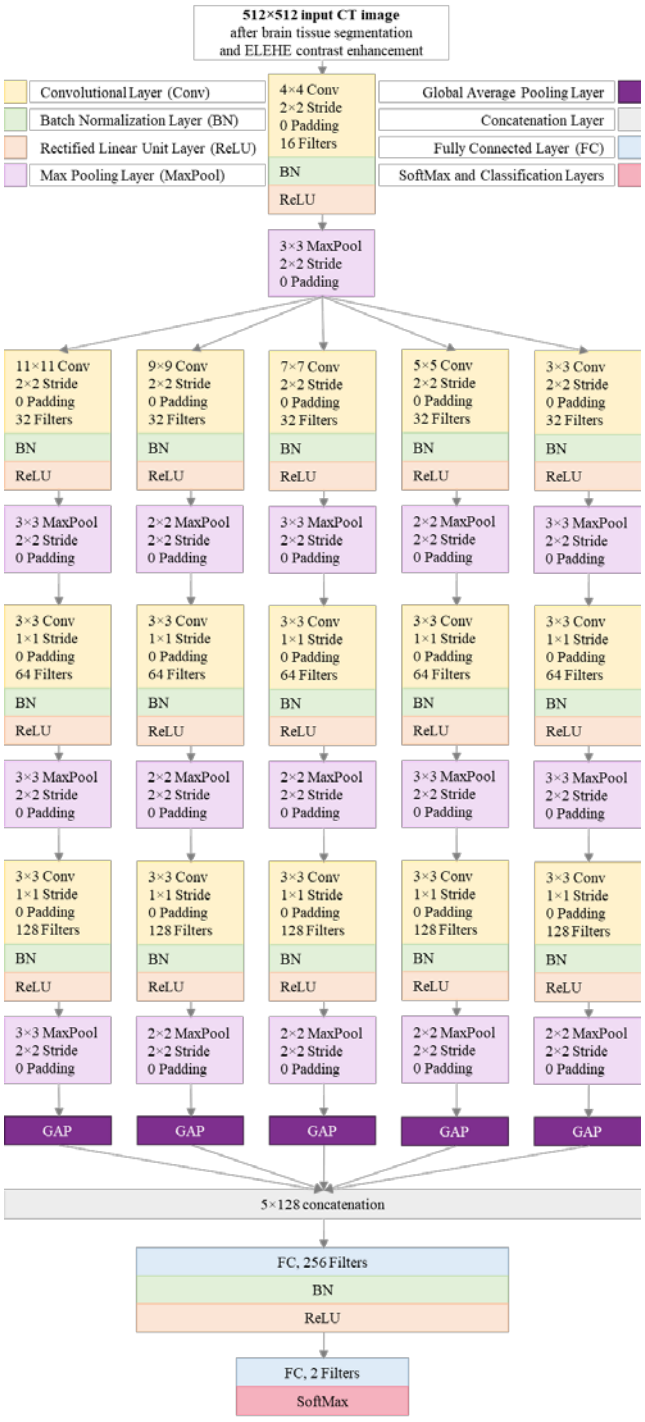


Fig. 9. The Proposed 5-Scale CNN Model (5S-CNN) Architecture for Ischemic Stroke Classification.

For comparison, the pre-trained models for tissue classification were fine-tuned using the approach adopted in the first and second components (i.e., CT slice classification and brain tissue segmentation). Specifically, the input layer of each pre-trained model was updated to match the DICOM brain CT image size of 512×512×1. Additionally, the filters' weights for the first convolutional layer were updated from 3-dimensional kernels to 2-dimensional kernels, which was achieved by taking the mean of each filter value across the

depth dimension, as in (1). The final classification layer was replaced with new classification layers targeted at brain tissue classification.

The same training and evaluation procedures were performed on each model, including the 5S-CNN, and each segmented CT slice was contrast-enhanced using ELEHE before being passed as an input to any model. As in the first and second components, K-fold cross-validation was used to validate the reliability and generalization of each model.

III. Dataset Collection, Annotation, and Preparation

The dataset used to develop and validate the proposed method was collected from King Fahad Medical City (KFMC) under Institutional Review Board (IRB) approval with log number (17-031). The dataset contains brain CT scans from more than 130 patients, consisting of proven cases of both normal and acute ischemic stroke scans. The collection process involved compiling a shortlist from clinical and radiologic databases of patients who presented at emergency rooms or clinics between January 2015 and January 2018 with symptoms and signs of stroke. Revision for inclusion and exclusion of the selected patients was performed, in which scans with artifacts (e.g., motion and metal) and findings of hemorrhagic stroke were excluded. The remaining scans were included for analysis and modeling. During the medical team's annotation process, records of included patients were annotated and categorized into either normal or acute based on review findings. Each scanned record was further annotated by specifying the slices of the patient's CT scan that were affected by the ischemic stroke.

Dataset preparation for the Brain CT Slice Classification Model included a subset of 1,130 CT images selected randomly from 100 patients. The dataset contained 570 images labeled as Brain CT Slice from 50 patients, and 560 images labeled as Not Brain CT Slice from 50 patients. For the Brain Tissue Segmentation Model, another subset of the collected dataset was prepared that included 365 CT images from 18 randomly selected patients. This subset was manually labeled using MATLAB Image Labeler under the supervision of the medical team. Pixel-wise labeling was used to label each selected image, wherein each pixel in every image was labeled as either Brain Tissue or Not Brain Tissue. In the Brain Tissue Classification Model, the collected dataset included 63 patients with an acute diagnosis, having 300 CT images labeled as Acute. For data balancing and cross-validation, 62 patients with a normal diagnosis were selected, having 300 CT images labeled as Normal.

IV. Data Augmentation

A large number of training images is required to train deep learning models, especially for image classification tasks. When only a small number of training images is available, model overfitting to the training images arises, which weakens the model's ability to adapt to new data. Different image augmentation techniques exist that can be used to improve the performance and generalization of deep learning models. These techniques rely on the creation of different forms of the original images used for training [36].

In our case, five augmentation techniques were used in the training of all models, including random reflection in both horizontal and vertical axes. The horizontal reflection applies the random reflection in the left-right direction of the image, while the vertical reflection applies the random reflection in the top-bottom direction. The other two techniques used were image translation on both directions of the input image, including the x-axis direction (horizontally) and the y-axis direction (vertically). It is necessary to specify a pixel range for this translation technique, which was set at 10 pixels for both translation directions. The fifth augmentation technique was random rotation, which was set in the range of 10º image rotation clockwise and anticlockwise.

V. Experiments and Results

All experiments were performed on a machine running the 64-bit Windows 10 operating system. The machine had an Intel® Core™ i7-8750H CPU @ 2.20GHz and 32GB of RAM, and it was equipped with an NVIDIA GeForce 1070 with 8GB of GPU memory. Model training and testing were implemented and evaluated using the 64-bit version of MATLAB R2020b.

To increase the generalization and reliability of the results, as well as due to data limitations, 5-fold cross-validation was used to validate the trained models. Depending on the type of model and its dataset, as in Table I, each model was trained and tested five times; every time, one-fold was used for testing, while the other four folds were used for training and validation. The data splitting approach was 70%, 10%, and 20% for training, validation, and testing, respectively.

TABLE I. Dataset Preparation for each Model

Model	Labeling Type	Labeling As	No. Images	No. Patients
Brain CT Slice Classification	Image as a class	Brain CT Slice	570	50
		Not Brain CT Slice	560	50
Brain Tissue Segmentation	Pixel as a class	Brain Tissue	365	18
		Not Brain Tissue		
Brain Tissue Classification	Image as a class	Acute	300	63
		Normal	300	62

In the Brain CT Slice Classification Model, 5-fold cross-validation was used to evaluate each of the fine-tuned pre-trained models (i.e., AlexNet, GoogleNet, ResNet18, and ResNet50). In each run of this approach, 226 images were used as a testing fold, while from the remaining images, 814 and 90 images were used for training and validation, respectively. For the Brain Tissue Segmentation Model, the adjustments to the pre-trained models, which included updating their structures to match the FCN-8 structure, were also evaluated using 5-fold cross-validation. In this case, each fold included 263, 29, and 73 images for training, validation, and testing, respectively.

Evaluation of the Brain Tissue Classification Model included evaluating the classification performance of the fine-tuned pre-trained models as well as the proposed 5S-CNN model in slice-level and patient-level classifications. Slice-level evaluation, considering that the same patient's slices are in only the training, validation, or testing dataset, demonstrates the slice-wise classification performance. However, patient-

level classification analyzes the performance for every slice of a patient and reports a single resulting class for that patient. Although patient-wise evaluation is the standard approach in the medical field [37], slice-level evaluation was also performed because a patient should be diagnosed with ischemic stroke if at least one slice of the patient's CT scan shows a sign of the ischemic stroke. The slice-level dataset splitting approach does not strictly comply with the 70%, 10%, and 20% method, especially when considering the use of cross-validation. The reason is that every patient will have a different number of CT slices that are affected by ischemic stroke. The patient-level acute vs. normal classification included 125 patients, and it was validated using 5-fold cross-validation. In each run of this approach, 25 patients were used as a testing fold, while from the remaining 100 patients, 90 were used for training and 10 for validation. ELEHE contrast enhancement was applied to each input image prior to Brain Tissue Classification.

Among the available optimization algorithms, the Adam optimization algorithm has been shown to outperform its counterparts. Therefore, all trained models were trained using the Adam optimization algorithm with a gradient decay factor of 0.9. The initial learning rate was set to 0.001 while the regularization factor was set to 0.0001. Each model was trained for 100 epochs with a minibatch size of 16 due to memory limitations.

For evaluation, the commonly employed performance measurements of accuracy, sensitivity, specificity, precision, and F1 score are used. However, the performance of segmentation models was evaluated in terms of pixel global accuracy, mean recall, mean Intersection over Union (IoU), Weighted Intersection Over Union (wIoU), and mean Boundary F1 (BF) score. Using 5-fold cross-validation, the mean and standard deviation of the resulting measurements over the five testing folds were reported. Performance measurements of the classification models were derived based on the concepts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as in (2–7).

For the evaluation of segmentation models, the pixel global accuracy measurement indicates the percentage of correctly identified pixels corresponding to the total number of pixels, regardless of the class type, as in (2). Mean recall measures the ratio of accurately classified pixels to the total number of pixels based on class type, which is averaged across both classes (see (3)). IoU computes the ratio of accurately classified pixels to the total number of ground truth and predicted pixels based on class type, as shown in (7). By averaging the resulting IoU over classes, the mean IoU can be obtained. In wIoU, the average IoU of each class is weighted by the total number of pixels in its corresponding class. The BF score calculates the predicted boundary of each class relative to its true boundary, as shown in (6). The mean BF score can then be obtained by averaging the resulting BF scores over classes.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}(\times 100\%) \quad (2)$$

$$Sensitivity = \frac{FP+FN}{TP+TN+FP+FN}(\times 100\%) \quad (3)$$

$$Specificity = \frac{TP}{TP+FN}(\times 100\%) \quad (4)$$

$$Precision = \frac{TN}{FP+TN}(\times 100\%) \quad (5)$$

$$F1\ Score = \frac{TP}{TP+FP}(\times 100\%) \quad (6)$$

$$\text{IoU} = \frac{2TP}{2TP+FP+FN}(\times 100\%) \quad (7)$$

A. Experimental Results for Brain CT Slice Classification

The experimental results for the brain CT slice classification task, which is the first step of the proposed methodology (as shown in Fig. 3), are given in Table II. Each in Table II presents the evaluation results for each of the fine-tuned pre-trained CNN models (i.e., AlexNet, GoogleNet, ResNet18, and ResNet50). For each evaluated model, the mean and standard deviation of the resulting measurements are computed using 5-fold cross-validation.

Based on the obtained results (see Table II), the selected model for the brain CT slice classification task was the fine-tuned ResNet50. This decision was made because ResNet50outperformed other models on each metric except sensitivity which was equal to ResNet18 but better in terms of standard deviation.

B. Experimental Results for Brain Tissue Segmentation

The brain tissue segmentation results, using FCN-8 with different fine-tuned backbone models, are shown in Table III. In this experiment, 365 images were used, in which every pixel within each image was labeled as either a brain tissue pixel or not a brain tissue pixel. The results of Table III show that both ResNet18 and ResNet50 produced similar results, but both outperformed other models.

FCN-8 with fine-tuned ResNet50 performed excellently in segmenting brain tissue pixels, achieving 99.70% global accuracy with a standard deviation of 0.04% (see Table III). However, certain pixels were misclassified. Therefore, as explained in section II, post-processing was applied after brain tissue segmentation; the area of the largest segmented connected group of pixels in the CT slice (which usually represents brain tissue) was compared to a threshold fixed at 1,000 pixels to either include or exclude the brain CT slice for further processing. The threshold was fixed based on the recommendations of the medical team. This comparison ensures that the area of the segmented brain tissue is sufficiently large for the subsequent brain tissue classification task. After this, it is necessary to fill the holes within the segmented brain tissue to ensure that misclassified brain tissue pixels are included in the largest segmented connected group of pixels. Fig. 10 shows experimental examples of segmented brain tissues using FCN-8 with fine-tuned ResNet50 followed by post-processing.

C. Experimental Results for Brain Tissue Classification

After brain tissue segmentation, the final task is to classify the segmented brain tissue as ischemic stroke or normal. Patient-wise and slice-wise classification experiments were performed for validation in this experiment. Each input image was enhanced using the ELEHE contrast enhancement technique prior to the training and testing of the fine-tuned pre-trained models and the 5S-CNN. As shown in Table IV, the 5S-CNN model with ELEHE contrast enhancement

outperformed the other models for every scenario, reflecting its power in terms of multi-scale feature learning. In addition, the decision made regarding the number of branches used in the proposed multi-scale CNN model (5S-CNN) was based on experimenting with the multi-scales of 2, 3, 4, and 5 branches. Five branches were found to yield the best results in all scenarios (patient-wise and slice-wise classification), as shown in Table V.

TABLE II. BRAIN CT SLICE CLASSIFICATION RESULTS (MEAN ± STANDARD DEVIATION) AVERAGE OVER THE FIVE FOLDS OF EACH MODEL

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
AlexNet	98.31 ± 0.63	98.33 ± 0.47	97.01 ± 1.45	97.99 ± 0.67	98.79 ± 0.53
GoogleNet	98.59 ± 0.41	98.80 ± 0.48	98.17 ± 0.74	98.21 ± 0.45	98.95 ± 0.65
ResNet18	99.04 ± 0.38	99.17 ± 0.88	98.92 ± 1.24	98.94 ± 1.20	99.04 ± 0.37
ResNet50	**99.21 ± 0.31**	**99.17 ± 0.42**	**99.25 ± 0.68**	**99.25 ± 0.68**	**99.21 ± 0.31**

TABLE III. BRAIN TISSUE SEGMENTATION RESULTS (MEAN ± STANDARD DEVIATION) AVERAGE OVER THE FIVE FOLDS OF EACH MODEL

Model	Global Accuracy	Mean Sensitivity	Mean IoU	Weighted IoU	Mean F1 Score
AlexNet	97.97 ± 2.60	98.43 ± 1.91	94.41 ± 6.81	96.32 ± 4.43	87.44 ± 13.69
GoogleNet	97.58 ± 3.28	98.23 ± 1.87	93.58 ± 8.04	95.72 ± 5.41	84.44 ± 19.43
ResNet18	99.63 ± 0.08	99.69 ± 0.04	98.85 ± 0.28	99.26 ± 0.15	98.87 ± 0.50
ResNet50	**99.70 ± 0.04**	**99.75 ± 0.04**	**99.06 ± 0.15**	**99.40 ± 0.08**	**98.93 ± 0.18**

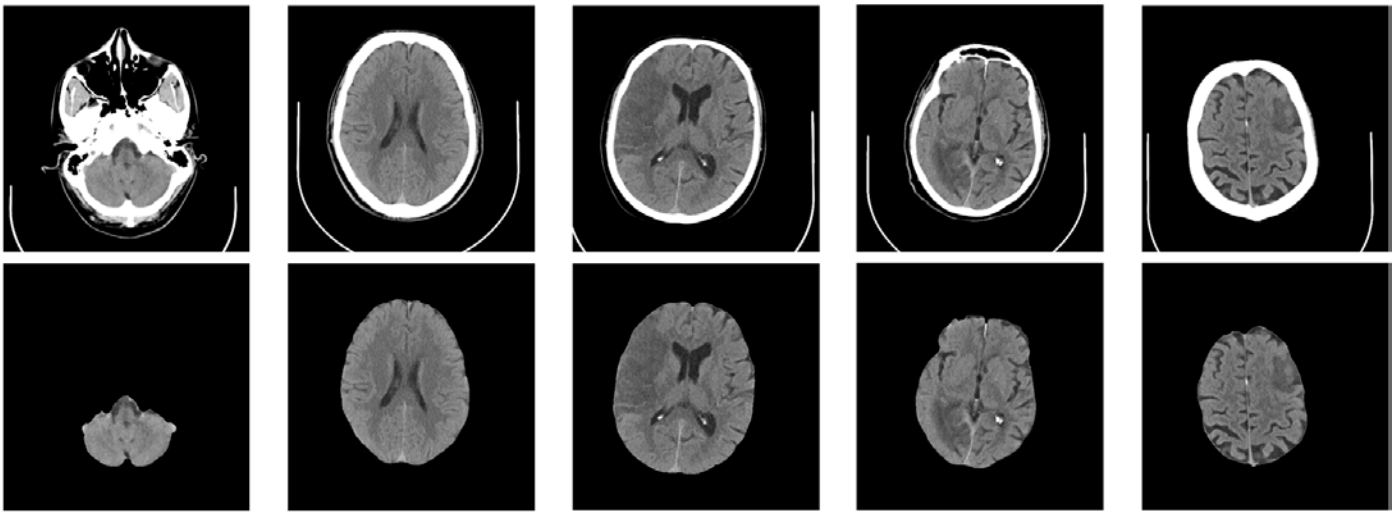


Fig. 10. Examples of Segmented Brain Tissues using FCN-8 with Fine-Tuned ResNet50 Followed by Post-processing: Original Image (Top), Segmented Image (Bottom).

TABLE IV. PATIENT-WISE AND SLICE-WISE BRAIN TISSUE CLASSIFICATION RESULTS (MEAN ± STANDARD DEVIATION) AVERAGE OVER THE FIVE FOLDS OF EACH MODEL

Scenario	Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Patient-wise classification	AlexNet	74.40 ± 8.29	84.57 ± 4.95	70.27 ± 9.25	56.67 ± 17.08	66.95 ± 12.94
	GoogleNet	75.20 ± 7.16	81.95 ± 11.07	74.36 ± 10.11	65.00 ± 22.36	69.80 ± 14.03
	ResNet18	79.20 ± 6.57	82.98 ± 6.39	77.59 ± 9.10	71.67 ± 13.94	76.30 ± 8.68
	ResNet50	81.60 ± 6.69	86.79 ± 7.58	79.35 ± 8.91	73.33 ± 14.91	78.68 ± 9.72
	5S-CNN	**87.20 ± 5.93**	**88.95 ± 3.67**	**87.44 ± 8.89**	**85.00 ± 12.36**	**86.19 ± 6.65**
Slice-wise classification	AlexNet	76.61 ± 6.34	78.11 ± 7.66	76.43 ± 12.17	70.73 ± 16.52	73.39 ± 9.04
	GoogleNet	79.30 ± 6.84	80.37 ± 8.35	79.09 ± 10.00	77.82 ± 9.87	78.57 ± 6.42
	ResNet18	82.61 ± 6.27	80.44 ± 5.79	85.50 ± 12.93	84.35 ± 15.78	81.40 ± 6.68
	ResNet50	84.14 ± 4.09	83.18 ± 6.01	85.34 ± **5.39**	85.08 ± **5.04**	83.96 ± 3.81
	5S-CNN	**90.51 ± 2.22**	**90.72 ± 5.38**	**91.38** ± 5.71	**90.15** ± 7.77	**90.11 ± 2.78**

TABLE V. IMPACT OF SCALES ON THE PROPOSED MULTI-SCALE CNN FOR PATIENT-WISE AND SLICE-WISE CLASSIFICATIONS (MEAN ± STANDARD DEVIATION) AVERAGE OVER THE FIVE FOLDS OF EACH MODEL

Scenario	Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Patient-wise classification	2S-CNN	73.60 ± 6.07	80.23 ± 7.71	71.90 ± **8.45**	61.67 ± 19.18	67.81 ± 12.31
	3S-CNN	71.20 ± 11.10	80.42 ± **5.63**	68.82 ± 13.75	51.67 ± 25.95	60.51 ± 19.73
	4S-CNN	78.40 ± 6.69	78.36 ± 8.15	79.59 ± 8.68	76.67 ± 13.36	77.05 ± 8.08
	5S-CNN	**87.20 ± 5.93**	**88.95** ± 8.67	**87.44** ± 8.89	**85.00 ± 12.36**	**86.19 ± 6.65**
Slice-wise classification	2S-CNN	83.42 ± 4.73	86.46 ± **3.57**	81.45 ± 8.15	78.82 ± 9.60	82.11 ± 5.07
	3S-CNN	82.44 ± 5.65	84.47 ± 9.43	82.08 ± 10.49	79.91 ± 13.25	81.19 ± 6.83
	4S-CNN	86.34 ± 3.29	88.08 ± 4.24	85.27 ± 7.32	83.48 ± 9.59	85.35 ± 4.55
	5S-CNN	**90.51 ± 2.22**	**90.72** ± 5.38	**91.38 ± 5.71**	**90.15 ± 7.77**	**90.11 ± 2.78**

VI. DISCUSSION

The first two tasks (i.e., CT slice classification and brain region segmentation) are relatively easy problems; ResNet50 and FCN-8 based on ResNet50 work adequately in both cases. However, the third task (i.e., classification of brain tissue as normal or ischemic stroke) is comparatively difficult due to the textural similarity between the normal region and the region affected by ischemic stroke. For this purpose, the 5S-CNN model is proposed. In this section, we discuss the classification results for fine-tuned pre-trained networks as well as the proposed 5S-CNN regarding the task of classifying ischemic stroke against normal cases in patient-wise and slice-wise classifications, the results for which are given in Table IV.

In brain tissue classification, AlexNet and GoogleNet achieved the lowest performance in both patient-wise and slice-wise classifications. Mean accuracies of 74.40% and 75.20% and standard deviations of 8.29% and 7.16% resulted from both fine-tuned models in patient-wise classification, while slice-wise classification achieved mean accuracies of 76.61% and 79.30% and standard deviations of 6.34% and 6.84%. For patient-wise classification using AlexNet, the values for average sensitivity, specificity, precision, and F1 score were 84.57%, 70.27%, 56.67%, and 66.95%, respectively, while GoogleNet achieved 81.95%, 74.36%, 65%, and 69.80% for these metrics. However, slice-wise classification resulted in an average sensitivity of 78.11%, specificity of 76.43%, precision of 70.73%, and F1-score of 73.39%, while GoogleNet achieved 80.37%, 79.09%, 77.82%, and 78.57% for these metrics, respectively. The standard deviation of sensitivity for AlexNet was 4.95% in the patient-wise scenario and 7.66% in the slice-wise scenario, while for GoogleNet, the values were 11.07% and 8.35%, respectively. Notably, the accuracy difference of 2.69% in the patient-wise scenario and 0.8% in the slice-wise scenario indicates the favorable generalization performance of GoogleNet. Compared to GoogleNet, ResNet18 and ResNet50 were associated with better performance, achieving mean accuracies of 79.20% and 81.60% in patient-wise classification and 82.61% and 84.14% in slice-wise classification, respectively.

For ResNet18, the average values for sensitivity, specificity, precision, and F1 score were 82.98%, 77.59%, 71.67%, and 76.30%, respectively in the patient-wise scenario, while the values for the same metrics were 80.44%, 85.50%, 84.35%, and 81.40% in the slice-wise scenario. ResNet50 yielded better performance values with an average sensitivity of 86.79%, specificity of 79.35%, precision of 73.33%, and F1 score of 78.68% in the patient-wise scenario, while the slice-wise scenario resulted in 83.18%, 85.34%, 85.08%, and 83.69%, respectively. Only in the slice-wise scenario, ResNet50 outperformed ResNet18 in all performance metrics except specificity, where the results indicated a specificity difference of 0.16% for ResNet18.

The proposed model, 5S-CNN, outperformed the other models on every metric in both patient-level and slice-level classification. Values of 87.20%, 88.95%, 87.44%, 85.00%, and 86.19% were achieved in the patient-wise scenario with respect to sensitivity, specificity, precision, and F1 score, while values of 90.51%, 90.72%, 91.38%, 91.15%, and 90.11% were achieved for slice-wise classification. Also, the standard deviation of the proposed model outperformed the other models on all metrics except specificity and precision, only in the slice-wise scenario. In this case, the standard deviation of 5S-CNN was 5.71% and 7.77% in these two metrics, whereas the specificity and precision of ResNet50 were 5.39% and 5.04%, respectively. Acute stroke is an early sign of ischemic stroke, and it is very difficult to identify due to the subtle differences that exist between normal brain regions and those affected by acute ischemic stroke. However, in this case, 5S-CNN yielded an excellent performance overall.

Table VI shows a comparison of model complexity for the ischemic stroke classification models based on the number of parameters and Floating-Point Operations (FLOPs). The number of parameters is the sum of all learnable weights and biases of all Conv and FC layers within a CNN model. By contrast, the number of FLOPs reflects the computations required for a single forward pass within the model. After adopting the pre-trained networks for brain CT images, AlexNet had the greatest number of parameters (approximately 57.7 million) and around 7.6 billion FLOPs. GoogleNet consisted of approximately 5.1 million parameters and 7.7 billion FLOPs. Notably, despite the substantial difference in the number of parameters, AlexNet and GoogleNet had almost the same number of FLOPs. This is attributable to the small kernel sizes and feature maps of GoogleNet, which contributed to the similar computational cost of AlexNet and GoogleNet. ResNet18 had approximately 11.2 million parameters and 9.1 billion FLOPs, meaning that its computational cost exceeds AlexNet and GoogleNet. ResNet50 had the largest number of FLOPs compared to the other models, amounting to

approximately 19.8 billion FLOPs, along with 23.2 million parameters. The significant number of FLOPs means that ResNet50 is substantially more computationally expensive and consumes more training time compared to its counterparts. The proposed model, 5S-CNN, has the lowest number of parameters (approximately 0.8 million) and FLOPs (approximately 0.6 billion). Therefore, in addition to outperforming the other models in terms of classification results, 5S-CNN is also more computationally efficient.

TABLE VI. COMPLEXITY OF MODELS USED FOR STROKE CLASSIFICATION

Model	Parameters (Millions)	FLOPs (Billions)
AlexNet	57.5	7.6
GoogleNet	5.1	7.7
ResNet18	11.2	9.1
ResNet50	23.5	19.8
5S-CNN	**0.8**	**0.6**

It is worth comparing the proposed method to those of Pereira et al. [15], Anis et al. [16], and Gautam et al. [17]. Each of these researchers also leveraged DL methods for ischemic stroke detection using brain CT images. The performance of these methods, as reported in the introduction, was determined based on each research group's private datasets. Therefore, to facilitate a fair comparison, we implemented these methods and trained them on the collected dataset based on the authors' recommendations. Both patient-wise and slice-wise brain tissue classification experiments were performed. The results and confusion matrices are shown in Table VII, as well as Fig. 11.

For the classification of acute ischemic stroke against normal cases, the results in Table VII show that 5S-CNN outperformed the other three methods on all performance metrics in both the patient-wise and slice-wise scenarios. Furthermore, the confusion matrices in Fig. 11(a) show the decisions that each method made regarding the optimal testing fold across the five cross-validated folds. In this fold, a total of 25 patients with 136 CT images were used for testing. Among those, 13 patients with 73 images had acute ischemic stroke, while 63 images were normal from the remaining normal patients.

In the patient-wise scenario, 5S-CNN correctly classified all acute patients as having acute ischemic stroke, whereas the other methods misclassified at least one acute patient as a normal patient. In this case, the classifier determines that a patient has acute ischemic stroke if one of the patient's CT slices is classified as having acute stroke. Due to this, the methods of Pereira et al. [15] and Anis et al. [16] were able to correctly classify 4 normal patients, while 8 were misclassified as having acute ischemic stroke. Also, the method proposed by Gautam et al. [17] misclassified 9 normal patients as having acute ischemic stroke. In the case of 5S-CNN, only 2 normal patients were misclassified as having acute ischemic stroke. As such, 5S-CNN outperformed all other models in this area.

A similar trend is seen in the scenario of slice-wise classification, as shown in Fig. 11(b). 5S-CNN correctly classified 69 acute stroke slices out of 73 images, while 21, 11, and 38 images were misclassified in Pereira et al. [15], Anis et al. [16], and Gautam et al. [17], respectively. In the classification of normal images, the compared methods misclassified more than 14 normal images as acute ischemic stroke, while 5S-CNN misclassified only 6 images out of the 63 normal tested images. Although our model misclassified 4 images as normal from the acute patients in slice-wise classification, it correctly classified all patients as suffering from acute ischemic stroke in the patient-wise scenario. The encoding of multi-scale information from the CT scans proves the potential of the 5S-CNN model to improve clinical decision-making.

Compared to state-of-the-art models, as well as similar works, the proposed method uses a fully automated approach to analyze brain CT images and determine whether it includes brain tissue or not. In turn, through the segmentation of brain tissue within the CT image, the method enables irrelevant objects and background to be eliminated, as well as ensuring the accurate segmentation of all brain tissue pixels using morphological operations. Following this, the method applies the ELEHE contrast enhancement technique to boost the interior details of the segmented brain tissue, thereby facilitating more effective classification. Finally, classification of the resulting brain tissue using the proposed model outperformed state-of-the-art CNN models, as well as models proposed in similar previous methods. All CT slices in the collected dataset were used without altering their original 16-bit greyscale range, thereby ensuring the stability of the original pixel values. Additionally, rather than resizing original images to fit the pre-trained networks, the pre-trained networks were fine-tuned to fit the original size of CT brain images. Taken together, the use of 5-fold cross-validation shows the generalization of the proposed 5S-CNN model, as well as its low susceptibility to overfitting.

Nevertheless, the 5S-CNN model is limited to deciding whether a brain tissue shows any signs of ischemic stroke without any localization of the stroke lesion within the brain CT slice. Another limitation of the proposed model is the possibility of misdiagnosing ischemic stroke in slices with three brain tissue areas separated by bones, which usually occur at the lower part of a brain CT scan. The developed segmentation module preserves the largest brain tissue area for further analysis and removes the rest. This decision was taken based on the medical team's recommendation for the current study, and it is a potential area for investigation in future works.

TABLE II. COMPARISON TO RELATED METHODS APPLIED ON THE COLLECTED DATASET IN PATIENT-WISE AND SLICE-WISE CLASSIFICATIONS (MEAN ± STANDARD DEVIATION) AVERAGE OVER THE FIVE FOLDS OF EACH MODEL

Scenario	Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Patient-wise classification	Pereira et al. [15]	63.20 ± 7.16	79.89 ± 12.33	60.43 ± 6.36	33.33 ± 21.25	43.16 ± 21.76
	Anis et al. [16]	81.60 ± 6.69	86.79 ± 7.58	79.35 ± 8.91	73.33 ± 14.91	78.68 ± 9.72
	Gautam et al. [17]	53.60 ± 3.58	63.33 ± 22.31	53.38 ± 2.16	21.67 ± 9.50	29.81 ± 9.81
	Proposed (5S-CNN)	**87.20 ± 5.93**	**88.95 ± 3.67**	**87.44 ± 8.89**	**85.00 ± 12.36**	**86.19 ± 6.65**
Slice-wise classification	Pereira et al. [15]	67.78 ± 5.60	69.76 ± 12.34	68.00 ± 6.90	65.26 ± 8.25	66.26 ± 4.85
	Anis et al. [16]	84.14 ± 4.09	83.18 ± 6.01	85.34 ± 5.39	85.08 ± 5.04	83.96 ± 3.81
	Gautam et al. [17]	56.18 ± 3.87	54.02 ± 6.67	59.83 ± 5.31	70.42 ± 12.31	60.51 ± 6.76
	Proposed (5S-CNN)	**90.51 ± 2.22**	**90.72 ± 5.38**	**91.38 ± 5.71**	**90.15 ± 7.77**	**90.11 ± 2.78**

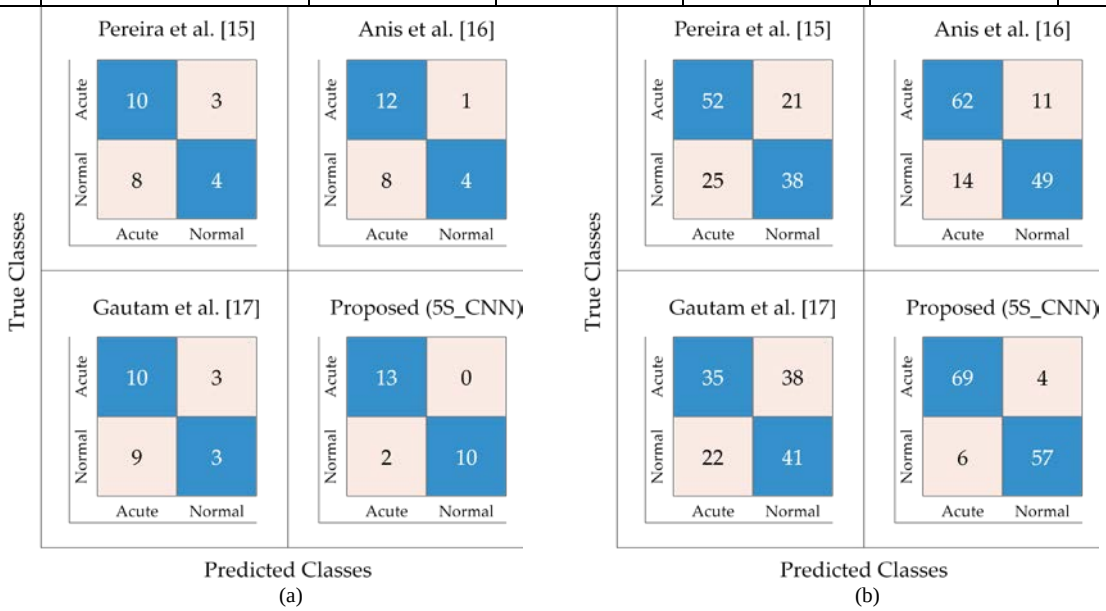


Fig. 11. Confusion Matrices of the Optimal Testing Fold from the Five Cross-Validated Folds: (a) Patient-Wise Confusion Matrices; (b) Slice-Wise Confusion Matrices.

VII. CONCLUSION

This paper proposed a novel, automated approach for acute ischemic stroke classification in brain CT images. A pre-processing technique that can be applied to the CT scans of patients with suspected ischemic stroke was presented, including the removal of CT slices that do not contain brain tissue, segmentation of brain tissue within the remaining slices, ELEHE contrast enhancement of the segmented brain tissue, and classification of brain tissue as an ischemic stroke or normal. A lightweight multi-scale CNN model (5S-CNN) was proposed for brain tissue classification on CT slices, to determine whether the patient is experiencing an ischemic stroke. Notably, this novel model outperformed state-of-the-art models. The model uses a 5-branch architecture, with different filter sizes for each branch, to learn features at different scales, which is crucial because ischemic stroke can have any regional size and appear at any location within brain tissue. A comparison with similar methods revealed that the proposed method outperforms the best-known current methods. The main focus of this research is to identify the presence of acute ischemic stroke on CT slices automatically. However, stroke lesion segmentation is essential for treatment decisions and management. We intend to investigate this area further in the future, ideally exploiting an expansion in the size of the existing collected datasets and samples.

ACKNOWLEDGMENT

The authors are thankful to the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia for funding this work through the research group no. RGP-1439-067. We would also like to thank Dr. Shanker Raja from King Fahad Medical City (KFMC) for his support during this research, as well as Dr. Abdulaziz AlSaad, Dr. Juman AlGhamdi, and Dr. Abeer AlDhawi from the Neurology Department at KFMC.

REFERENCES

[1] World Health Organization, "The top 10 causes of death," 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] N. H. Rajini and R. Bhavani, "Computer aided detection of ischemic stroke using segmentation and texture features," Measurement, pp. 1865-1874, 2013.

[3] A. R. Asirvatham and M. Z. Marwan, "Stroke in Saudi Arabia: a review of the recent literature," Pan African Medical Journal, vol. 17, no. 1, 2014.

[4] F. Al-Senani, M. Al-Johani, M. Salawati, A. Alhazzani, L. B. Morgenstern, V. S. Ravest, M. Cuche and S. Eggington, "An Epidemiological Model for First Stroke in Saudi Arabia," Journal of Stroke and Cerebrovascular Diseases, vol. 29, no. 1, p. 104465, 2020.

[5] K. K. Andersen, T. S. Olsen, C. Dehlendroff and L. P. Kammersgaard, "Hemorrhagic and Ischemic Strokes Compared: Stroke Severity, Mortality, and Risk Factors," Stroke: , pp. 2068-2072, 2009.

[6] A. R. Xavier, A. I. Qureshi, J. F. Kirmani, A. M. Yahia and R. Bakshi, "Neuroimaging of Stroke: A Review," Southern Medical Journal, vol. 96, no. 4, pp. 367-379, 2003.

[7] C. K. Hansen, A. Christensen, H. Rodgers, I. Havsteen, C. Kruuse and H. Christensen, "Does the Primary Imaging Modality—Computed Tomography or Magnetic Resonance Imaging—Influence Stroke Physicians' Certainty on Whether or Not to Give Thrombolysis to Randomized Acute Stroke Patients?," Journal of Stroke and Cerebrovascular Diseases, 2017.

[8] Z. Calic, C. Cappelen-Smith, C. S. Anderson, W. Xuan and D. J. Cordato, "Cerebellar infarction and factors associated with delayed presentation and misdiagnosis," Cerebrovascular Diseases, vol. 42, no. 5-6, pp. 476-484, 2016.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. v. d. Laak, B. v. Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[10] G. Kaur and J. Chhaterji, "A Survey on Medical Image Segmentation," International Journal of Science and Research (IJSR), vol. 6, no. 5, pp. 2319-7064, 2017.

[11] H. R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa and K. Mori, "Deep learning and its application to medical image segmentation," Medical Imaging Technology, vol. 36, no. 2, pp. 63-71, 2018.

[12] C.-M. Lo, P.-H. Hung and K. L.-C. Hsieh, "Computer-aided detection of hyperacute stroke based on relative radiomic patterns in computed tomography," Applied Sciences, vol. 9, no. 8, p. 1668, 2019.

[13] G. Wu, J. Lin, X. Chen, Z. Li, Y. Wang, J. Zhao and J. Yu, "Early identification of ischemic stroke in noncontrast computed tomography," Biomedical Signal Processing and Control, vol. 52, pp. 41-52, 2019.

[14] C.-L. Chin, B.-J. Lin, G.-R. Wu, T.-C. Weng, C.-S. Yang, R.-C. Su and Y.-J. Pan, "An Automated Early Ischemic Stroke Detection System using CNN Deep Learning Algorithm," 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), pp. 368-372, 2017.

[15] D. R. Pereira, P. P. R. Filho, G. H. d. Rosa, J. P. Papa and V. H. C. d. Albuquerque, "Stroke Lesion Detection Using Convolutional Neural Networks," 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-6, 2018.

[16] A. A. M. Suberi, W. N. W. Zakaria, R. Tomari, A. Nazari, M. N. H. Mohd and N. F. N. Fuad, "Deep Transfer Learning Application for Automated Ischemic Classification in Posterior Fossa CT Images," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 8, pp. 459-465, 2019.

[17] A. Gautam and B. Raman, "Towards effective classification of brain hemorrhagic and ischemic stroke using CNN," Biomedical Signal Processing and Control, vol. 63, p. 102178, 2021.

[18] X. Bai, Y. Zhang, F. Zhou and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," Information Fusion, vol. 22, pp. 105-118, 2015.

[19] S. Zhang, M. Zhang, S. Ma, Q. Wang, Y. Qu, Z. Sun and T. Yang, "Research Progress of Deep Learning in the Diagnosis and Prevention of Stroke," BioMed Research International, 2021.

[20] M. S. Sirsat, E. Ferme and J. Camara, "Machine Learning for Brain Stroke: A Review," Journal of Stroke and Cerebrovascular Diseases, vol. 29, no. 10, p. 105162, 2020.

[21] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[22] L. M. Allen, A. N. Hasso, J. Handwerker and H. Farid, "Sequence-specific MR Imaging Findings That Are Useful in Dating Ischemic Stroke," Radiographics, vol. 32, no. 5, pp. 1285-1297, 2012.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[24] K. H. X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[25] P. Soille, Morphological Image Analysis: Principles and Applications, Springer Science & Business Media, 2004.

[26] W. Shi and H. Liu, "Modified U-Net Architecture for Ischemic Stroke Lesion Segmentation and Detection," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC 2019), vol. 1, pp. 1068-1071, 2019.

[27] A. Manvel, K. Vladimir, T. Alexander and U. Dmitry, "Radiologist-Level Stroke Classification on Non-contrast CT Scans with Deep U-Net," International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 820-828, 2019.

[28] H. Kuang, B. K. Menon and W. Qiu, "Segmenting Hemorrhagic and Ischemic Infarct Simultaneously From Follow-Up Non-Contrast CT Images in Patients With Acute Ischemic Stroke," IEEE Access, vol. 7, pp. 39842-39851, 2019.

[29] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015.

[30] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Pearson Education Limited, 2018.

[31] B. Kurt, V. V. Nabiyev and K. Turhan, "Medical Images Enhancement by using Anisotropic Filter and CLAHE," 2012 International Symposium on Innovations in Intelligent Systems and Applications, pp. 1-4, 2012.

[32] S.-C. Huang, F.-C. Cheng and Y.-S. Chiu, "Efficient Contrast Enhancement Using Adaptive Gamma Correction With Weighting Distribution," IEEE TRANSACTIONS ON IMAGE PROCESSING, vol. 22, no. 3, pp. 1032-1041, 2013.

[33] Z. Al-Ameen and G. Sulong, "A New Algorithm for Improving the Low Contrast of Computed Tomography Images Using Tuned Brightness Controlled Single-Scale Retinex," Scanning, vol. 37, no. 2, pp. 116-125, 2015.

[34] T. V, K. S. Sim and E. K. Wong, "Brain Early Infarct Detection Using Gamma Correction Extreme-Level Eliminating With Weighting Distribution," Scanning, vol. 38, no. 6, pp. 842-856, 2016.

[35] T.-L. Tan, K.-S. Sim and A.-K. Chong, "Contrast Enhancement of CT Brain Images for Detection of Ischemic Stroke," 2012 International Conference on Biomedical Engineering (ICoBE), pp. 385-388, 2012.

[36] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation," Journal of Big Data, vol. 6, no. 1, pp. 1-48, 2019.

[37] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning," Radiology, vol. 295, no. 1, pp. 4-15, 2020.

Learning Cultural Heritage History in Muzium Negara through Role-playing Game

Nor Aiza Moketar[1], Nurul Hidayah Mat Zain[2], Siti Nuramalina Johari[3], Khyrina Airin Fariza Abu Samah[4]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Melaka
Kampus Jasin, Melaka, Malaysia

Lala Septem Riza[5]
Department of Computer Science Education
Universitas Pendidikan Indonesia, Bandung, Indonesia

Massila Kamalrudin[6]
Innovative Software System & Services
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

***Abstract*—The traditional classroom-based teaching and learning of the History subject are ineffective and less interactive, influencing the students' interest and motivation to learn history. Therefore, museum-based learning was proposed to supplement classroom-based learning for effective teaching and learning of the History subject. However, the excursions to the museum are often hindered by issues caused by the geographical location, the museum's policies, and student commitments. The hindrances motivated the researchers to design and develop a role-playing game (RPG) in Muzium Negara (National Museum of Malaysia) known as 'Waktu Silam' to enhance students' interest, motivation, and knowledge on the cultural and historical heritage of Malaysia. A survey questionnaire was distributed to assess the enjoyment level provided by the game. The results showed that 84.8% of participants had experienced the element of enjoyment in this game. This study anticipated enhancing the student interest and knowledge in history, enhancing visitors' experience, and promoting tourism to Muzium Negara. Additionally, the project is expected to include multiplayer functionality to add more interactivity to the game in future works.**

***Keywords*—Muzium Negara; history; role-playing games; gamification; museum-based learning; enjoyable**

I. Introduction

The History subject was made compulsory and a core subject for students by the Malaysian Ministry of Education in 1989 [1]. The subject has become a compulsory-pass subject for secondary school students sitting for the Malaysian Certificate of Education examination (Sijil Peperiksaan Malaysia/SPM) since 2013 [2][3][4]. This step proved that the History subject is vital for several academic reasons. The History subject's main objective is to foster patriotism spirit towards the nation, creating the spirit of unity among people of different races and embedding a sense of pride to be Malaysians among students [1][2][3][4]. The content of the curriculum has undergone rapid changes since the colonial era until Malaysia achieved independence, and the entire chronology has been recorded in the syllabus.

The syllabus of the History subject in the textbook comprises many facts and figures. Students need to memorize the facts to establish the connection between the chronology and to understand the concept of history [2][5]. As a result, the learners feel disinterested and unenjoyable in learning history as they cannot sense or comprehend the significance of such historical events [6]. Furthermore, the traditional classroom-based learning when teaching history is a teacher-centred approach that is proven ineffective and less interactive in enhancing students' interest, motivation, and knowledge [7]. Traditional storytelling when teaching history can cause boredom among the students.

Additionally, a study by Jaafar and Mohd Noor [7], and Napiah, Awang, Ahmad and Che Dahalan [2] stated that learning history is abstract in nature compared to visiting museums or historical places that are more realistic [2][8]. Students will find it more meaningful and practical to understand the history if they gain the experience and engage with the artefacts in the museum. Napiah et al. [2] showed that learning history in a real museum setting contributes a positive effect and helps students understand history better than in the classroom setting [2]. This scenario is an effective way to assist students in remembering chronological history without solely memorizing the facts.

Museum-based learning was proposed to supplement classroom-based learning for effective teaching and learning in history [2][9]. Students are exposed to the genuine historical artefacts exhibited in the museum through this approach. They will gain new learning experiences by seeing, touching, and interacting with historical materials that are not available in classroom settings. Hands-on history learning will also be more fun and interesting. Furthermore, learning history in a museum setting also enhances the students' understanding of the historical concepts taught in the classroom [8][10]. However, the excursions to the museum are often hindered by several issues due to the geographical location, the museum's policies, and student commitments. Therefore, not every student will obtain the opportunity to visit the museum to explore and experience the exhibition.

The issue has motivated the current study's researchers to develop an RPG known as 'Waktu Silam' to give the targeted users the sense and enjoyment of virtually learning the historical content in the national museum of Malaysia, Muzium Negara. 'Waktu Silam' is in the Malay language, which means "the old days". The researchers implemented game-based learning through the RPG genre to learn historical contents in

This study is sponsored by the Indonesian Ministry of Research and Technology and Universitas Pendidikan Indonesia.

Muzium Negara. Game-based learning is deemed an exciting and effective way of learning when adequately constructed using learning principles as a goal. Besides, the proposed games can enhance students' motivation, encourage engagement, and promote learning [11]. The nature of history as a subject makes it essential for current and emerging instruments established to support the main delivery of instruction [5]. Players with or without the intention to learn history may find history easy and be excited to learn with this game.

The upcoming sections of this paper begin with Section 2 that outlines the study's background that focuses on Muzium Negara's overview, the introduction of game-based learning and RPG, and related works. Subsequently, Section 3 presents the study's design and development, Section 4 elaborates on the results and discussions, while Section 5 concludes the paper.

II. Background of Study

A. Muzium Negara

Muzium Negara is the national museum of Malaysia, strategically situated in the heart of the capital city of Kuala Lumpur. The museum was built resembling the Malay palace-style as a symbol of the guardian of the nation's history. The structure is three-storeys high with 109.7 metres long, 15.1 metres wide, and 37.6 metres at the central point. The building was officially opened on 31st August 1963. The galleries in the museum have been upgraded to represent an exhilarating and state-of-the-art approach to exploring Malaysia's history from the pre-historical era to the present time.

The museum houses four main exhibition galleries, as shown in Fig. 1, namely, the Prehistoric Era, Malay Kingdoms, Colonial Era, and Malaysia Today galleries [12][13]. The Prehistoric Era gallery showcases the evolution of the earth's surface until the origin of Malaysia's earliest inhabitant. Additionally, the gallery outlines the discovery of the Palaeolithic age (200000 years ago) stone tools to the Hindu-Buddhist temple and relics found in Lembah Bujang. The Malay Kingdom gallery traces the historical timeline of the first Malay Kingdoms in the archipelago, specifically the Malay Peninsula. The gallery also highlighted the glorifying days of the Malacca Malay Kingdom in the 15th century.

Conversely, the exhibition presents the historical chronicles of the control and administration of foreign powers: the Portuguese, Dutch, British, and Japanese, and the subsequent effects on the nation's political, social, and economic situations in the Colonial Era gallery. Finally, the Malaysia Today gallery walks the visitors through the arduous path trodden by the peoples' relentless struggle for independence and the formation of a new nation. The gallery demonstrates the transformation process and achievement gained since Malaya's independence in 1957.

Muzium Negara welcomes many tourists daily, including foreigners and Malaysian students. The museum inspires a more comprehensive understanding of Malaysia and its multiracial composition through collections, exhibitions, research, publications, educational and public programmes. Muzium Negara also bears the responsibility of nurturing, protecting, and publishing information on cultural history and natural heritage. Thus, Muzium Negara plays a crucial role in history authentication and preservation.

B. Game-based Learning and Role-Playing Game

Game-based learning (RPG) is considered an effective approach for teaching and learning purposes. The application of digital games in education helps to develop interest and motivate learners [14]. It can motivate students to focus their attention on education in an enjoyable and engaging way. In addition, game-based learning also provides an opportunity for students to experience a new learning method where students can be more active in learning sessions. Students gain a chance to experience, take risks, and learn without fear of failure from real-life consequences [11]. Some studies established that games could produce positive learning outcomes, efficient in promoting learning and retention, as well as provide an engaging experience compared to traditional instructional methods [15][16]. For example, a study by Wan Fatimah, Afza, and Mohd Hezri Amir [17] have developed an RPG prototype named Maths Quest to engage and assist children in learning mathematics. The heuristic evaluation in terms of learnability, satisfaction, screen design and performance effectiveness has been conducted and received positive feedback from the participants. A different study by Sung and Hwang [18] also proposed a game-based learning approach to guide students in science courses. The result from the experiment shows that the game helps in promoting students' learning attitudes and motivation as well as improving their learning achievement and self-efficacy.

The RPG is among game genres where the player controls a fictional character (or characters) that embarks on a pursuit in a fantasy world. The RPGs varied range share a parallel emphasis by providing the player with a role that advances and progresses via playing and storytelling experience. However, a robust variation exists across and within the format. The RPG also offers a remarkable opportunity to assess most of the crucial questions in-game studies [14]. Besides, RPGs involve a good storyline and interactions with other objects or characters. The storylines and interactions will provide users with the intended experience made by the developer. RPGs can also be used to develop educational games with such features.

The RPGs provide learning environments that engage collaborative problem solving and distribute apprenticeship when playing and influence digital and print literacy advancement, besides science, math, and computational literacy. Hammer, Schrier, Bowman, and Kaufman [19] highlighted that the RPGs are deeply related to constructivism and sociocultural learning theory. The learning in constructivism happens through hands-on experimentation with a new situation, whereas the learning in sociocultural takes place through the adoption of new social roles. The underlying educational theories for each feature makes it appropriate for learning and explain how it is employed in education [19]. These reasons denote that RPGs can be considered a good educational platform although underrated.

Fig. 1. The Four Main Galleries in Muzium Negara – (a) Prehistoric Era, (b) The Entrance view of Malay Kingdom Gallery, (c) The Portuguese Section in the Colonial Era Gallery, and (d) Multiracial Section in the Malaysia Today Gallery.

III. Related Work

Devising simulated worlds generates an illustrious, sophisticated approach to the design concepts behind virtual worlds. A virtual environment is denoted as the most frequently utilized alternate phrase for platforms otherwise acknowledged as virtual worlds [20]. An avatar or character that mimics the user in the virtual environment can facilitate the user to observe the virtual three-dimensional (3D) environment and induce immersion towards the user. The use of avatar is the medium that allows users to manoeuvre objects in the virtual environment [21]. The virtual environment can be categorized into two general types: multi-user virtual environment (MUVE) and Massively Multiplayer Online Role-Playing Game (MMORPG). According to Döpker, Brockmann and Stieglitz [22], the significant dissimilarity in MUVEs is that the users do not have a specific goal to attain or a start-finish character as in MMORPGs. Thus, integrating RPG elements into a virtual museum can generate users' interest in learning cultural heritage in engaging ways.

The museum's virtualization includes tasks mapping that must be undertaken in a virtual reality museum context for individual tasks in a 3D game [23]. A study by Prasetyo and Suyoto [24] found that implementing the gamification methods at museums can encourage the community to be motivated to visit the museum. Similarly, a study by Araujo, Koenigschulte, and Erb [25] also revealed that implementing interactive games for museum environments can enhance visitor experiences. Another research from López-Martínez, Carrera, and Iglesias [26] found that applying game concepts to museums can contribute to higher levels of cognitive engagement among users. A study by Cosovi´c & Brki´c [27] also shows that game-based learning is a resolution to transform a traditional museum into a virtual museum and encourage active learning where game-based learning can assist in preserving the cultural heritage. All these approaches integrate game elements into virtual reality museums.

Lepouras and Vassilakis [23] pointed out that education through entertainment is crucial to enhancing user engagement in learning. Approaches utilizing game-integrated entertainment aspects can enhance the user's learning process. Apart from that, digital games can also be used as a persuasive tool to persuade people to learn or improve their knowledge [28]. Therefore, game elements have been applied in various subjects such as Islamic teaching [28], programming [29], and history [30]. In term of History subject, a study by Lee, Talib, Zainon, and Lim [30] have designed and implemented a framework using RPG on the mobile application. The implementation is based on the narrative story of Merong Mahawangsa, a legendary Malaysian warrior. In contrast, our work allows the player to learn historical knowledge from Malaysia's national museum virtually.

IV. Methodology

The researchers implemented the Agile Development Methodology for the development of this project. The phases of the agile development model involved Plan, Design, Develop, Test, Release, and Feedback. Although relatively new, the methodology helps create the project based on thorough evaluation and understanding. The methodology is

based on implementation over the documentation with consumer involvement and can solve both problems and agility adjustments. Besides, the methodology uses and works through an iterative development framework known as Scrum. Scrum functions in the game development methodology by breaking down game development into a series of tasks named "sprints". The game developers split the game into clusters of associated jobs or features to ease the work with sprints. Additionally, the model stresses small phases with minimal planning and gives moderate access to frequent change requirements in the processes.

The following is the brief explanations of each phase on the Agile Development Methodology Fig. 2:

- Plan: At this stage, the researchers go through the brainstorming session to idealize the game concept, scope and goal. The initial user requirements were also analyzed and documented. The researchers also identified the software and hardware requirements for the development of the project.
- Design: The researchers focused on the plot and storyboard design as defined in the planning stage at the design stage. The details of the design phase are discussed in the next sub-section.
- Develop: The project's development was started once the plot and storyboard designs were agreed upon and confirmed.
- Test: In the testing stage, the verification and validation of the developed project were conducted. The testing ensured that the project had fulfilled the requirements and eliminated any errors or bugs. It is to prevent any undesirable issues from occurring.
- Release: The project was released into production once the testing phase was cleared.
- Feedback: At this stage, the researcher used the convenience sampling method to collect the users' feedback about the developed project. The researchers have distributed survey questionnaires to young adults aged between 17 and 25 years old. The questions were adapted from the eGameFlow model [31]. There are eight dimensions of scale in this model which are the Concentration, Clear Goal, Feedback, Challenge, Autonomy, Immersion, Social Interaction and Knowledge Improvement. Table I shows the description of each dimension.

A. Design Phase

The design phase focused on the plot and storyboard design. The researchers applied the narrative structure mechanism guided by Freytag's pyramid to design the game's plotline. Freytag's pyramid was initially identified from a successful theatrical tragedy and has been widely applied by game designers [32]. The narratives mechanism is extensively utilized in RPG and plays a crucial role in developing an engaging and meaningful game [33]. Based on Freytag's pyramid, the stories can be segregated into five acts: Exposition, Rising Action, Climax, Falling Action, and Conclusion. Table II describes the details of each act's narrative purpose in Freytag's pyramid. The researchers adopted Freytag's pyramid to design the narrative structure of the game plot as in Fig. 3.

The game's overall acts are displayed in Fig. 4. First, the player can choose to start the game from the main menu. Then, a cut-scene that introduces the main character and the mission to be accomplished is shown. Next, the player will be prompted to begin the journey to Muzium Negara and explore the exhibition galleries. The player can explore any gallery in the Muzium Negara and interact with the exhibited artefacts. Detailed history information is displayed for each artefact. The game's climax starts when the player is suddenly knocked out and enters a new world, the Prehistoric Era. A non-player character in this new imaginary world will explain the mission to be accomplished to exit the world. Pop-up details are available along the mission mentioning the world's history and the artefacts found. The players can choose to proceed to the next level or end the game once the mission is completed. The following levels include the imaginary world related to the galleries in Muzium Negara, namely, the Malay Kingdom, Colonial Era, and Malaysia Today.

TABLE I. THE DESCRIPTION OF SCALE DIMENSION IN THE EGAMEFLOW MODEL

Scale Dimension	Description
Concentration	The game must include activities that stimulate the player's concentration while avoiding the stress associated with learning overload, which might cause the player to lose focus on the game.
Clear Goal	The game's objectives and tasks should be clearly stated at the beginning.
Feedback	Feedback enables a player to evaluate the knowledge gap between their present level of understanding and the knowledge required to complete the game's task.
Challenge	The game should include challenges that are appropriate for the player's skill level, with the difficulty of these challenges increasing as the player's skill level improves.
Autonomy	The learner should enjoy taking the lead in game play and having complete control over his or her decisions.
Immersion	The game should lead the player into a state of immersion.
Social Interaction	The game's tasks should become a way for users to socialise.
Knowledge improvement	The game should improve the player's level of knowledge and skills while meeting the goal of the curriculum.

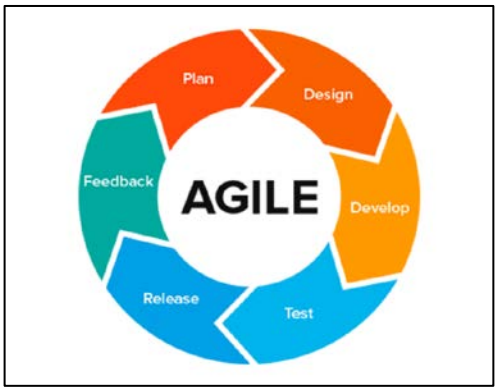


Fig. 2. Agile Development Methodology.

TABLE II. THE NARRATIVE PURPOSE OF EACH ACT IN FREYTAG'S PYRAMID

Act	**Narrative Purpose**
Exposition	Describes the setting, protagonist, and primary conflict. Ends with the inciting moment that drives the story forward.
Rising Action	Develops the primary conflict and relates with secondary conflicts. The protagonist overcomes minor obstacles.
Climax	The story's turning point. The exact nature of the primary conflict crystallizes, the antagonist is revealed, and the path ahead is made clear.
Falling Action	Resolves the primary conflict. Leads to a moment of final suspense, where the outcome is in doubt.
Conclusion	The story returns to a state of normality. Ties up loose ends.

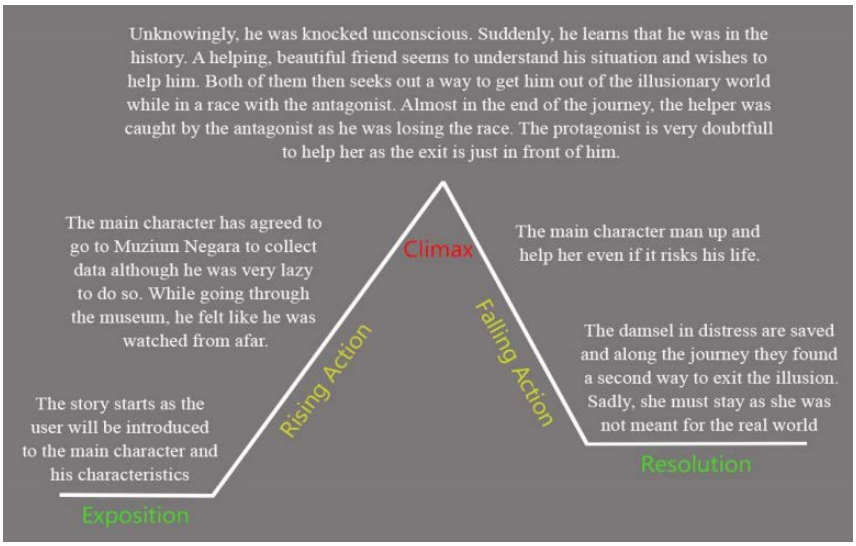


Fig. 3. The Plot Design.

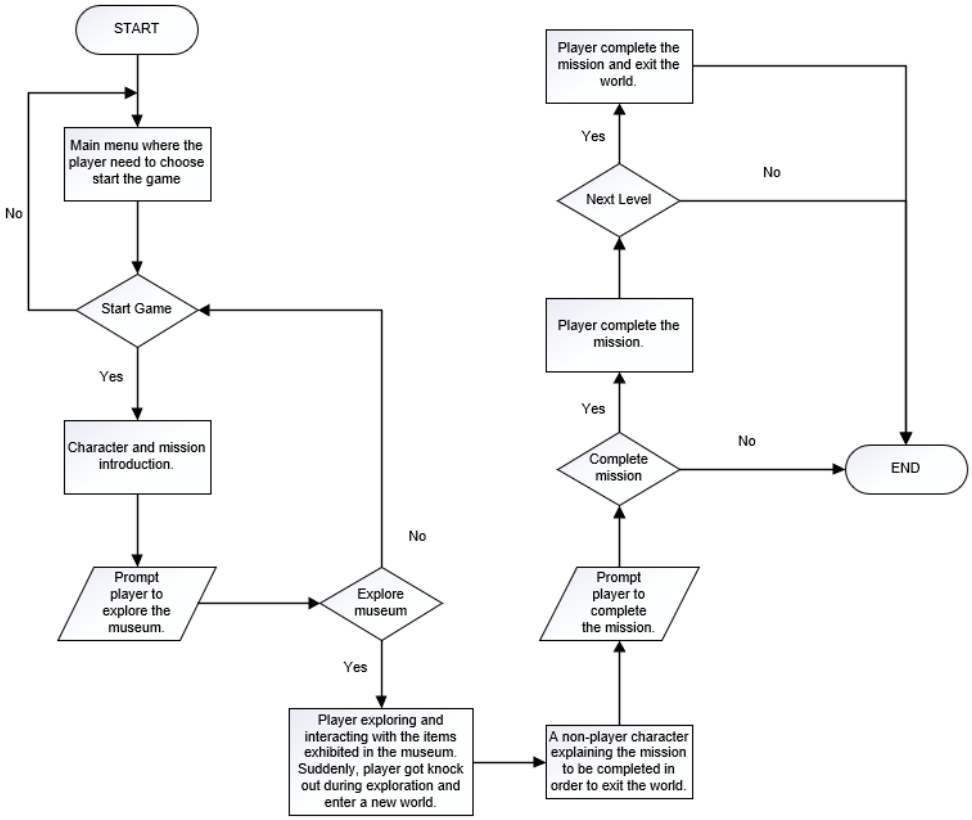


Fig. 4. The Overall Flow of the Game.

B. Development Phase

The project's development started with the plot's completion and storyboard design. All the project's models, including the museum and artefacts, were developed using Blender, an open-source 3D creation suite. Fig. 5 shows the museum's exterior design development and its artefacts modelled using Blender. Conversely, the environment scene for the game was developed using Unity, a cross-platform game engine developed by Unity Technologies. Fig. 6 shows the glimpses of the scene produced during the design and development phase.

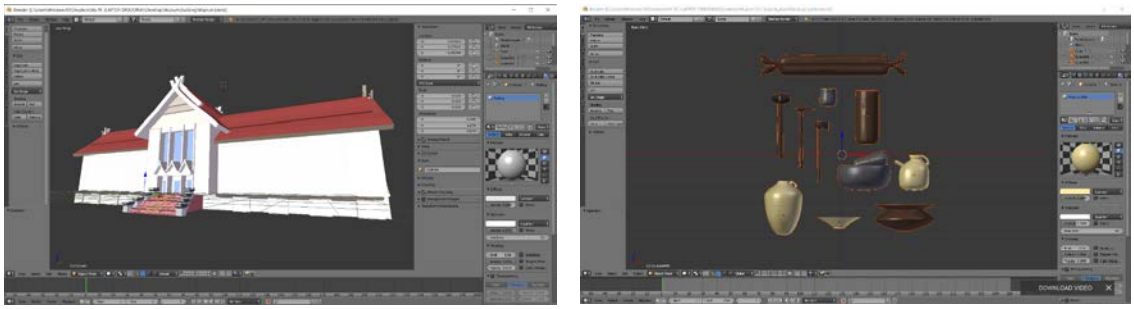
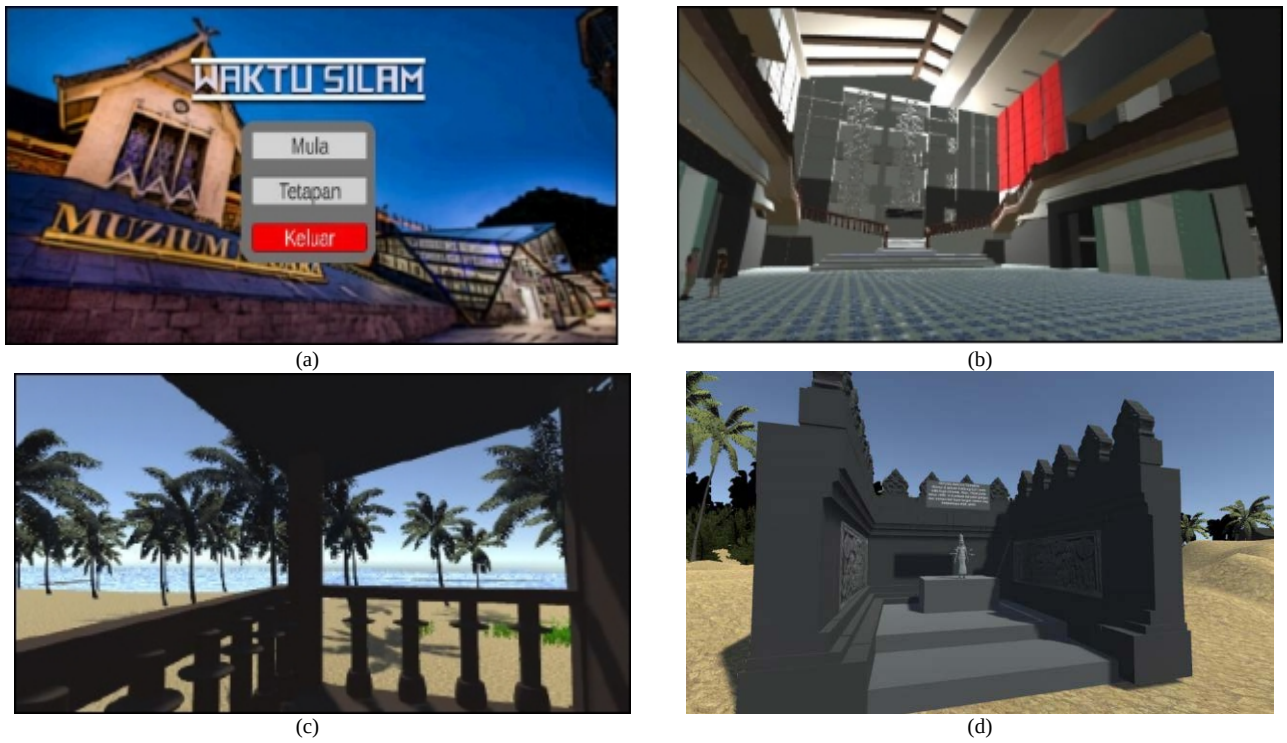


Fig. 5. The Development of 3D Models using Blender.



(a) (b)

(c) (d)

Fig. 6. The Scene Environment Developed using Unity – (a) The Exterior Design of Muzium Negara, (b) The Interior Lobby of the Museum, (c) The Malay Kingdom Scene and (d) The Artefacts and Monument.

V. Result and Discussion

The researchers adapted the eGameFlow model to evaluate the performance of the developed game. eGameFlow model acts as a scale to measure the respondents' enjoyment experience through its enjoyment factors. The factors adapted incorporate concentration, goal clarity, feedback, challenge, control, immersion, social interaction, and knowledge improvement. Nevertheless, social interaction was not included as the evaluation element because the games were neither developed nor created for multiplayer.

According to the survey results, 32 respondents took part in the evaluation. The majority of respondents were aged 21 to 25 years old, whereas 23 out of 32 respondents (71.9%) were male. Table II presents the average mean for each element when the descriptive analysis was performed. Next, the analysis calculated the average mean of each element into a total average to determine the evaluation's outcome. The total mean is 4.24, or 84.8% of the respondent's enjoyment while playing the game. The analysis results from the survey demonstrate that most respondents agreed that the '*Waktu Silam'* game is enjoyable to be played. In addition, the '*Waktu Silam'* game was identified to efficiently increase the knowledge on knowing and memorizing history content in the museum. Furthermore, this game can enhance the player's knowledge about Malaysia's history.

TABLE III. RESULT OF THE AVERAGE MEAN FOR EACH EGAMEFLOW ELEMENTS

Elements	Item	Questions	Mean	Average Mean
Concentration	C1	The game grabs my attention	4.21	4.13
	C2	The game provides content that stimulates my attention	4.21	
	C3	Most of the gaming activities are related to the learning task	4.28	
	C4	No distraction from the task is highlighted	3.93	
	C5	Generally speaking, I can remain concentrated in the game	4.06	
	C6	I am not distracted from tasks that the player should concentrate on	4.19	
	C7	I am not burdened with tasks that seem unrelated	4.09	
	C8	Workload in the game is adequate	4.06	
Goal Clarity	G1	Overall game goals were presented in the beginning of the game	4.09	4.16
	G2	Overall game goals were presented clearly	4.16	
	G3	Intermediate goals were presented in the beginning of each scene	4.03	
	G4	Intermediate goals were presented clearly	4.28	
	G5	I understand the learning goals through the game	4.22	
Feedback	F1	I receive feedback on my progress in the game	4.19	4.10
	F2	I receive immediate feedback on my actions	3.81	
	F3	I am notified of new tasks immediately	4.09	
	F4	I am notified of new events immediately	4.06	
	F5	I receive information on my success (or failure) of intermediate goals immediately	4.25	
	F6	I receive information on my status, such as score or level	4.19	
Challenge	H1	I enjoy the game without feeling bored or anxious	4.38	4.12
	H2	The challenge is adequate, neither too difficult nor too easy	4.09	
	H3	The game provides ‘hints’ in text that help me overcome the challenges	4.06	
	H4	The game provides ‘online support’ that helps me overcome the challenges	3.67	
	H5	The game provides video or audio auxiliaries that help me overcome the challenges	4.19	
	H6	My skill gradually improves through the course of overcoming the challenges	4.19	
	H7	I am encouraged by the improvement of my skills	4.16	
	H8	The difficulty of challenges increases as my skills improved	3.94	
	H9	The game provides new challenges with an appropriate pacing	4.19	
	H10	The game provides different levels of challenges that tailor to different players	4.28	
Autonomy	A1	I feel a sense of control the menu (such as start, stop, save, etc.)	4.84	4.63
	A2	I feel a sense of control over actions of roles or objects	4.34	
	A3	I feel a sense of control over interactions between roles or objects	4.69	
	A4	The game does not allow players to make errors to a degree that they cannot progress in the game	4.59	
	A5	The game supports my recovery from errors	4.15	
	A6	I feel that I can use strategies freely	4.69	
	A7	I feel a sense of control and impact over the game	4.69	
	A8	I know next step in the game	4.84	
	A9	I feel a sense of control over the game	4.84	
Immersion	I1	I forget about time passing while playing the game	4.09	4.23
	I2	I become unaware of my surroundings while playing the game	4.38	
	I3	I temporarily forget worries about everyday life while playing the game	4.34	
	I4	I experience an altered sense of time	4.15	
	I5	I can become involved in the game	4.22	
	I6	I feel emotionally involved in the game	4.38	
	I7	I feel viscerally involved in the game	4.06	
Knowledge Improvement	K1	The game increases my knowledge	4.06	4.31
	K2	I catch the basic ideas of the knowledge taught	4.06	
	K3	I try to apply the knowledge in the game	4.38	
	K4	The game motivates the player to integrate the knowledge taught	4.69	
	K5	I want to know more about the knowledge taught	4.38	
		Average Mean		**4.24 (84.8%)**

VI. Conclusion

The researchers designed and developed an RPG named '*Waktu Silam'* as an educational game where the player can virtually learn the historical content in Muzium Negara. The main objective for the game's development is to enhance students' interest, motivation, and knowledge of the Malaysian cultural and historical heritage. Unlike previous research, this initiative allows the player to virtually engage with the artefact in the Muzium Negara and immerse themselves in the virtual world to learn about the past. The evaluation of the game's enjoyment by adapting the eGameFlow model has been conducted. The results showed that the participants experienced about 84.8% enjoyment while playing the game.

For future work, we plan to include the game's console version and add multiplayer functionality to improve the game's interactivity. Additionally, the researchers also plan to collaborate with Muzium Negara to obtain more information for the game. The collaboration will increase the quality of details and information on the history and enhance the game's quality. The researchers envision that the collaboration will indirectly help boost tourism to Muzium Negara.

Acknowledgment

The authors would like to acknowledge the Indonesian Ministry of Research and Technology and Universitas Pendidikan Indonesia for funding this work.

References

[1] R. Ahmada, A. Rahim, A. A. Seman, and M. J. Salleh, "Malaysian secondary school history curriculum and its contribution towards racial integration," Procedia - Soc. Behav. Sci., vol. 7, no. 2, pp. 488–493, 2010, doi: 10.1016/j.sbspro.2010.10.066.

[2] L. Napiah, M. Awang, A. Razaq Ahmad, and S. Che Dahalan, "Museum Based Learning in History Education to Enhance Patriotism among Students," in Proceedings of The 2nd International Conference on Sustainable Development & Multi-Ethnic Society, 2019, vol. 2, pp. 95–99, doi: 10.32698/GCS.0178.

[3] K. G. Kaspin, M. M. Noor, and M. M. Awang, "Perspektif Pelajar Terhadap Kurikulum Sejarah Peringkat Menengah di Malaysia," J. Pemikir Pendidik., vol. 9, no. December 2018, pp. 12–31, 2018.

[4] N. Syazwani and A. Talib, "Kaedah Pembelajaran Sejarah Berasaskan Lawatan Ke Muzium History Learning Method Based on Museum Visits," Insa. Online J. Lang. Commun. Humanit., vol. 2, no. June, pp. 45–57, 2019.

[5] V. Zirawaga, A. Olusanya, and T. Maduki, "Gaming in education: Using games a support tool to teach History," J. Educ. Pract., vol. 8, no. 15, pp. 55–64, 2017, [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1143830.pdf.

[6] G. P. Kusuma, L. K. Putera Suryapranata, E. K. Wigati, and Y. Utomo, "Enhancing Historical Learning Using Role-Playing Game on Mobile Platform," Procedia Comput. Sci., vol. 179, no. 2019, pp. 886–893, 2021, doi: 10.1016/j.procs.2021.01.078.

[7] C. Chee-Huay and Y. Kee-Jiar, "Why Students Fail in History: A Minor Case Study in Malaysia and Solutions from Cognitive Psychology Perspective," Mediterr. J. Soc. Sci., vol. 7, no. 1, pp. 517–526, Dec. 2016, doi: 10.5901/mjss.2016.v7n1p517.

[8] S. A. Jaafar and A. Mohd Noor, "Pelaksanaan Pengajaran Dan Pembelajaran Sejarah Di Sekolah-Sekolah Di Malaysia, 1957 - 1989," SEJARAH, vol. 25, no. 2, pp. 40–57, Dec. 2016, doi: 10.22452/sejarah.vol25no2.3.

[9] A. B. S. Kechot, "Proses Pendidikan Muzium: Satu Kajian Awal," J. Melayu, vol. 5, no. December 2009, pp. 285–293, 2010.

[10] M. I. H. S. Azman Ligun, Mohd Mahzan Awang, Abdul Razaq Ahmad, "Muzium Sebagai Instrumen Pembelajaran Sejarah Luar Bilik Darjah," J. Kurikulum Pengajaran Asia Pasifik, vol. 5, no. 1, pp. 19–30, 2017.

[11] A. Pho and A. Dinscore, "Game-Based Learning Overview and Definition," Tips Trends Instr. Technol. Commitee, no. Spring 2015, pp. 1–5, 2015, [Online]. Available: https://acrl.ala.org/IS/wp-content/uploads/2014/05/spring2015.pdf.

[12] C. Carlos Augusto Bahamón, "Strategies for significant learning in the museum, based on interactive experiences," ACM Int. Conf. Proceeding Ser., pp. 2–5, 2019, doi: 10.1145/3358961.3358989.

[13] "Department of Museums Malaysia." http://www.jmm.gov.my/en/museum/muzium-negara (accessed Jun. 01, 2021).

[14] J. M. Spector, "Emerging educational technologies: Tensions and synergy," J. King Saud Univ. - Comput. Inf. Sci., vol. 26, no. 1, pp. 5–10, 2014, doi: 10.1016/j.jksuci.2013.10.009.

[15] J. P. Rowe, E. V. Lobene, B. W. Mott, and J. C. Lester, "Play in the museum: Design and development of a game-based learning exhibit for informal science education," Int. J. Gaming Comput. Simulations, vol. 9, no. 3, pp. 96–113, 2017, doi: 10.4018/IJGCMS.2017070104.

[16] G. Petri, C. G. von Wangenheim, J. C. R. Hauck, and A. F. Borgatto, "Effectiveness of games in software project management education: An experimental study," J. Univers. Comput. Sci., vol. 25, no. 7, pp. 840–864, 2019.

[17] A. Wan Fatimah, W, S. Afza, and A. L. Mohd Hezri Amir, "Role-playing game-based learning in Mathematics.," Electron. J. Math. Technol., vol. 4, no. 2, pp. 185–196, 2010.

[18] H.-Y. Sung and G.-J. Hwang, "A collaborative game-based learning approach to improving students' learning performance in science courses," Comput. Educ., vol. 63, pp. 43–51, Apr. 2013, doi: 10.1016/j.compedu.2012.11.019.

[19] J. Hammer, A. To, K. Schrier, S. L. Bowman, and G. Kaufman, "Learning and Role-Playing Games," in Role-Playing Game Studies, no. April, New York: Routledge, 2018.: Routledge, 2018, pp. 283–299.

[20] C. Girvan, "What is a virtual world? Definition and classification," Educ. Technol. Res. Dev., vol. 66, no. 5, pp. 1087–1100, Oct. 2018, doi: 10.1007/s11423-018-9577-y.

[21] D. Dewez, L. Hoyet, A. Lécuyer, and F. A. Argelaguet Sanz, "Towards 'Avatar-Friendly' 3D Manipulation Techniques: Bridging the Gap Between Sense of Embodiment and Interaction in Virtual Reality," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021, pp. 1–14, doi: 10.1145/3411764.3445379.

[22] A. Döpker, T. Brockmann, and S. Stieglitz, "Use Cases for Gamification in Virtual Museums," in Proceedings of the Jahrestagung der Gesellschaft für Informatik, 2013, pp. 2308–232.

[23] G. Lepouras and C. Vassilakis, "Virtual museums for all: employing game technology for edutainment," Virtual Real., vol. 8, no. 2, pp. 96–106, Jun. 2004, doi: 10.1007/s10055-004-0141-1.

[24] N. A. Prasetyo and S. Suyoto, "Design Mobile App for Increase the Visitor Museum using Gamification Method," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 16, no. 6, p. 2791, Dec. 2018, doi: 10.12928/telkomnika.v16i6.10384.

[25] L. M. de Araujo, A. Koenigschulte, and U. Erb, "Enhancing Visitors' Experience - A Serious Game for Museum Environment," in Edulearn10: International Conference on Education and New Learning Technologies, 2010, no. May.

[26] A. López-Martínez, álvaro Carrera, and C. A. Iglesias, "Empowering museum experiences applying gamification techniques based on linked data and smart Objects," Appl. Sci., vol. 10, no. 16, 2020, doi: 10.3390/APP10165419.

[27] M. Ćosović and B. R. Brkić, "Game-Based Learning in Museums—Cultural Heritage Applications," Information, vol. 11, no. 1, p. 22, Dec. 2019, doi: 10.3390/info11010022.

[28] M. S. A. Aziz, P. Auyphorn, and M. S. Hamzah, "Exploring the use of digital games as a persuasive tool in teaching Islamic knowledge for muslim children," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 109–113, 2019, doi: 10.14569/ijacsa.2019.0100616.

[29] R. Ibrahim, N. Z. A. Rahim, D. W. H. Ten, R. C. M. Yusoff, N. Maarop, and S. Yaacob, "Student's opinions on online educational games for learning programming introductory," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 352–340, 2018, doi: 10.14569/IJACSA.2018.090647.

[30] G. H. Lee, A. Z. Talib, W. M. N. W. Zainon, and C. K. Lim, "Learning history using role-playing game (RPG) on mobile platform," Lect. Notes Electr. Eng., vol. 279 LNEE, no. January 2014, pp. 729–734, 2014, doi: 10.1007/978-3-642-41674-3_104.

[31] F.-L. Fu, R.-C. Su, and S.-C. Yu, "EGameFlow: A scale to measure learners' enjoyment of e-learning games," Comput. Educ., vol. 52, no. 1, pp. 101–112, Jan. 2009, doi: 10.1016/j.compedu.2008.07.004.

[32] B. Rolfe, C. M. Jones, and H. M. Wallace, "Designing Dramatic Play: Story and Game Structure," in Proceedings of the 2010 British Computer Society Conference on Human-Computer Interaction, BCS-HCI 2010, Sep. 2010, no. September, pp. 448–452, doi: 10.14236/ewic/HCI2010.54.

[33] C. Moser and X. Fang, "Narrative Structure and Player Experience in Role-Playing Games," Int. J. Hum. Comput. Interact., vol. 31, no. 2, pp. 146–156, 2015, doi: 10.1080/10447318.2014.986639.

Smart Irrigation and Precision Farming of Paddy Field using Unmanned Ground Vehicle and Internet of Things System

Srinivas A[1]
Department of Mechanical Engineering
PES University
Bengaluru, India

J Sangeetha[2]
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology
Bengaluru, India

***Abstract*—Paddy is one of the largest consumed staple foods across the globe, especially in Asian countries. With the population growing larger and the agricultural land shrinking, there is a need to increase the yield of the crop to meet the ever-growing food demand. The yield of paddy largely depends on the irrigation of the paddy field, that is maintaining the optimum water level in the paddy field. The solution to this irrigation problem has been proposed in this paper, by addressing various challenges in implementing the Unmanned Ground Vehicle (UGV) and Internet of Things (IoT) system in paddy cultivation. A UGV which carries the sensors were used to collect the sensor data (water level, rainwater, humidity, temperature, light intensity) from the paddy field, which is controlled by cloud-based solution and by the mobile application-based solution. The data was then processed and used to control the water valves which can again be controlled by using cloud and mobile application. Water level maintained by using the mobile application-based solution, cloud-based solution and by following the traditional method of irrigation was compared and the cloud-based solution was found to be more efficient. Thereby providing a solution which reduces the manpower required for the process of irrigation when compared to the traditional irrigation method, also reducing the water wastage, therefore conserving water.**

Keywords*—Sensor; cloud; mobile application; agriculture; water valve

I. Introduction

The Internet of Things (IoT) has entered almost all the areas in today's world. The concept of one thing being connected to other things. The ability to work with coordination, by communicating with one another over the Internet, has found a lot of application in fields like industry [1], smart city [2], connected building and campuses [3], smart home [4], health care [5], logistics [6], connected car [7]. Many research works are being done in various fields and domains [8], one of the fields where extensive research is being done is in agriculture [9].

The ever-growing population of the world, the diminishing natural resources, the depleting water resources has made food security a global concern. With only 11% of the earth surface [10] available for feeding 7.7 billion people across the globe [11], there is a need to implement new technologies in the field of agriculture. The leading food crops in terms of global consumption are paddy, wheat, and corn. Paddy is the most consumed, about an average of 50% of the daily caloric supply in Asian countries such as India, China, Indonesia, Japan, Philippines, Thailand etc. [12]. With urbanization, there is a competition for water sharing between cities, industries, and agriculture. It is estimated by [13] that the production of paddy must be increased by 50 to 60 per cent without increasing the land and water usage for paddy for the next 30 years. With many of the country economy depending mainly on paddy production, there is a need to implement innovations in the field of agriculture to cope up with all these challenges. The business data platform Statistica [14] estimates that about 3.8 billion people across the globe will have access to mobile phones and Internet by 2021, the use of Internet of thing can be the game-changer to boost the paddy production for all the developing nations.

Robots are machines that perform certain tasks and are usually controlled by a computer program. There are various kinds of robots such as autonomous robots (Robots which have central processing unit and can perform set of tasks on their own) and remote-controlled robots (Robots which are instructed by humans using a remote to perform tasks). Robots can also be classified into Unmanned Ariel Vehicle (UAV) and Unmanned Ground Vehicle (UGV) based on whether the unmanned robot moves in the air and moves on land respectively. Robots have entered agriculture to help farmers in various repetitive tasks or tasks that are difficult to be performed by humans. They are already being tested on applications such as spraying fungicides [15], pesticides [16], they are also used in mushroom picking [17]. There has been widespread research across the globe on robotics, with rapid research and development of robots, the markets will exhibit rapid growth during the coming decades [18].

II. Literature Review

The major factor influencing the yield of any crop is irrigation. In the paddy field, to increase the yield it is required to maintain standing water throughout the growth period, so the main aspect of irrigation system is maintaining the water level in the paddy field to the optimal level at all times [19]. Lots of IoT solutions has been developed by researchers across the globe for smart irrigation systems. One of them being the Wireless Sensor Networks (WSN) used in [20] where several sensor nodes are placed in various locations of the agricultural

land and the soil moisture, temperature, humidity, and various other parameters are being measured. The Same concept of WSN was used in [21] to obtain temperature, humidity, phosphorous, nitrogen, calcium, soil temperature, moisture, and PH data in the sugarcane field. In [22] two different types of wireless sensor network were used to obtain real-time data for irrigation in nurseries, the wireless sensor network can be used in all the above-mentioned field. It can also be used in paddy cultivation as in [23] where wireless sensor network was utilized to monitor and irrigate the paddy field. This may lead to additional work thereby increasing the cost of production as the wireless sensors have to be removed from its place each time when heavy machinery is used for various activities in the paddy field.

The novelty of this paper is in developing a reliable solution for paddy cultivation using IoT system by investigating the problems in placing the sensor to monitor the paddy field. Due to harsh environmental conditions in the paddy field, like very loose soil, the constant presence of water etc. It is very difficult to employ the method of using multiple sensor nodes to monitor the soil state. Placing multiple sensor nodes will be very difficult for the sensor to be maintained properly in the harsh condition of the constantly wet paddy field. It would also cause a lot of difficulties while harvesting the paddy field because the sensors must be removed whenever the crop is harvested by big machinery [24]. This paper also aims to develop the solution with the help of UGV so that it reduces the manpower required for irrigation thereby overcoming the problem of shortage of workforce in the agricultural sector.

The paper is divided into sections, where Section 3 deals with the methodology adopted in solving the problem. Later, the result is discussed in Section 4, followed by conclusions and future work in Section 5.

III. Methodology

Maintaining an optimum level of water and irrigating the paddy field with the right amount is very important to increase the yield of paddy. To perform this task a smart irrigation system for paddy field is important. The main challenges in developing a smart irrigation system for the paddy field are as follows.

- Collection of sensor data.
- Implementing the irrigation solution in the paddy field.

A. Collection of Sensor Data

Collection of data such as real-time field water level, rain status, temperature, humidity in the paddy field has many challenges because of the harsh condition in the paddy field. The method of placing a network of sensors as done in nurseries [25] will not work in particular to the paddy field, as the paddy field is exposed to heavy machinery for ploughing, transplanting, harvesting etc. Placing a network of a wireless sensor for measuring water level, rainwater, humidity, temperature, the light intensity is a big challenge. Instead of reducing the labour cost, this will lead to more cost of labour because the sensor nodes must be removed and placed each time when heavy machinery is used on the field. To solve this problem, the method adopted here is to divide the paddy field into sections and use a UGV which carries sensors on it, to move around in the paddy field collecting data. This not only solves the problem of placing sensor but also reduces the number of sensors needed, thereby reducing the cost, and making this method more feasible and reliable. The description of the sensors used is listed in Table I.

TABLE I. Sensors used in IoT Paddy Irrigation System

SN	*Name*	*Description*
1	DHT22	This sensor uses thermistor and capacitive humidity sensor to measure the temperature and humidity of the surrounding air respectively.
2	LDR	It is a light-sensitive device whose resistivity is a function of the incident electromagnetic radiation.
3	Water level sensor	This sensor gives an analogue signal that depends on the conductance which varies with the water level. Its outputs value varies between 0 to 1024.
4	Raindrop sensor	In this sensor, as the raindrops fall on the circuit board, it creates a path of parallel resistance which is measured using op-amp.

B. Implementing the Irrigation Solution in the Paddy Field

The method implemented here is developed to reduce the required manpower in paddy cultivation and thereby reducing the cost and thus helping the farmer. Paddy irrigation is labor-intensive, as it requires the farmer to constantly check the water level in different sections of the paddy field and maintain the optimal water level in each section, by operating the valves in each section. These works can be achieved by adopting a mobile application-based method and by using an independent cloud-based method.

1) Mobile application-based method: In this method, a mobile application acts as one place switch to monitor the paddy field water level. The mobile application is used to regulate the water level in the field by remotely operating the valves in different sections. Thereby, greatly helping the farmer by reducing the time for operating the valves and directing the water flow in the paddy field. A flow chart representing the mobile application- based method is shown in Fig. 1. In this, the farmer first inputs the login credentials in the mobile application. After logging in, the farmer will be given the option of controlling either the UGV or the water valve. In the UGV control screen, the farmer can control the UGV to move left, right, forward, backward. The UGV moves as instructed and sends the sensor data to the mobile application, where the farmer can see the water level at that location. In the water valve control screen, the farmer can switch the valve on or off.

2) Independent cloud-based method: A flow chart representing the independent cloud-based method is shown in Fig. 2. In this method, the cloud acts as the central system coordinating with the local server and pushing notifications to the mobile application for farmers. The UGV carrying the sensors will be controlled by a local server, the local server then collects the data from the UGV, processes the data and stores it. The local server then updates the variables in the

cloud database, based on this the cloud platform will switch the valve on or off. Thereby, maintaining an optimal level of water at all times. At the end of each day, the local server also pushes the day's sensor data to the cloud storage.

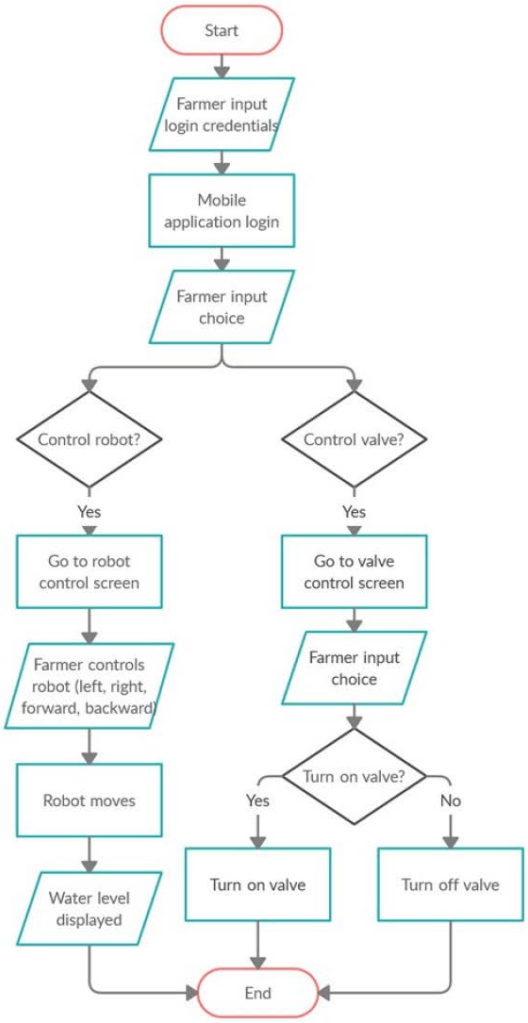


Fig. 1. Flowchart of the Mobile Application-based Method.

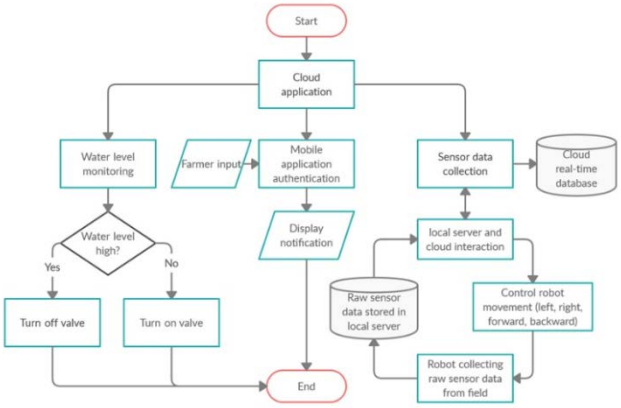


Fig. 2. Flowchart of the Independent Cloud-based Method.

IV. Result and Discussion

To develop an IoT solution for the problems faced by the farmers in paddy cultivation. It has been identified that the key to increasing the paddy crop yield, is to closely monitor the water level in the field and irrigate the land with the correct amount of water at various stages of the paddy growth. As a result, the authors have primarily focused on five main components to achieve this goal and they are listed below:

- Water level monitoring system.
- Water Valve system.
- Unmanned Ground Vehicle.
- Mobile application-based solution.
- Independent cloud-based solution.

On the road to developing a solution, some more hardware components that were used are listed in Table II. The description of each component of the solution that was developed is discussed in the following sub-sections.

TABLE II. Some more Components used in IoT Paddy Irrigation System

SN	*Name*	*Description*
1	Arduino UNO	It is a microcontroller having 6 analogue input pins and 14 digital input/output pins.
2	Raspberry Pi 3 Model B+	It is a single-board computer with wireless, LAN and Bluetooth connectivity.
3	Node Microcontroller Unit (Node MCU)	It is a microcontroller having low-cost open-source firmware and board based on the eLua project.
4	ESP8266	It is a low-cost Wi-Fi microchip with microcontroller capability and a full TCP/IP stack.
5	L293d IC motor driver	It is a Motor Driver module which can control the speed and direction of two motors simultaneously.
6	Relay Module	It can be controlled with low voltage. It is an electrically operated switch which allows the current to pass through when turned on.
7	Solenoid valve	It is an electromechanically controlled valve which consists of a solenoid and a movable ferromagnetic core in its centre.

A. Water Level Monitoring System

Water is a very important natural resource; water is the lifeline for any agricultural activity. The scarcity of water resources in many countries has forced everyone to look for means of utilizing the water resource with utmost care [26]. To utilize the water resources fruitfully, it is necessary to provide the right amount of water for agricultural activity, especially for the cultivation of water demanding crops like paddy, which calls for the need of good water management system for paddy cultivation. The best way to have good water management system is to have a good water level monitoring system.

In paddy cultivation, water plays a vital role in proper yield, as water can both be constructive (i.e., more yield) and destructive (i.e., poor yield). Too less water may lead to dying of crops and cracking of land. Very high water may lead to a poor yield of paddy, it may also lead to the soil nutrients being washed out and causes soil salinity. Therefore, water level

monitoring is crucial for paddy cultivation. The paddy cultivation always requires standing water. This standing water needs to be at different levels in different stages of the paddy life cycle which is obtained from [27]. Therefore, it is very important to check the paddy field water level at various stages of its life cycle.

To monitor the paddy field, the water level sensor is used. The water level sensor is connected to the Arduino UNO via an analogue input pin. The sensor measures the water level and gives output in the form of analogue signals which varies from 0 to 1024, which is mapped to 0 to 5 centimeter of water level in the Arduino using proportionality relation (1). This measurement is then taken consecutively for 20 times in one location and then averaged to get the most accurate value of water level at that location. In this way, the water can be monitored most accurately. Fig. 3 gives the pictorial representation of the water level monitoring system.

$$water\ level(cm) = \frac{analog\ pin\ input * 5}{1023} \quad (1)$$

B. Water Valve System

Valves are very important for the controlled flow of water. In the agriculture field, valves are vital for controlling, stopping, and directing the water flow for irrigation. Valves play the role of end actuator by working as the output device for the input from a water level monitoring system. The water level monitoring system acts as the sensing system in the irrigation of paddy fields.

As paddy fields are very large, it is usually divided into many small sections by constructing bunds. Water is then directed to each of the sections by opening a small portion of the bunds when the required level of water is reached. The portion of the bunds is closed, and other sections are irrigated in the same way. This involves a lot of manpower and wastage of precious time of the farmers. This can otherwise be directed to other important agricultural activities such as weeding, fertilizing process etc., it also leads to excess or low water being irrigated which may lead to a lower yield of paddy. Hence this loss of time and money the farmer and the poor yield can be prevented by using a valve system that is remotely operated either by mobile application or by cloud services.

To develop this valve system a low-cost solenoid valve is used. This valve is connected to the Relay module which is controlled by Node MCU as shown in Fig. 4. Node MCU receives a signal from either the mobile application or cloud platform to switch on or off the valve, and then the MCU switches the relay module which then leads to the opening or closing the valve. In this way, the exact level of water given by [27] in the paddy field is maintained.

Fig. 3. Pictorial Representation of the Water Level Monitoring System for Paddy Cultivation.

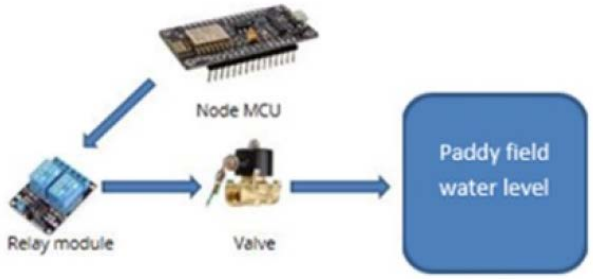


Fig. 4. Pictorial Representation of the Valve System for Paddy Cultivation.

C. Unmanned Ground Vehicle

Paddy field has harsh conditions that are not suitable for sensors, hence the sensors must be protected, it also poses the challenge of placing the sensors in the right locations, to get an accurate reading. Placing the sensors on the ground at a different location to monitor water level, temperature, humidity etc. will not be suitable for paddy field because of various difficulties involved. Hence, there is a need for a robot to carry sensors. As our primary objective is to measure the water level, aerial robots will not serve the purpose, as it is not possible to measure the water level on the paddy field from air. Hence a UGV was developed for this purpose.

The UGV was designed to face the challenges in the paddy field. The design procedure consisted of first identifying the difficulties the UGV would face such as muddy, slippery soil, water standing up to a level. Then various sensors that the UGV must carry was considered, motor's speed to drive the UGV, the battery size, the weight of all the components were considered. The chassis of the UGV was designed considering all these factors. The material was assigned as aluminum because of its lightweight property, as the weight was a major factor in design. As the weight increases, the motor size increases, and the battery size should have to be increased. The wheel of the UGV was decided to be a tracked wheel since it provides a large surface area. Computer-Aided Drawing (CAD) software 'Solid Edge' was used to visualize each part of the chassis. Later all the parts were assembled in the CAD software, shown in Fig. 5. Then based on the CAD software visualization, a model was built.

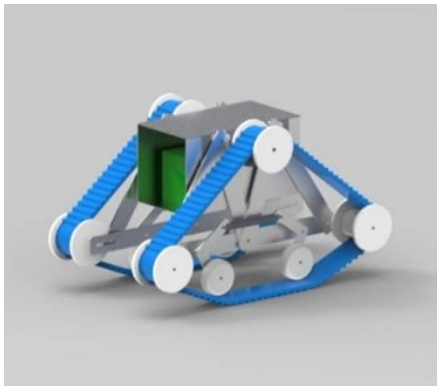


Fig. 5. Picture of UGV Assembled in CAD Software Solid Edge.

The UGV is driven by two DC motor, with the help of L293d IC motor driver which is controlled by Arduino UNO. The UGV carries sensors DHT22 (Humidity and temperature sensor), water level sensor, raindrop sensor, LDR. The pictorial representation of all the sensors carried by the UGV is shown in Fig. 6 and the model of the UGV after fabrication is shown in Fig. 7.

D. Mobile Application-based Solution

The smartphone industry is a growing market in the world, with more and more smartphones penetration into developing country like India, it is expected to reach a user base of 442 million by 2022 [28]. Hence, it has been identified that mobile application-based solution to be a feasible solution for the problem of irrigation [29].

There are lots of work involved in irrigating a paddy field such as, constantly checking the water level in different sections of the paddy field, making sure that the water is maintained at the correct level in each section, opening and closing the valves and directing the water to all the sections. This may not be a big problem for a very small field, but for a very large field, it is very time consuming and may even require a separate full-time worker just for performing these tasks. Employing a separate worker just to perform these tasks incurs additional expenditure to the farmer which can otherwise be prevented by employing technology. Hence the use of a smartphone with a mobile application to perform these tasks will help reduce the unwanted expenditure in employing workers.

The mobile application is a web-based application which is hosted by Google's Firebase. It has three functionalities they are,

- Operate the UGV that carries the sensor.
- Switching on or off the valves.
- Receive any notifications from the cloud platform.

The mobile application has a login screen which requires an email id and password to log in. The credentials are authenticated by the Google firebase's authentication, where it assigns Unique Identification (UID) for each user hence the details of the farmers are stored securely. It has a screen for receiving the notification regarding updates on the field and other related news such as government schemes, which is shown in Fig. 8, firebase has the option of pushing a user-specific notification by using UID that is assigned to each user.

The mobile application has a screen for operating the UGV, where there is information regarding the water level at the current location of UGV and it has buttons for maneuvering the UGV left, right and to move forward, backward and stop. It also has a button for switching the mode of the UGV control to the mobile application-based method or independent cloud-based method. It has another screen which has buttons for switching the valves on and off, this is shown in Fig. 9.

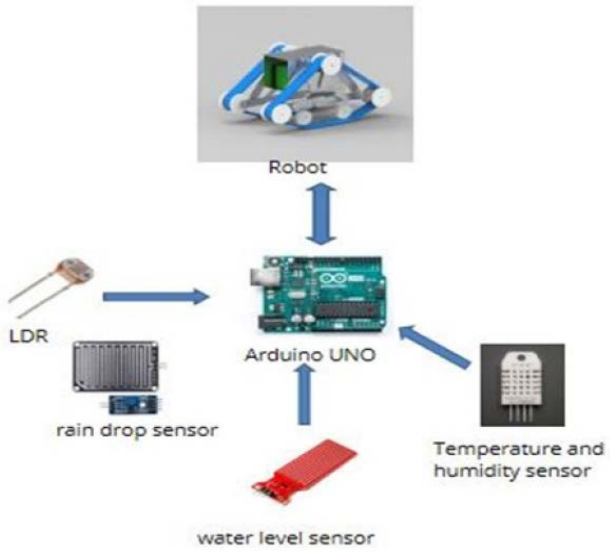


Fig. 6. Pictorial Representation of all the Sensors Carried by the UGV.

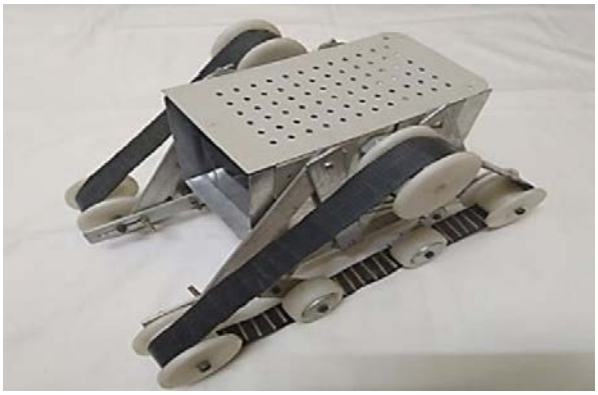


Fig. 7. The Model of the UGV after Fabrication.

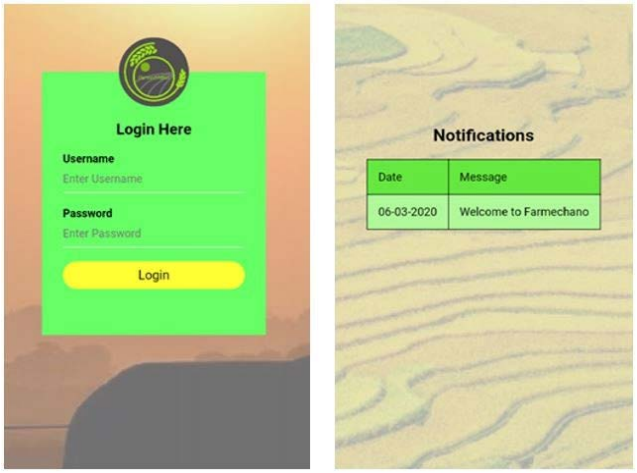


Fig. 8. Snapshot of Login Screen towards Left and Notification Screen towards the Right of Mobile Application.

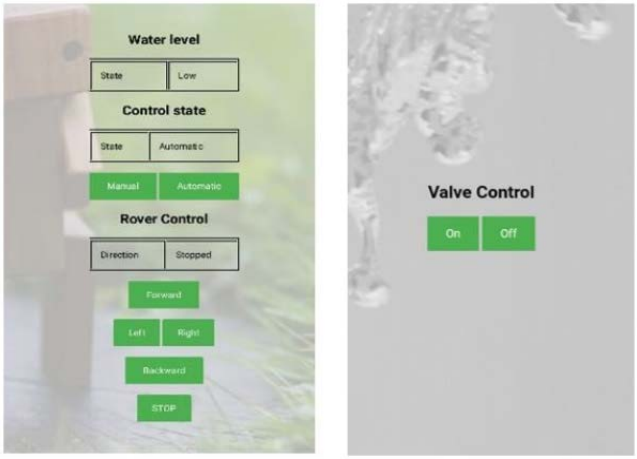


Fig. 9. Snapshot of the Screen for UGV Control in the Left and Screen of Valve Control in the Right in the Mobile Application.

A pictorial representation of a mobile application-based control system is shown in Fig. 10. In this when the buttons in the mobile application are pressed, the corresponding variable of the UGV in Google's Firebase real-time database is updated. These variables, which is constantly being monitored by the local server (Raspberry Pi) as well as the Node MCU of the valve control system receives the command. The local server then sends this command to the UGV using Wi-Fi as the channel for communication. The UGV receives this command with the help of ESP8266 connected to the Arduino Uno. Based on the command, the Arduino controls the driving motor, on the other hand, the Node MCU of the valve control system controls the valve based on the command.

E. Independent Cloud-based Solution

Paddy cultivation is labour intensive, a decrease of workers available for agricultural activity has become a major problem in recent years. As more and more agricultural labourers are migrating to cities in search of better livelihood, financial and social status [30]. This has risen an alarm in the agricultural sector, as farmers are not getting sufficient labourers to carry out agricultural activities. The poor income of farmers and the rising cost of labourers for agricultural activities, has also become a concern. Hence this problem can be solved by bringing in automation in the cultivation of paddy field and integrating with the cloud to develop an end-to-end IoT solution.

The main objective of an independent cloud-based system is listed below:

- To monitor the water level of the paddy field
- To control the UGV for collecting paddy field data
- To control the water level of the paddy field
- To store the collected sensor data

The pictorial representation of the independent cloud-based control system is shown in Fig. 11. In this Raspberry Pi is employed, which is the local server placed in the storeroom next to the paddy field which is considered as the control room. The local server further divides the sections of the paddy fields created by constructing bunds into small pieces virtually, it then instructs the UGV which carries the water level monitoring system on it, to move towards each of these sections and carry out the reading, the water level monitoring system along with various other sensors attached to the UGV (as discussed in section IV C) sends the readings to the local server through Wi-Fi. The local server then processes these raw sensor data and stores the values in the local variables such as water level, temperature, humidity, light intensity, rain. The local server then compares the values for water level with the predefined values which are recommended by the Tamil Nadu Agricultural University (TNAU) where the optimal level of water for different stages of the paddy crop growth can be obtained for different ecosystems [27]. The value of variable water level (wat lev) in each section of the paddy field, temperature (temp), humidity (hum), rain, light intensity (LDR) of the field as processed and displayed by the local server is shown in Fig. 12.

The local server then logs all the processed sensor data in a local file of .CSV format which has columns of time, water level, LDR, rain, humidity, temperature, for every one-hour. This .CSV file is then uploaded to Google firebase's Fire-store at the end of each day, which can later be used for checking the health of various sensors.

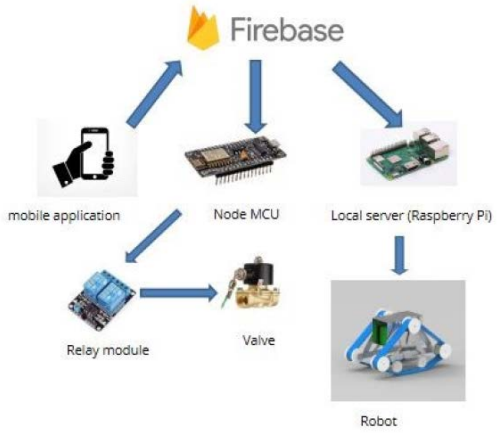


Fig. 10. Pictorial Representation of the Mobile Application-based Control System.

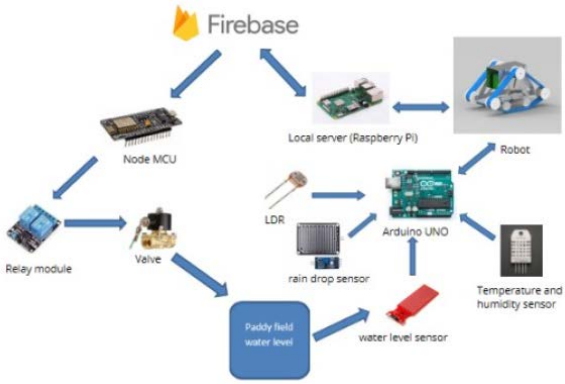


Fig. 11. Pictorial Representation of the Independent Cloud-based Control System.

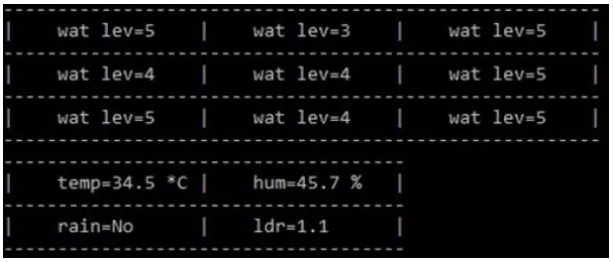


Fig. 12. Snapshot of the Local Variable Values of Water Level for each Section of the Paddy Field along with Temperature, Humidity, Rain Status, Light Intensity (LDR) as Processed and Displayed by the Local Server.

The water level and rain status are updated on the respective variable in Google's firebase real-time database. Based on the water level and rain status the firebase updates the variable 'valve' in the firebase Realtime database. This variable 'valve' being monitored by the valve system controls the valves for various sections of the paddy field. Thus, the independent cloud-based solution can monitor and control the water level of the paddy field.

As mentioned earlier, to increase the paddy crop yield monitoring the water level and irrigating the paddy field with the right amount of water at a different stage of the paddy growth is important. First, the water level that was maintained in a paddy field by following the traditional method of irrigation, where farmers direct the flow of water to different sections manually and maintain water level by eye judgment was recorded, for each hour of the day from 06:30 Hrs. to 18:30 Hrs. Then water level that was maintained in the paddy field by following a mobile application-based solution was recorded for each hour of the day from 06:30 Hrs. to 18:30 Hrs. Then water level that was maintained in the paddy field by following independent cloud-based solution was recorded for each hour of the day from 06:30 Hrs. to 18:30 Hrs. Then the water level values that were collected by using all three methods were plotted against time of the day as shown in Fig. 13. The optimum water level is represented by a constant line in the graph.

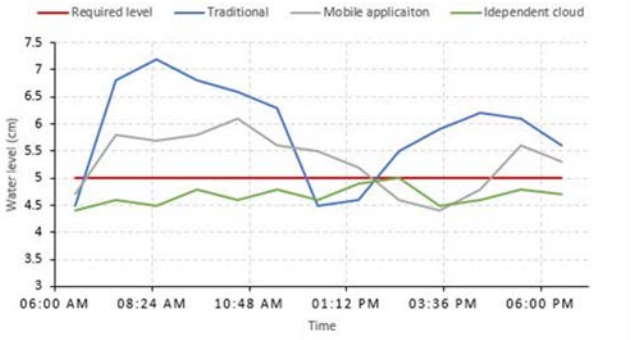


Fig. 13. A Plot of Water Level in cm vs Time for various Methods followed for Paddy Cultivation.

From the graph, the independent cloud-based solution could maintain the water level in the paddy field at a near-constant required level when compared to the mobile application-based solution and traditional method. It was observed that in the traditional method, farmers tend to use more water than recommended since they have no means to measure the exact level of water and just go by eye perspective, whereas when using the mobile application-based solution, there is means to know the water level, thereby the water level is fairly close to the recommended level but there are large deviations a time, as it is still operated by a human.

V. Conclusion and Future Work

A novel solution for irrigation of paddy field was presented with Unmanned Ground Vehicle (UGV) and Internet of Things (IoT) by investigating various problems arising in placing sensors in the paddy field. This solution enables maintaining the water level in the paddy field in the optimal level which is an important parameter for a good yield of paddy crop, throughout the growth cycle in the most cost-effective manner. The solution presented in this paper is feasible to integrate into the present paddy field with ease. It also helps in saving the time and manpower required for paddy cultivation. The solution presented here can be used in places where water resources are scarce and stringent monitoring of water is required. It minimizes the wastage of water and thereby conserving the precious water resources. The UGV can further be modified to carry and spray pesticides, fertilizers etc. making it a multirole UGV for use in paddy field. The path planning for the UGV can be made more dynamic by implementing machine learning and artificial intelligence. The solution developed here can be modified and tested for use in other type of crops.

References

[1] Aazam, Mohammad, Sherali Zeadally, and Khaled A. Harras. "Deploying fog computing in industrial internet of things and industry 4.0." IEEE Transactions on Industrial Informatics 14, no. 10 (2018): 4674-4682.

[2] Sanchez, Luis, Luis Muñoz, Jose Antonio Galache, Pablo Sotres, Juan R. Santana, Veronica Gutierrez, Rajiv Ramdhany et al. "SmartSantander: IoT experimentation over a smart city testbed." Computer Networks 61 (2014): 217-238.

[3] Sastra, Nyoman Putra, and Dewa Made Wiharta. "Environmental monitoring as an IoT application in building smart campus of Universitas Udayana." In 2016 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS), pp. 85-88. IEEE, 2016.

[4] Park, Eunil, Yongwoo Cho, Jinyoung Han, and Sang Jib Kwon. "Comprehensive approaches to user acceptance of Internet of Things in a smart home environment." IEEE Internet of Things Journal 4, no. 6 (2017): 2342-2350.

[5] Islam, SM Riazul, Daehan Kwak, MD Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. "The internet of things for health care: a comprehensive survey." IEEE access 3 (2015): 678-708.

[6] Li, Lin. "Application of the internet of thing in green agricultural products supply chain management." In 2011 Fourth International Conference on Intelligent Computation Technology and Automation, vol. 1, pp. 1022-1025. IEEE, 2011.

[7] Husni, Emir, Galuh Boy Hertantyo, Daniel Wahyu Wicaksono, Faisal Candrasyah Hasibuan, Andri Ulus Rahayu, and Muhamad Agus Triawan. "Applied Internet of Things (IoT): car monitoring system using IBM BlueMix." In 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 417-422. IEEE, 2016.

[8] Stankovic, John A. "Research directions for the internet of things." IEEE internet of things journal 1, no. 1 (2014): 3-9.

[9] Tzounis, Antonis, Nikolaos Katsoulas, Thomas Bartzanas, and Constantinos Kittas. "Internet of Things in agriculture, recent advances and future challenges." Biosystems engineering 164 (2017): 31-48.

[10] Crop production and natural resource use: http://www.fao.org/3/y4252e/y4252e06.htm.

[11] Population: https://www.un.org/en/global-issues/population

[12] Maclean, Jay L., David Charles Dawe, and Gene P. Hettel, eds. Rice almanac: Source book for the most important economic activity on earth. Int. Rice Res. Inst., 2002.

[13] Balasubramanian, V., M. Bell, and M. Sombilla. "Yield, profit, and knowledge gaps in rice farming: Causes and development of mitigation measures." BRIDGING THE RICE YIELD GAP IN THE ASIA-PACIFIC REGION (2000): 163.

[14] Number of smartphone users worldwide from 2016 to 2021 (in billions): https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[15] Singh, S., T. F. Burks, and W. S. Lee. "Autonomous robotic vehicle development for greenhouse spraying." Transactions of the ASAE 48, no. 6 (2005): 2355-2361.

[16] Sammons, Philip J., Tomonari Furukawa, and Andrew Bulgin. "Autonomous pesticide spraying robot for use in a greenhouse." In Australian Conference on Robotics and Automation, vol. 1, no. 9. 2005.

[17] Yaghoubi, Sajjad, Negar Ali Akbarzadeh, Shadi Sadeghi Bazargani, Sama Sadeghi Bazargani, Marjan Bamizan, and Maryam Irani Asl. "Autonomous robots for agricultural tasks and farm assignment and future trends in agro robots." International Journal of Mechanical and Mechatronics Engineering 13, no. 3 (2013): 1-6.

[18] Bogue, Robert. "Robots poised to revolutionise agriculture." Industrial Robot: An International Journal (2016).

[19] Brown, K. W., F. T. Turner, J. C. Thomas, L. E. Deuel, and M. E. Keener. "Water balance of flooded rice paddies." Agricultural Water Management 1, no. 3 (1977): 277-291.

[20] Kaewmard, Nattapol, and Saiyan Saiyod. "Sensor data collection and irrigation control on vegetable crop using smart phone and wireless sensor networks for smart farm." In 2014 IEEE Conference on Wireless Sensors (ICWiSE), pp. 106-112. IEEE, 2014.

[21] Abhijith, H. V., Darpan A. Jain, and U. Adithya Athreya Rao. "Intelligent agriculture mechanism using internet of things." In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2185-2188. IEEE, 2017.

[22] Lea-Cox, John D., George F. Kantor, and Andrew G. Ristvey. "Using wireless sensor technology to schedule irrigations and minimize water use in nursery and greenhouse production systems." In Comb. Proc. Int. Pl. Prop. Soc, vol. 58, pp. 512-518. 2008.

[23] Xiao, Kehui, Deqin Xiao, and Xiwen Luo. "Smart water-saving irrigation system in precision agriculture based on wireless sensor network." Transactions of the Chinese society of Agricultural Engineering 26, no. 11 (2010): 170-175.

[24] Muazu, A., A. Yahya, W. I. W. Ishak, and S. Khairunniza Bejo. "Machinery utilization and production cost of paddy cultivation under wetland direct seeding conditions in Malaysia." Engineering in agriculture, environment and food 8, no. 4 (2015): 289-297.

[25] Coates, Robert W., Michael J. Delwiche, Alan Broad, and Mark Holler. "Wireless sensor network with irrigation valve control." Computers and electronics in agriculture 96 (2013): 13-22.

[26] García, Laura, Lorena Parra, Jose M. Jimenez, Jaime Lloret, and Pascal Lorenz. "IoT-based smart irrigation systems: An overview on the recent trends on sensors and IoT systems for irrigation in precision agriculture." Sensors 20, no. 4 (2020): 1042.

[27] Expert system for paddy: http://www.agritech.tnau.ac.in/expert_system/paddy/cultivationpractices3.html

[28] Number of smartphone users in India: https://www.statista.com/statistics/467163/forecast-of-smartphone-users-in-india/

[29] Saad, Abdelmadjid, and Abdoulaye Gamatié. "Water management in agriculture: A survey on current challenges and technological solutions." IEEE Access 8 (2020): 38082-38097.

[30] Doddamani, K. N. "A study on migration of agriculture labourers from hyderabad Karnataka area to Maharashtra." IOSR-JHSS 19 (2014): 68-71.

Adaptive Trajectory Control Design for Bilateral Robotic Arm with Enforced Sensorless and Acceleration based Force Control Technique

Nuratiqa Natrah Mansor, Muhammad Herman Jamaluddin, Ahmad Zaki Shukor
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia

***Abstract*—This study offers an approach for tackling the issue of instability on the computed force generated on a joint of a robotic arm by improving the model of a bilateral master-slave haptic system with an adaptive technique known as Reaction Force Observer (RFOB). The purpose of recommended modelling is to correct unsought signals coming from the employed standard controller and the surroundings produced within the moving joint of the articulated robotic arm. RFOB is employed to adjust the signal interference by modifying its position response to obtain the desired final location. The investigation and observation were carried out in two separate tests to evaluate the outcomes of the recommended integration technique with the former system that only enforced Disturbance Observer (DOB). Generated feedbacks produced from the organised experiments are measured inside a simulation platform. All numerical records and signal charts illustrate the durability of the proposed method since the system integrated with acceleration-based force control is more precise and quicker.**

***Keywords*—Force and position controller; reaction force observer; bilateral control robotic arm; sensorless; system response**

I. Introduction

A bidirectional master-slave industrial robotic arm manipulator system is a revolutionary technology that was beyond imagination a century ago. It permits explorers and adventurers to reach the places unavailable to them. The places might be inaccessible, harmful or isolated. By the time this research is studied, there have been many applications in multiple fields, including surgical operation, exploration in the deep ocean and outer space operations, and coping with volatile or high emission activities [1]. Robots are programmable machines. It can cope in a different environment with unique features, mobile and easy to manoeuvre. Therefore, this study utilised a youBot made by German robots' manufacturer, KUKA, as the device to showcase and operate as a bidirectional haptic system.

According to prior studies conducted by researchers all around the world, old traditional approaches had technical limitations. The techniques focus on enhancing the control system itself through the use of premade and essential equipment such as a keyboard, joystick, data glove, basic manipulator connection, and the use of force sensors [2]. Common force sensors appear to have several restrictions and disadvantages for the system. The system has particular uncertainty, instability, and delays [2-4]. On an actual industrial robot arm, not much of the previous study uses contemporary control techniques to enhance the system feedback. It is impracticable and wasteful not to use these strategies, which have been shown to improve responses on control systems in multiple previous studies [3,5]. The combination of force control and position control into the design of bilateral robotic arms should be emphasised to discover the disparity, uniqueness, and uncommon industrial task handled by the arm manipulator with other ordinary and smaller haptic devices [6,9].

Therefore, it is a refreshing attempt to build and model a bilateral control system on industrial robots. In the past 40 years, haptic technology has evolved across engineering studies and many other research areas, including arts and design. Several researchers have studied its control system, auxiliary, communication and wearable devices, as there is a diversity of many possibilities, opening doors for incoming haptic technology [6-7]. For instance, [7,8] discovered that control action and response inside the said system might increase as far as 90% accuracy compared to the conventional approach that did not utilise any adaptive control technique. The standard system without the adjusted controller parameters needs to battle around 25% to attain the controller goal [4]. Following the deployment of DOB and RFOB, it acquired efficacy and simultaneously avoided any infallibility on the control process. With the facts, the integration of both adaptive techniques to the new design of the control system should be commissioned and emphasised. The benefits of implementing these stated observers into a system are projected to boost the feedback of bilateral haptic inside the system. It is also to eliminate all unwanted signals and disturbances that occur while operating.

The main objective of this study is to be set and incorporate a type of sensorless force control into the robotic arm simulation and observe the most acceptable parameters to arrange for the robot to operate and work optimally in a bilateral way. The outcome reported in this article is focused on employing two versions of robust control tools to increase the efficacy and discover the capability of the system to adjust its operation to obtain the optimal potential mode of operation [6-8]. The second section will delve into the robust acceleration control and the block diagram of acceleration-based force control, review the series of steps of procedures and methodologies performed on each experiment. In the following section later, all recorded data and information observed from every experiment are tallied and illustrated in

Zamalah Scheme, Universiti Teknikal Malaysia Melaka (UTeM).

graph version. Next, Section III will discuss the feedback of the aforementioned master-slave manipulator control system recorded from the simulation. Meanwhile, Section IV summarises the essential findings and conveys the recommendations to acknowledge the limitations and improve future work.

II. Methodology

A. Control Design based on Disturbance Compensation

The fundamental purpose of this project is to employ a software simulation for constructing a model of a haptic robotic arm to work in bilateral master-slave interaction. As being discussed in the introduction part, system response in normal circumstances carries noise and suffers constraints during operation. The constraints can be in the bandwidth of the filter due to the wide frequency range or internal stability. Hypothetically, there are concerns and limitations to achieving the robustness as it is generally mediocre and tough to maintain the system. One of the solutions to handle the challenge is introducing a control tool into a control system [8]. This control structure can boost system infallibility by erasing uncertainties and undesirable information within the system. The role of this inner-loop output-feedback controller is to discard outer disturbances and make up the outer-loop baseline controller resilient against the plant's uncertainties [7].

Employing force sensors in a machine or equipment brings several significant drawbacks, although it is responsible for measuring the force acting on a specific object. The conventional sensors are not exceptionally durable, expensive, and restricted capability to detect the bandwidth. This observer can be enforced into a system loop to replace the traditional instruments for estimating force measurement and remain sensorless [5]. Moreover, the developed system is suitable to perform navigation and task manipulation in a connected teleoperation setup. The dynamic characteristics of DOB can improve the restrictions and inadequate ability encountered by standard basic controllers. Compared to the other two controllers, such as Proportional and Integral controllers (PI) and Proportional Integral Derivate Controller (PID), Proportional and Derivative Controller (PD) is well matched to pair with the observer to construct a new design of closed-loop control system for the master-slave robotic arm.

Incorporating DOB in the design system can measure disturbance force, F_{dis}, and give compensating current, $I_{cmp,}$ achieving robust motion control in unstable plants. The proposed tool is another notable technique for measuring force-producing and estimating disturbances. Consequently, filtered data by the observer paired with a fed-back input signal will be utilised to nominalise the inner loop, satisfy causality, and adjust motion control. Therefore, it will deliver high precision readings for precise position tracking. Whereas the equation is equivalently illustrated into block diagrams as shown in Fig. 1 and Fig. 2.

Systematically, the integrated technique supports the system to predefined performance criteria and achieves firm acceleration control. Fig. 2 summarises the schematic of the acceleration-based position control block diagram.

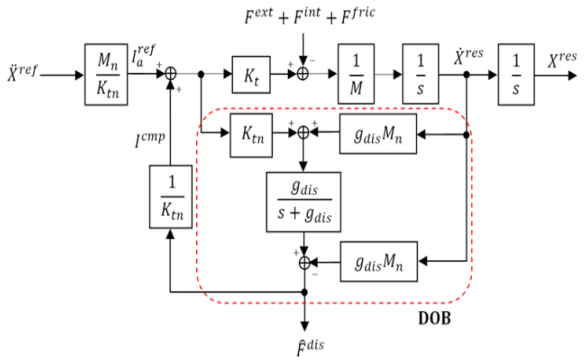


Fig. 1. A Block Diagram for a DOB based Open-loop Control System.

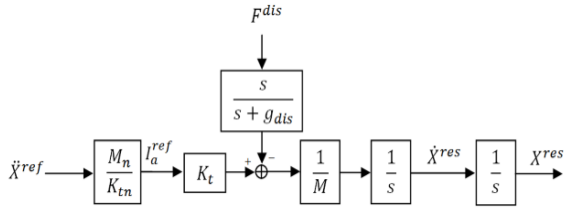


Fig. 2. Robust Acceleration Control.

The open-loop sensitivity and the co-sensitivity transfer function is derived in equation (1) below to denote uncertain and nominal plant models. The following transfer functions are the components of *Fdis*:

$$F_{dis} = F_{ext} + F_{int} + F_{fric} + (M - M_n)\, s^2 X_{res} + (K_{tn} - K_t)\, I_a^{ref} \quad (1)$$

The force-induced consist of modelling errors of nominal mass, M_n and nominal thrust coefficient, K_{tn} in overall is represented as disturbance force, F_{dis} equation. Meanwhile, Interactive force, F_{int} components are the Coriolis term, centrifugal term, and also gravity term. A low pass filter (LPF) and the inverse of a nominal plant model are required to shape the DOB model. The disturbance force F_{dis} is approximated as the following equation:

$$F_{dis} = \frac{g_{dis}}{s+g_{dis}} F_{dis} \quad (2)$$

The equation for position controller, C_p in differential mode is derived from summation of both position gain, K_p and velocity gain, K_v.

$$C_p = K_p + sK_v \quad (3)$$

Then, the corresponding position response, x_{res}, is expanded into:

$$x_{res} = \frac{C_p}{s^2}(x_{cmd} - x_{res}) \quad (4)$$

Where x_{res} in (4) has been rearranged and translated into (5), as seen below:

$$\frac{x_{res}}{x_{cmd}} = \frac{C_p}{s^2 + C_p}$$

$$\frac{x_{res}}{x_{cmd}} = \frac{sK_v + K_p}{s^2 + sK_v + K_p}$$

$$\frac{x_{res}}{x_{cmd}} = \frac{2\xi\omega_n s + {\omega_n}^2}{s^2 + 2\xi\omega_n s + {\omega_n}^2} \quad (5)$$

From the above equation, the damping ratio value, ξ, can be tuned to 1.0 to obtain a critical damping effect, whereas natural angular frequency, ω_n is a substitution of $\sqrt{K_p}$ or ½ K_v.

B. Control Design based on Reaction Force Estimation

Aside from the original purpose of DOB, it can be employed with RFOB to estimate the reaction force. Past studies have shown that the RFOB can estimate a wider band and excellent range of situations (Mansor et al., 2017). Consequently, force sensors are replaceable to be employed in bilateral manipulator's systems. The component of estimation also requires the identification of friction force F_{fric} and the interaction force F_{int}. Furthermore, the design of RFOB is competent to estimate the exterior force given out by the disturbance of the components in the type of acting force or force response [4-5]. To describe the process of this control loop technique and arrangement within the loop system, outlined block diagrams is illustrated in Fig. 3 and Fig. 4, respectively.

The cut-off frequency, g for RFOB, is similar to the DOB. Thus, the computed external force $\hat{F}_{ext}$ is equate as follows:

$$\hat{F}_{ext} = \frac{g_{dis}}{s+g_{dis}} F_{ext} \tag{6}$$

As the value for force controller, C_f is associated with the force gain, K_f, hence the estimated force response, F_{res} is described as follows:

$$\hat{F}_{res} = \frac{C_f Z_c g_{dis}}{s^2(s+g_{dis})}(F_{cmd} - \hat{F}_{res}) \tag{7}$$

Which can be transformed into (8):

$$\frac{\hat{F}_{res}}{F_{cmd}} = \frac{1}{\frac{s^2(s+g_{dis})}{C_f Z_c g_{dis}}+1} \tag{8}$$

C. Design of Bilateral Master-Slave System with Adaptive Control Technique

Fig. 5 displays the whole close-loop bilateral control system following the employment of both control loop techniques into the system. The observers' data will merge and feed back into the input signal for every passing process. The generated feedback will correct any internal modelling error or interruption and emerge into one input.

There are two modes in the bilateral control system, which are Differential Mode, $\ddot{x}_{dif}$ and Common Mode, $\ddot{x}_{com}$. The first mode is the product of position controller, C_p with the difference between the position of the master-slave system, and the latter is the product of force controller, C_f with the differences between the forces computed in master-slave. The equation for both modes follows the equation (9) and (10) below.

$$\ddot{x}_{dif}^{ref} = C_p(s)(x_s^{res} - x_m^{res}) \tag{9}$$

$$\ddot{x}_{com}^{ref} = C_f(s)(f_s^{ext} - f_m^{ext}) \tag{10}$$

D. Procedures for Work Simulation

This study's analysis and experiment are carried out through a simulation platform inside a robotic simulation software called Virtual Robotic Experimentation Platform (VREP). The software has a built-in KUKA youBot in its library and is ready to be integrated with various programmable tasks and coding languages. Furthermore, simulation scenes, models, and object characteristics are simple to manage as it has formed a plethora of choices and functionalities. The first task to be considered to prove whether the proposed system is ideal must follow the law of action and reaction in bilateral communication. After the system works according to the law, the experiment will be carried out to analyse the system with adaptive techniques. All starting values for each parameter and setting are presumed to be related to the real-time experiment. To construct the bilateral way of communication inside the KUKA youBot, the controllers, input and output arrangement are shown in Fig. 6.

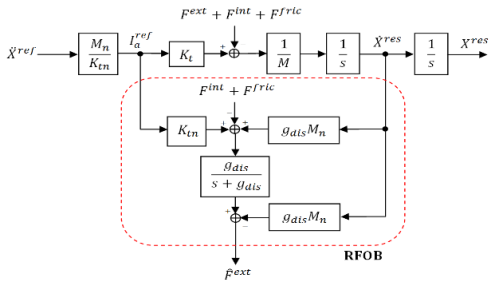


Fig. 3. A Block Diagram for RFOB based on an Open-Loop Control System.

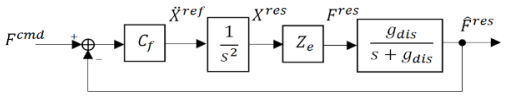


Fig. 4. Acceleration based Force Control System Block Diagram.

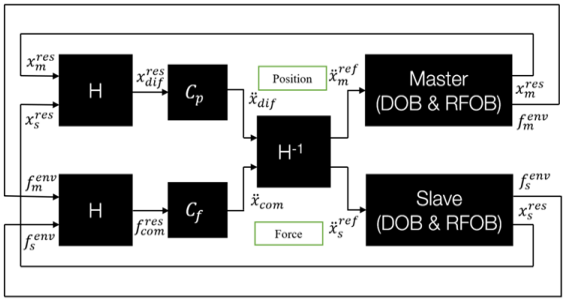


Fig. 5. A Simplified Block Diagram of Master-Slave Bilateral Control System with Proposed Tools.

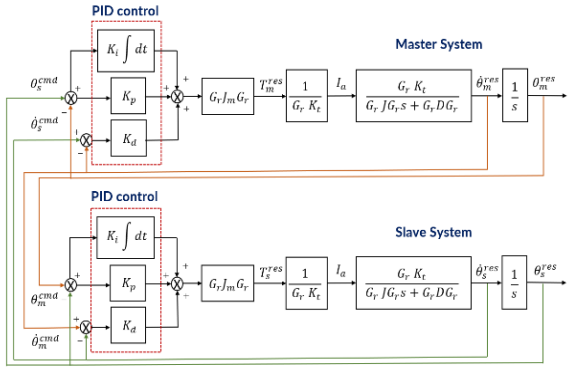


Fig. 6. General Block Diagram for a Bilateral Robotic Arm with Two Separate Subsystems.

The remote Application Programming Interface (API) function will be used to communicate between both robots with the programmed environment. To simplify the study work, only reading in a single joint are monitored even though the robot has 5 degrees of freedom (DOF) and multiple operatable joints, as indicated in Fig. 7 below. This joint on the waist part is labelled as 'Joint0' (located at the first joint on the lower robot component).

The reason to focus on a single link of the manipulator is to reduce the complexity of managing the trajectory control and operating successive joints and degrees of freedom during tests. The environment and setup for all experiments are modelled in VREP. To illustrate the simulation processes, the steps are outlined in Fig. 8 and shown in the following Fig. 9a and Fig. 9b.

In summary, this section explained the methodology to carry the experiment, general block diagrams for every proposed design of bilateral control system and introduced observer, modelling equations and pictorial illustration of the robot operating in simulated software. Detail procedures and control setup have been described to explain every test that has been carried out.

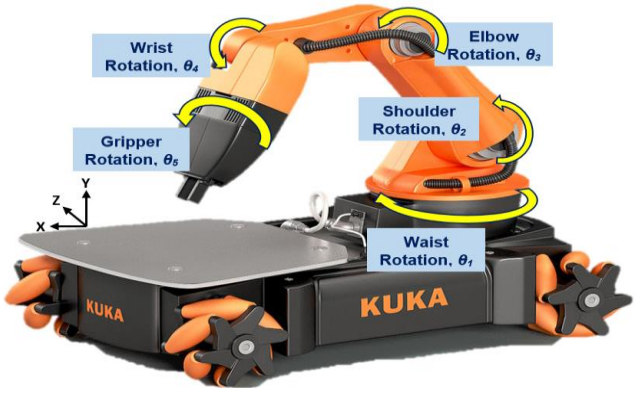


Fig. 7. Visualisation for Five Degrees of Freedom (DOF) of the YouBot Arm.

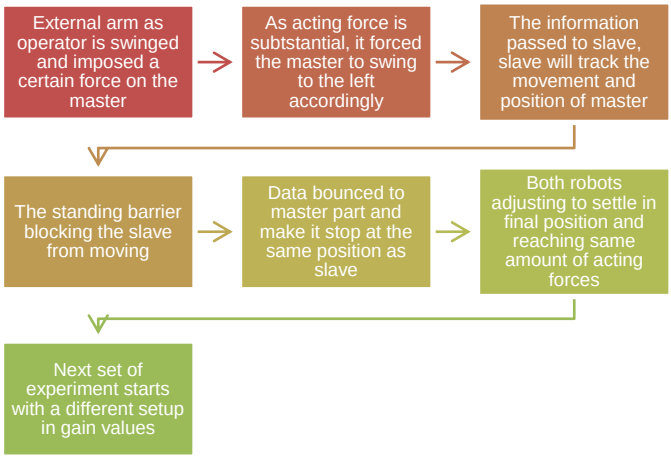
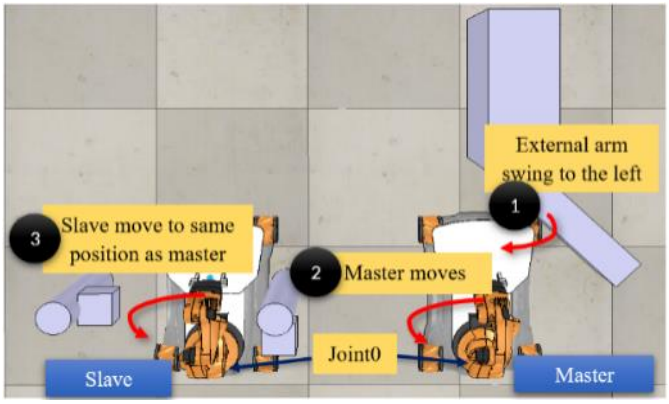
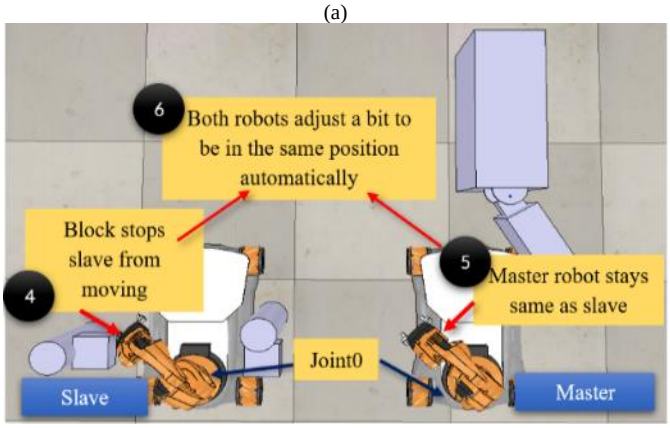


Fig. 8. Procedure Undergoes by the Robotic Manipulator.



(a)



(b)

Fig. 9. (a) and (b). Overall view and Operation Steps for the Bilateral Control System.

III. Results and Analysis

This section discusses the proposed system's outcome, and output feedback gathered after running separate tests. All findings from experiments have been acquired and assembled into table form and constructed into graph form to observe the disparity between all versions of control systems. Each experiment has been conducted according to the parameters and variable setlist. For every refreshed assessment set, all steps are repeating for three times to compute the mean values before being illustrated in the graph version. These independent variables of K_p and K_d values are presented in Table I below. For RFOB, the priority is to validate the Differential Mode Law of the bilateral master-slave telerobotic arm manipulator system. The test is to verify that the system abiding by the law of subtraction between master and slave for position reading is equal to zero. The test also demonstrates that the integration of two types of different observers is capable of enhancing the system response and lowering the noise value within the internal system. The selected gain values are based on the experimental validation approach. Numerous trials have been performed on an extensive range of variables to determine appropriate lowest and highest values that work compatibly with the proposed system.

TABLE I. IDENTIFIED SET OF CONTROLLER GAIN VALUES

ω_n	$K_p(\omega_n)^2$	$K_d(2\omega_n)$
1	1	2
2	4	4
5	25	10
10	100	20
20	400	40
50	2500	100
100	10000	200
200	40000	400
500	250000	1000

A. Evaluation on Force Control

The first analysis is performed with RFOB is paired to the inner loop output feedback of the working system. The proposed control loop is implemented to monitor the force reaction happening in the system. The recorded force reading generated on both subsystems is displayed in Fig. 10 to Fig. 18, respectively.

Entire graphs from Fig. 10 to Fig. 18 shows the feedback in forces for master and slave youBot in a different set of settings, corresponding to every value of ω_n as listed in Table I. There are two types of forces produced in every graph. The first generated formed in blue colour signifies the torque reading yield from the master; meanwhile, the second line in orange represents the torque induced from the joint at slave robot. It should be noted that the maximum torque for Joint0 for both subsystems is set at 8N.

Referring to graphs in Fig. 10 to Fig. 14, when the force in the master began to grow into 8Nm after being pushed by the external manipulator, the reading value in the slave varied and generated a series of gradual increments in force reading. At this point, the slave did not move and remained in its original position at 0˚. This is because the gain values of K_p and K_d are too small and insufficient to increase the controller's sensitivity in the proposed bilateral control system. Although the force reading in the master climbed up for a certain period after being pushed, the feedback and data circulated from the master subsystem were considered ignored by the slave. This is because the slave cannot read the exact information passed through its subsystem. Force generated on the said joint is collectively unstable and fluctuating.

Nevertheless, as exposed in Fig. 10 to Fig. 12, force reading on both systems is restored to zero after the external manipulator returned to its early position, stopping it from pushing the master youBot arm ahead. There are some instabilities traced in force reading of master and slave as shown in Fig. 13 and Fig. 14 after the value of K_p and K_d is increased up to K_p = 100, K_d = 20 and K_p = 400, K_d = 40. At this point, force reading became unstable and wavering, evidently referred to the current scenario in the simulation workspace. Both robots attempt to identify each other's final pose when the slave robot quakes after being in contact with the obstruction block.

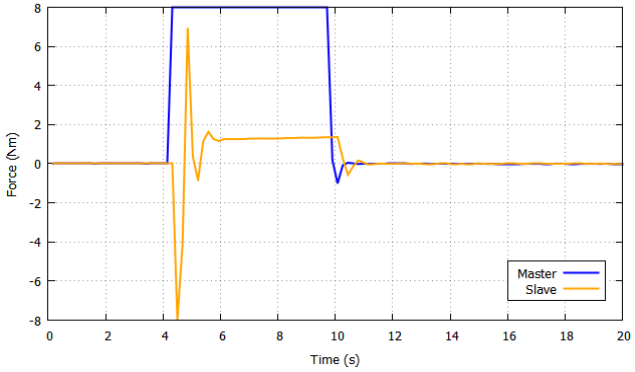


Fig. 10. Generated Force on both Subsystems across Period for ω_n = 1.

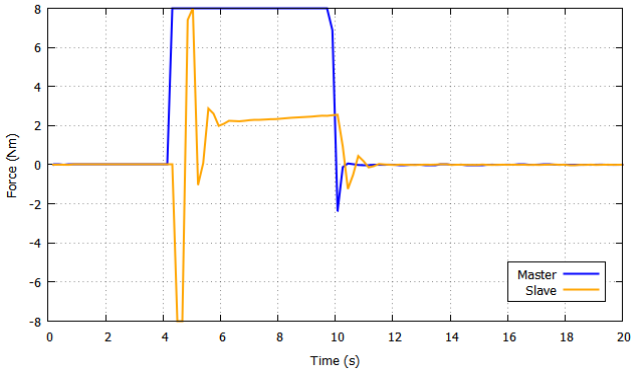


Fig. 11. Generated Force on both Subsystems across Period for ω_n = 2.

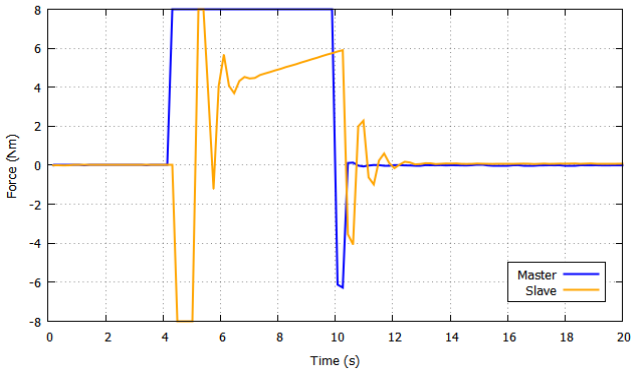


Fig. 12. Generated Force on both Subsystems across Period for ω_n = 5.

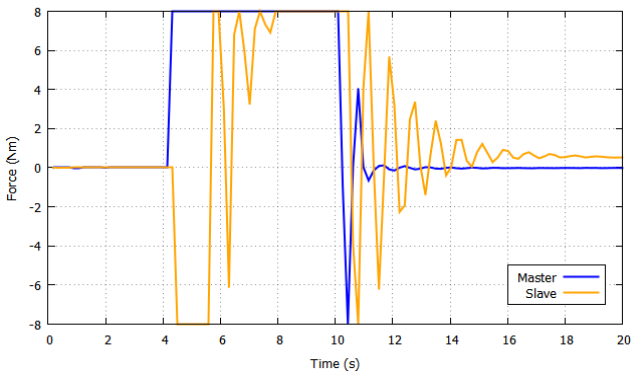


Fig. 13. Generated Force on both Subsystems across Period for ω_n = 10.

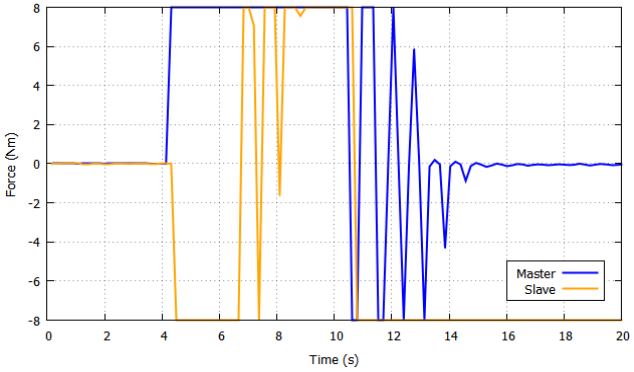


Fig. 14. Generated Force on both Subsystems across Period for ω_n = 20.

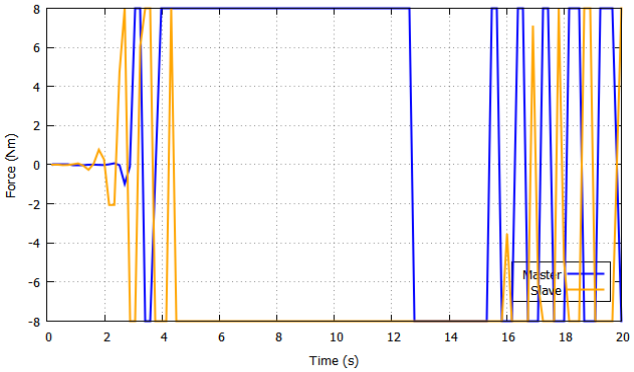


Fig. 15. Generated Force on both Subsystems across Period for ω_n = 50.

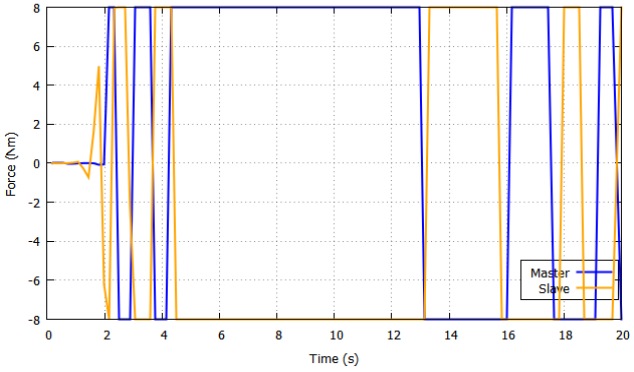


Fig. 16. Generated Force on both Subsystems across Period for ω_n = 100.

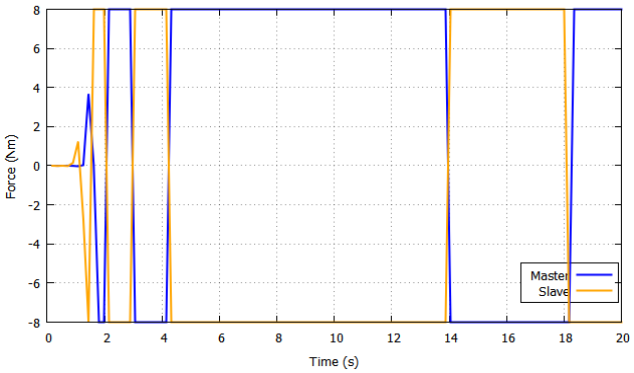


Fig. 17. Generated Force on both Subsystems across Period for ω_n = 200.

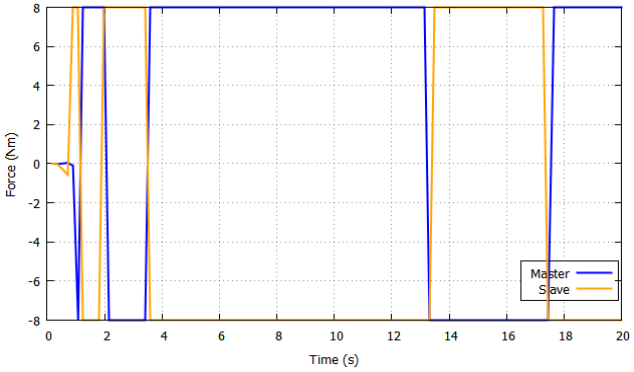


Fig. 18. Generated Force on both Subsystems across Period for ω_n = 500.

Whereas in Fig. 15 to Fig. 18, force reading at slave's joint shows a finer trend of the desired output and corresponds to the reading in master's joint compared to the prior. The gain value for each parameter has been more extensive. At t > *3.5s*, the external arm began to press on the surface of the master youBot. After receiving information from the partner identified as master youBot, the slave youBot will detect the same applied force and proceed ahead. When the block stopped the slave, it instantly applied a counterforce response and attempted to move forward. The shape of the graph can describe this condition in Fig. 15 to Fig. 18. Graphs revealed that the master youBot arm used its maximum torque of +8N to go further. However, the slave youBot arm attempted to withstand greater force coming from the block, resulting in -8N of torque reading in the experiment. The greater the gain values of K_p and K_d were thrust to the system, the greater the connection between the magnitudes of the input signal and the magnitude of the output signal in a steady state. In short, force reading in blue showed that the master youBot arm delivered a maximum torque valued at +8N to move. In contrast, the slave youBot arm tried to counter the enormous force from the block, which resulting -8N of opposing torque value in the experiment. After t > *12s*, the external manipulator returned to its original position and ceased to exert 8N of torque for pushing the master youBot arm. At this moment, there is no outside force acting on the master youBot to propel it forward. Nonetheless, the force reading can be traced on both sides of master and slave as these two are swinging back and forth before settling on their initial position, which is at 0˚. Taking the force reading from the graph and simulation of the system, the robots oscillated for a short time as they strive to settle and eradicate the value of the disturbance before returning to the position in proceeding.

B. Evaluation of Position Control

The second experiment aimed to recognise position control for both single links of the youBot arm (at Joint0) when RFOB is implemented to enhance and operate the system. Thus, all graphs from Fig. 19 to Fig. 27 display the position readings of Joint0 for each master and slave robotic arm.

Every graph from Fig. 19 to Fig. 27 above depicted the position feedback of the single joint in master and slave youBot in a separate set of parameters ranging from ω_n = 1 to ω_n = 500, associated to a list of ωn determined in Table I. The blue line indicates the reference angle, also known as step

input, while the green line in the graphs indicates the joint position of the master, and lastly, the red line signifies the joint position of a slave.

Initially, both Joint0's master and slave manipulator positions are set at 0°. Referring to graphs in Fig. 19 to Fig. 20, the reading for the position of Joint0 in master youBot reached its peak time at around t > *9.5s*, and the situation was maintained until t = *20s*. The position of slave youBot touched the highest at 15° in Fig. 21, 5° in Fig. 20 and 2° in Fig. 19. In comparison, the position of a slave has minimally increased after t > *5s*, although it did not have much difference compared to the original position. The position of the master reached its peak time values around 70° at *9.5s*, as portrayed in Fig. 19 to Fig. 21, while *10s* for Fig. 22 to Fig. 23 and 78° at *13.5s* in Fig. 24.

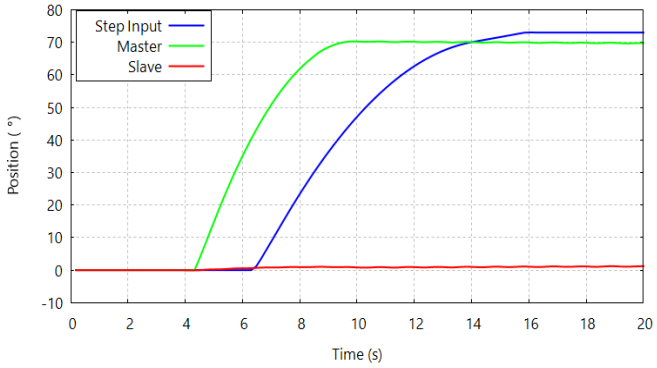


Fig. 19. Position Tracked on both Subsystems across the Period for $\omega_n = 1$.

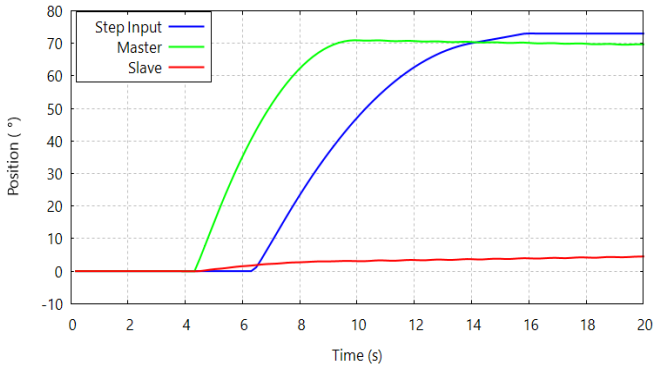


Fig. 20. Position Tracked on both Subsystems across the Period for ωn = 2.

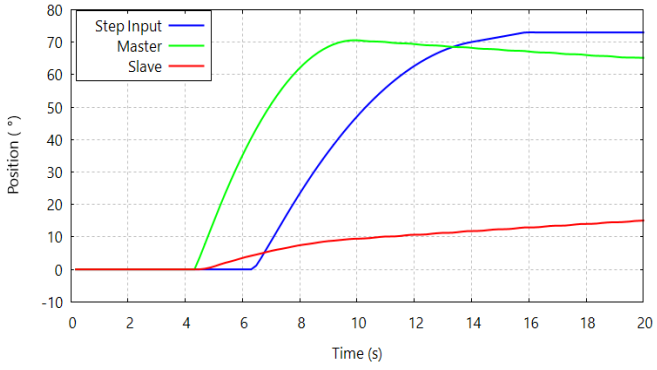


Fig. 21. Position Tracked on both Subsystems across the Period for ωn = 5.

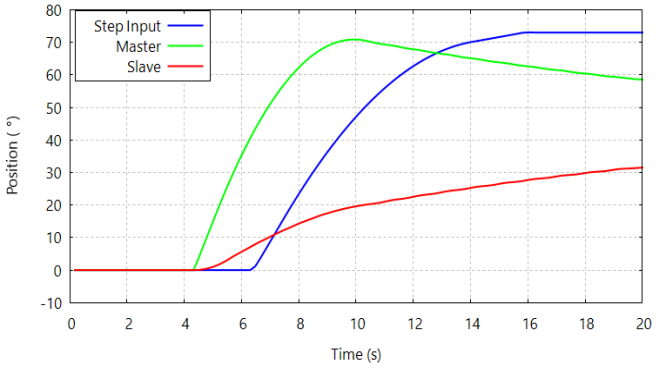


Fig. 22. Position Tracked on both Subsystems across the Period for ωn = 10.

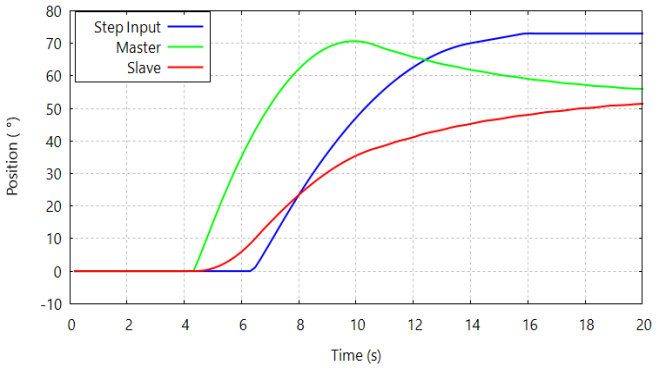


Fig. 23. Position Tracked on both Subsystems across the Period for ωn = 20.

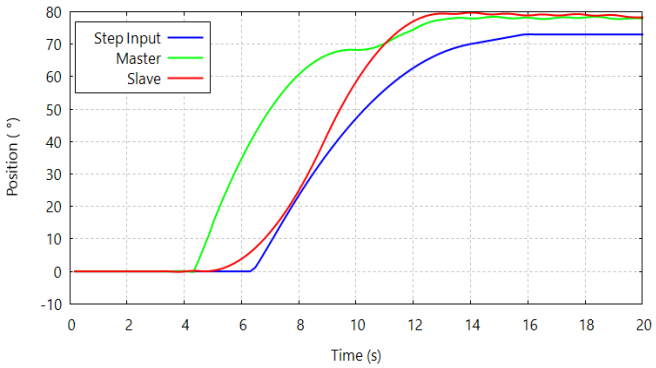


Fig. 24. Position Tracked on both Subsystems across the Period for ωn = 50.

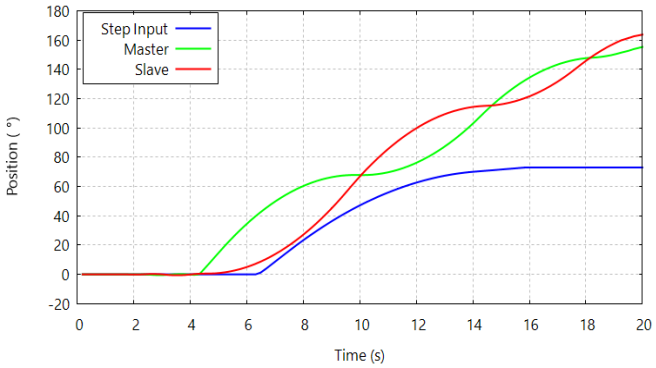


Fig. 25. Position Tracked on both Subsystems across the Period for ωn = 100.

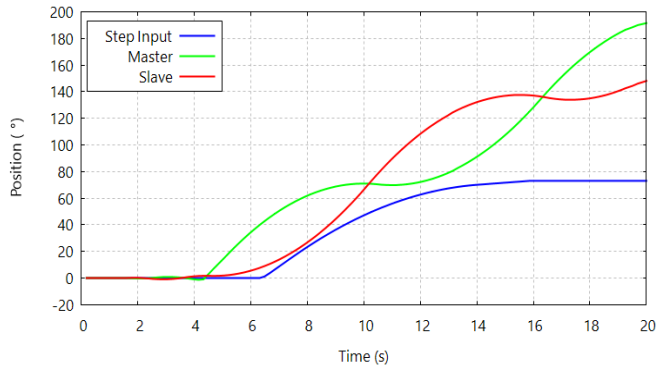


Fig. 26. Position Tracked on both Subsystems across the Period for ωn = 200.

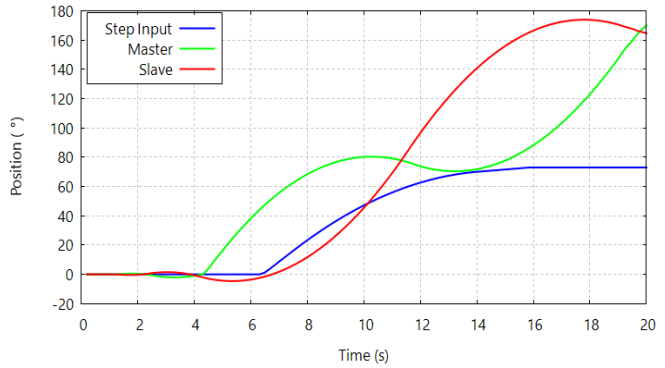


Fig. 27. Position Tracked on both Subsystems across the Period for ωn = 500.

While for Fig. 25 to Fig. 27, both master and slave alternately swing while gradually increasing on its position. The reading from the graphs kept going up because the master arm received an external force that made it move forward. However, the position of the slave according to Fig. 19 until Fig. 21 changed in minimal value compared to the master arm robot because the gain value is not adequate, having neither effects of the percentage of overshoot nor settling time. Referring to Fig. 24 and Fig. 27, when the value of gain is increased, the angle of position in the slave robot started to have changed, moved further and stretched at the top position, recorded at 80° for stable condition and 178° for unstable condition. On the other hand, the reading for the master's position began to drop, similar to the slave's position reading, as shown in Fig. 24. Following the graphs in Fig. 25 until Fig. 27, it can be demonstrated that slave youBot was trying to move further to track the position of master and master concurrently trying to catch the latest post of the slave.

Nevertheless, delays are noticed during the communication process of the two subsystems, thus making the position reading fluctuate until the end of the experiment. The position value of master and slave surpassed the reference angle when the value of ω_n is 50 until the value of ω_n was set at 500. Graph reading in Fig. 24 also presented that the differences in the value of error between the position of master and slave with reference angle are the lowest compared to others which are noted at -6° to -7°, and the most significant error is observed from the graph in Fig. 19 with the value of -70°. Based on the output reading, the percentage of accuracy for each design control is computed and tabulated into a table form. Table II below shows that control systems using RFOB have achieved greater accuracy compared to other systems, while Table III organised the transient response of various frequencies.

TABLE II. ACCURACY RECORDED IN EVERY DESIGNED SYSTEM

ω_n	$K_p(\omega_n)^2$	$K_d(2\omega_n)$	Accuracy *(%)*
1	1	2	1.78
2	4	4	6.49
5	25	10	23.09
10	100	20	53.79
20	400	40	91.82
50	2500	100	98.94
100	10000	200	99.54
200	40000	400	92.66
500	250000	1000	95.31

TABLE III. COMPARISON OF TRANSIENT RESPONSE FOR VARY FREQUENCIES

ω_n	Peak Position (°)		Peak Time (s)		Delay (s)	
	Master	Slave	Master	Slave	Master	Slave
1	70	2	9.5	20.0	4.5	6.0
2	70	5	9.5	20.0	4.5	5.0
5	70	15	9.5	20.0	4.5	5.0
10	70	32	10.0	20.0	4.5	5.0
20	70	52	10.0	20.0	4.5	5.0
50	80	80	13.5	13.5	4.5	5.0
100	158	162	20.0	20.0	4.5	5.0
200	190	170	20.0	20.0	4.5	5.0
500	170	178	20.0	17.5	4.5	7.0

From nine designed systems, six of them achieved more than 90% of accuracy. The system achieved the most accuracy with ω_n = 50, steadily at 99.78%, followed by ω_n = 500 and ω_n = 20, with each of them reaching more than 95% accuracy. Meanwhile, the system with the most negligible accuracy for differential mode law is recorded at 3.56% when the value of ω_n = 1. Therefore, the best design of the proposed system to acquire the most satisfactory position control is ω_n = 500. This is because master and slave robots attained the same final position after 15s, had a minor steady-state error, and achieved critical damping. For common mode law, the ideal design for the controller is when ω_n = 50. As prove, the summation of torque reading observed at both joint of master and slave system revealed to be equal to zero and experienced more minor disturbance as seen in Fig. 19.

IV. DISCUSSION AND CONCLUSION

To sum up, the outstanding response derived from all analyses for both experiments, the results are outlined in Table IV accordingly. The most satisfactory outcome is highlighted as the best parameter for each mode law.

TABLE V. RESPONSE IN BOTH LAW OF BILATERAL CONTROL SYSTEM

	Common Mode	**Differential Mode**
Compatible	K_p=250000 and K_d=1000	K_p=2500 and K_d=100
Incompatible	K_p=1 and K_d=2	K_p=250000 and K_d=1000
Delay and overshoot		-0.5s, from 11s to 15s
Accuracy		99.78%
Total equation	Summation of forces between master and the slave is zero	Position difference between master and slave is almost zero
Force pattern	Fewer vibrations, more stable	

According to the above summary in Table II, K_p=2500 and K_d=100 are considered as the best gain for the integrated system. Once RFOB is employed into the system, six systems achieved greater than 90% accuracy, compared to only five systems that reached more than 90% accuracy without the RFOB. Above all, the form of noise produced at Joint0 in the second experiment is much more refined than in the first experiment. This proves RFOB has advantages in improving the system stability and eliminating periodic oscillation. Furthermore, the recommended technique is intended to remove unwanted signals such as Coriolis forces, F_c, viscous damping friction, and gravity forces produced internally, especially on the motor located at a specific link. To achieve an accurate estimation of forces to be delivered to the operator, the dynamical effects within the force signal must be adjusted.

The RFOB experiment required around 11s to 15s to attain an overshoot in terms of the time delay. In RFOB configuration, the settling time generated by slave youBot is considerably better and sharper. Furthermore, the amplitude of the curve formed is relatively constant and persistent until the target joint reaches its final position. Nevertheless, the system is underdamped for a moment before it progressively climbs to reach overshoots and peak times. For the record, underdamped is a condition in which the system oscillates slower at a low-frequency rate and takes longer to get a steady-state. This situation occurred whenever the value of the PD controller increased. As a result, the stability of the control system may be derived to be conditional, based on the value of gain and threshold. In overall, the proposed designs can reach stability in a certain level of gain levels but can deteriorate when the gain value is unsuitable.

The main objective to validate a master-slave control system with DOB and RFOB implementation abiding by the law of bilateral control system has been successfully confirmed. The stated target is justified by running several sets of parameters and collecting the output response for the analysis. The suggested technique has met the capacity in enhancing the whole system performance. This study also demonstrates that a control system can be sensorless to measure dimension when active reacting forces are in contact with the system. The proposed technique is also applicable to multiple applications of industrial robots aside from position and force tracking. Using the observers to replace old-style force sensors on a device or equipment increased the system reaction and improved internal uncertainty across the system.

In conclusion, few analyses are performed to determine every single type of common controller response, rankings in the percentage of accuracy, overshoot, settling time and delay. The best kind of controller was chosen based on its performance in all three studies. This conclusion is backed by the fact that both robots oscillated at a reduced angle for a time in a control system without applying the DOB approach before they stabilised and came down to remain in a single spot. Whereas the idea for introducing RFOB into the system will remove the unwanted disturbances and errors that arrive before being feed at the DOB loop. Different experiments with diverse parameters have been conducted to oversee the system's latency after applying DOB and RFOB as part of the control loop technique. This work also assesses the system reaction for each suggested design of the bilateral system with a varied set of controllers. Indirectly, the primary purposes of this study were successfully attained.

Several analyses were performed to determine each controller's performance, standings in the accuracy, time-delay, and settling time. The best controller was chosen based on its performance in all three studies. This conclusion is supported by the fact that both robots oscillated at a smaller angle for a while in a control system without employing the DOB technique before they stabilised and came down to remain in a single place. Whereas the method to add RFOB into the system will subtract the uncertainties coming on the input of DOB. To observe the potential and advantage of applying DOB and RFOB onto the system response, separated tests with different parameters have been performed to examine and evaluate the system reaction for each proposed controller design. Indirectly, the second and third objectives of this study were successfully achieved.

Several proposals may be offered and studied in the near future research to enhance the system and make it more resilient. First, video and visual input and recording might be implemented by putting a high-speed camera at the tip of the robot arm as an added sensor for tracking. The extracted data from raw images and videos will undergo image processing technique into a type of visual force to suppress the signal error and remove the unwanted noise [10]. Besides, the relevant data gained from the pictorial records will be utilised to increase the coordination of the system's trajectory [11]. The visual data feeding into the reaction force estimation loop system will be used for the soft navigation system. Assuredly, the research should be undertaken on actual hardware and real interaction to assess the response and stability in a real-time experiment.

Additionally, this project can be one of the working mobile robots dispatched to risky or work in remote regions due to its versatility. Aside from that, a specific experiment is recommended to execute by employing image processing technology on Linux based operating system with a robot operating system (ROS) connected to the entire robotic arm. The information processed from visual data can be fed into the system for more outstanding object tracking and positioning accuracy. The limitations of executing the current technique on a simulator and virtual platform would lower the capacity of the produced feedback and overall system performance. This control procedure is likely to be more responsive in

actual ROS communications. It is easy to track any unaligned output tracking or missteps while executing the robot's task. Not just that, implementing a real-time based system for subsequent study can boost the time responsiveness on the machine and actively doing force and position tracking.

ACKNOWLEDGMENT

The authors would like to acknowledge the funding supplied by Universiti Teknikal Malaysia Melaka (UTeM) under the Scheme of Zamalah to perform the study. The authors would also like to convey our credits to the Robotics and Industrial Automation research group within the Faculty of Electrical Engineering.

REFERENCES

[1] N. Hiroshi, O. Kiyoshi, Y. Yuki, K. Naoki, M. Toshimasa and T. Akifumi, "Force Sensorless Fine Force Control Based on Notch-Type Friction-Free Disturbance Observers," IEEJ J. Ind. Appl., vol. 7, 2018, pp. 117-126.

[2] Al. Kolsanov, S. Chaplygin, S. Rovnov and A. Ivaschenko, "Augmented Reality Application for Hand Motor Skills Rehabilitation," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 11, no. 4, 2020.

[3] G. Liming, Y. Jianjun & Q. Yingjie, "Torque control based direct teaching for industrial robots considering temperature-load effects on joint friction," Ind. Rob., vol. 46 no. 5, 2019, pp. 699-710.

[4] U. Kodai, S. Tomoki, T. Kenta, S. Sho and T. Toshiaki, "Reaction Force Estimation of Electro-hydrostatic Actuator Using Reaction Force Observer," IEEJ J. Ind. Appl., vol. 7, 2018, pp. 250-258.

[5] J. Seul and L. Joon, "Similarity Analysis Between a Nonmodel-Based Disturbance Observer and a Time-Delayed Controller for Robot Manipulators in Cartesian Space," IEEE Access, vol. 9, 2021, pp. 122299-122307.

[6] S. Choi and K. J. Kuchenbecker, "Vibrotactile Display: Perception, Technology, and Applications," Proc. -2013 IEEE, vol. 101, no. 9, 2013, pp. 2093-2104.

[7] E. Sariyildiz, and K. Ohnishi, "A Comparison Study for Force Sensor and Reaction Force Observer-based Robust Force Control Systems", Proc. -2014 IEEE 23rd Int. Conf. Ind. Electron., 2014, pp. 1156-1161.

[8] E. Sariyildiz, S. Hangai, T. Uzunovic and T. Nozaki, "Discrete-Time Analysis and Synthesis of Disturbance Observer-Based Robust Force Control Systems," IEEE Access, vol. 9, 2021, pp. 148911-148924.

[9] N. N. Mansor, M. H. Jamaluddin and A. Z. Shukor, "Concept and application of virtual reality haptic technology: A review", J. Theor. Appl. Inf. Technol., vol. 95, no. 14, 2017, pp. 3320-3336.

[10] A. Protsenko, "The Development of a Fault Detection and Identification System for Executive Units of Manipulators Using Technical Vision," Int. Multi-Conf. Ind. Eng. Mod. Technol. FarEastCon, 2020, pp. 1-5.

[11] Z. I. Bell, P. Deptula, H. Chen, E. A. Doucette and W. E. Dixon, "Velocity and Path Reconstruction of a Moving Object Using a Moving Camera," Proc. Am. Control Conf., 2018, pp. 5256-5261.

Assessment System of Local Government Projects Prototype in Indonesia

Herri Setiawan[1], Husnawati[2], Tasmi[3]
Department of Informatics, Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia[1]
Department of Computer System, Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia[2, 3]

***Abstract*—The purpose of this research is to build an application that is used for project evaluation and provide recommendations on project performance in local government agencies. In this study, project evaluation was carried out using the Group Decision Making (GDM) model based on the Group Decision Support System (GDSM) concept. The project output and outcome parameters used by the Decision Maker (DM) use a hybrid of the Multi-Criteria Decision Making (MCDM) and Project Management Body of Knowledge (PMBOK) methods to reduce subjectivity in scoring qualitative data and to determine project ratings from all DMs. Copeland Score voting method. The results of application computing on the implementation of Group Decision Support System (GDSS) and MCDM indicate that the project ranking process will be faster and more accurate. The results of the sensitivity test show that two criteria have a great influence on project performance so that they have a very important role in project evaluation.**

Keywords*—Group Decision Making (GDM); Group Decision Support System (GDSS); MULTI-CRITERIA DECISION MAKING (MCDM); Project Management Body of Knowledge (PMBOK); local government

I. INTRODUCTION

Projects in government agencies are part of the program, and consist of a set of actions to mobilize resources, either in the form of personnel (HR), capital goods including equipment and technology, funds, or a hybrid of some or all of these types of resources as inputs to produce outputs. in the form of goods/services. Measurement of project performance in government agencies, including the current project, has a weakness because it does not reflect the project evaluation that is generally carried out, which is only based on the percentage of achievement of the planned level of achievement of each project performance indicator as determined through the successful realization of the indicators in question [1].

The calculation of the percentage of achievement of the planned level of activity achievement used is based on the absorbed funds and the realization of output between the realization and the plan, which is stated in the administrative document of the Government Agency Performance Accountability Report (LAKIP). There can be no correlation between the output produced and the expected outcome. In the current LAKIP measurement method, including project evaluation, the criteria used to measure organizational performance are only limited to quantitative criteria, namely timeliness of implementation and effectiveness in the use of financial resources. Research conducted [2] developed a description of how to determine KPIs in ICT projects to get ideas and solutions in evaluating ICT projects.

Based on the problems an institutional decision-making system is needed which is an activity that can be carried out by individuals, groups, or organizations. Also, special steps need to be taken so that the group's decisions can be agreed upon and binding on the various parties involved. According to the viewpoint of software engineering, instances of existing applications that utilize notoriety as well as trust approaches incorporate shared, peer to peer (P2P) communications, internet business, e-advertising, multi-specialist frameworks, web search tools and Group Decision Making (GDM) situations [3][4]. The decision is the result of an evaluation of the selection of the best alternative, which involves the relevant parties. With the number of considerations and desires that must be considered, decision-making needs to listen to the considerations of many people.

One solution that is widely offered in making decisions using computing is the Group Decision Support System (GDSS). The model was made with various approaches, one of which used an approach to GDSS [5]. This model is formulated regarding social choice theory. The model is structured using a voting mechanism, in such a way that it allows each decision-maker to express their choice. Research shows that this model can accommodate Multi-Criteria Decision Making (MCDM). In modeling, the choice (vote) of the DM is considered. MCDM has played a role as an instrument for many individuals to choose candidates or alternatives. From its meaning, multi-criteria decision-making (MCDM) is used to determine the best alternative from several alternatives based on certain criteria [6].

There have been many previous studies using this method such as [7][8][9][10][11][12]. One of the popular methods used for this is the Analytic Hierarchy Process (AHP). This method uses human perception input and can process qualitative data, resulting in a comprehensive decision-making model. This is because AHP can solve multi-objective and multi-criteria problems based on the comparison of preferences of each element in the hierarchy. However, the AHP method has a weakness because the main input is in the form of perception so that it involves subjectivity, this will be a problem if the DM gives a wrong assessment. Another popular method used within the project scope is the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). This method is used because it is closely related to the benefits (benefits) and costs (costs) in a project. This method is based on the concept that the best-chosen alternative not only has the shortest distance

from the positive ideal solution (benefit) but also has the longest distance from the negative ideal solution (cost). The main weakness of the TOPSIS method is that it does not provide elicitation of weights and check the consistency of the assessment [13].

This study will create a new model to facilitate the achievement of consensus among DM while respecting different preferences, interests, and values. GDSS is expected to expand DM capabilities, but not to replace DM assessment. What is to be realized is "a new decision model that is implemented in a computer-based system that supports a group of people who are members of the same task or goal and have one particular tool that functions to interconnect the people in the group". This is known as the Group Decision Support System (GDSS) [14].

The purpose of this research is to develop a Group Decision Support System (GDSS) model for project evaluation, which not only fulfills administrative or normative needs but also provides a more objective evaluation. In this study, assessment criteria will be used that can represent the assessment in the scope of project evaluation, especially evaluation of projects in local government agencies. Considering aspects of the applicable legislation, Decision Makers (DM) are involved, namely: Government Institution Executives, Project Management Work Units, Business Process Owners Units, and the Community represented by DPRD, to provide assessments and evaluations of project implementation in institutions local government. Then, in the evaluation process, a GDSS concept was used. The GDSS concept can overcome inconsistencies that may occur in decision making because with GDSS decisions are made based on a mathematical calculation model. Project evaluation is carried out by DMs on output and outcome parameters, using a hybrid MCDM method based on the established criteria. In the weighting of the criteria used weighting techniques in the AHP method. Furthermore, the weight of the resulting criteria is used as input for the TOPSIS method to generate project rankings for each DM. At this stage, scoring of project qualitative data is carried out based on the Project Management Body of Knowledge (PMBOK) to reduce DM subjectivity. As the last step, to determine the project ranking of all DMs, the Copeland Score voting method was used.

II. Literature Review

The appraisal process is an important step in evaluation because it underlies the successful evaluation of a project. Through a systematic literature review, several project evaluations models have been identified and analyzed. The results show that the effectiveness and efficiency of evaluation are increasingly important for a project [15]. In this regard, several researchers have implemented MCDM methods, both in the DSS or SPKK [16][8][9][10]. Many organizations realize the importance of the MCDM method because the use of the MCDM method increases effectiveness in decision-making. As in the research conducted [8], it is stated that the first step to reducing the risk of project failure is to choose the optimal project. The effectiveness of the criteria in selecting the most optimal project was identified and defined by the MCDM approach, using the AHP and TOPSIS methods. The use of the proposed model can help companies facilitate a systematic approach in making the right project selection decisions.

Effective and efficient project selection has an important meaning in every organization because the decision-making process to assess the feasibility of a project is very complex [16]. The method used in this research is AHP and Moora. Modeling is based on various types of logic, considering the existence of various criteria, the objectives of the decision-maker, and the nature of the complexity of the evaluation process. According to him, the main advantage of MCDM is that it provides decision-making by analyzing complex problems, can aggregate criteria in the evaluation process, and provide scope for decision-makers to actively participate in the decision-making process.

Research [9] proposes the application of the MCDM framework to monitor and measure ongoing project performance. Linear Programming (LP) and MCDM methods are used to evaluate decision-making on the selection of project priorities. An MCDM approach is also proposed [17] to evaluate Product Development (PD) effectively. After the criteria hierarchy is built, the weight of the criteria is calculated using the AHP method. The Vikor method is used to rank the results at a later stage. The results of the sensitivity analysis show that the AHP-Vikor integrated model can accommodate the evaluation of the criteria weights. And the results of empirical studies show that the proposed evaluation framework can solve the problems that arise.

The Project Management Body of Knowledge (PMBOK) was developed by the Project Management Institute (PMI), an organization in America specializing in project management development. PMBOK is a guide that contains knowledge in project management and is always updated within a certain time. Project management is the application of knowledge, skills, tools, and techniques in project activities to meet project needs. PMBOK generally develops 9 (nine) areas that must be understood in project management, namely: Project Integration Management, Project Scope Management, Project Time Management, Project Cost Management, Project Quality Management, Project Human Resource Management, Project Communication Management, Project Risk Management, Project Procurement Management [18]. According to [19], managing ICT projects is seen as a challenging activity, because it involves a balanced portion of tangible and intangible resources. Various studies of independent institutions in America and Europe say that more than 70% of ICT projects are considered a failure, in the sense of not meeting the targets set previously at the planning stage. One of the causes of this failure is due to the indiscipline of stakeholders who are directly involved with ICT projects in complying with the standard project implementation and control methodologies that have been outlined.

Copeland Score is one of the voting methods whose technique is based on reducing the frequency of wins with the frequency of defeats from pairwise comparisons [20]. group decision-making to determine gene abnormalities in cancer.

III. METHODOLOGY

This research will use program/activity data in districts/cities of South Sumatra Province, Indonesia. This study builds a project evaluation GDSS model in local government agencies in determining the best project ranking, using a hybrid method of MCDM and PMBOK. The framework of this research can be seen in Fig. 1.

A. Proposed Model

The proposed model is a Group Decision Making (GDM) model based on the GDSS concept, using the methods in Multi-Attribute Decision Making (MADM) to determine the best project from several alternatives based on several predetermined criteria. Then the MADM combination method in the SPKK will be developed based on the Analytical Hierarchy Process (AHP) method, Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and Copeland Score, each of which has its role.

Based on Fig. 1, to determine the weight of the criteria, the weighting technique contained in the AHP method is used, then the results of the weighting of these criteria will be input in the TOPSIS calculation which is used to determine the ranking as a result of evaluating the performance of these activities. At the scoring stage, project qualitative data is based on the Project Management Body of Knowledge (PMBOK) to reduce DM subjectivity.

From the TOPSIS calculation, the project ranking for each DM is generated, and to unify the differences in preferences between DMs, the Copeland Score method is used as a voting method to determine the best project ranking from all decision-makers. Fig. 2. is a GDM model for the evaluation of ICT projects in government agencies.

In the GDM component, each component plays a role in the decision-making process as described. Each DM performs a weighting against the assessment criteria it has. The next stage is that each DM assigns a score to the alternative (ICT project) so that the project ranking of each DM is generated. The final step is to rank projects from all DMs using the voting method as a result of group decisions. Fig. 3. describes the decision-making process carried out individually and in groups in the GDM model.

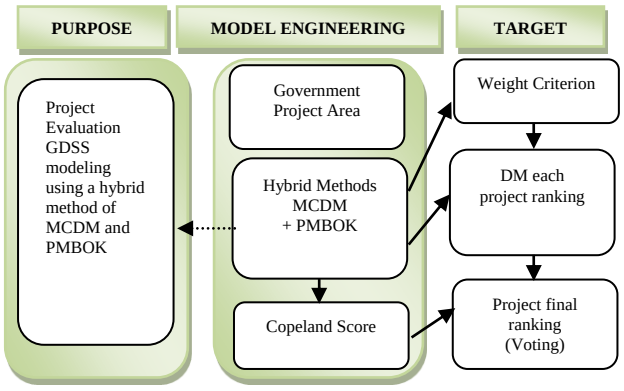


Fig. 1. Framework.

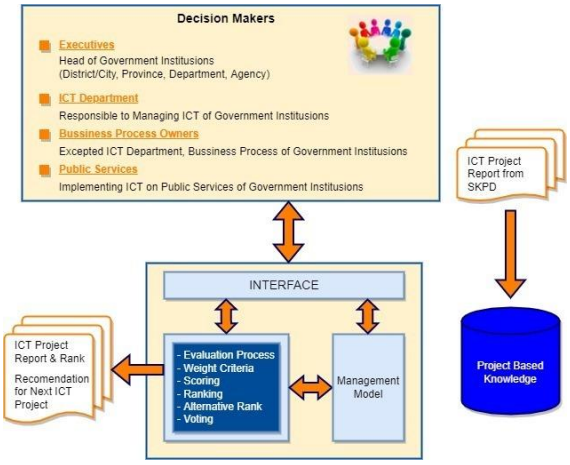


Fig. 2. GDM Model for Evaluation of the Proposed Local Government Information and Communication Technology (ICT) Project.

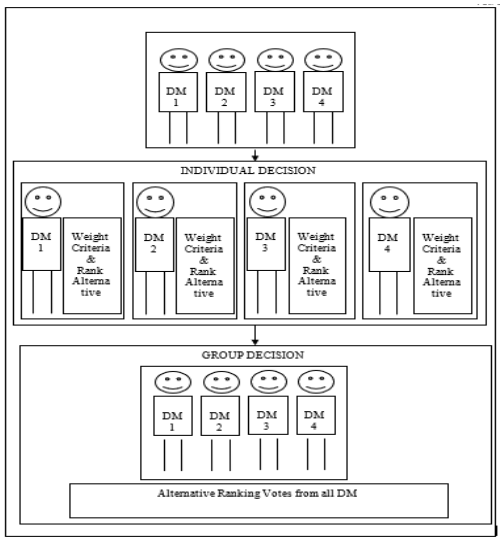


Fig. 3. Individual and Group Decision Making Process in GDM Model.

B. Calculating Weight Algorithm

In this section, a combination of the AHP and TOPSIS methods is carried out with the preparation of a pairwise comparison matrix and weighting the criteria, with the goal of determining whether matrix A is consistent or not as shown in Fig. 4.

1) Each Decision Maker (DM_i) has its assessment criteria (C_{ij}).

2) In process two, matrix A is squared, and in-process three matrix B is calculated. Matrix B is the sum of the elements in the same row of matrix A. Based on Matrix B, the eigenvectors are calculated so that Matrix E is obtained, described in process four.

3) The process of five, six, seven, and eight is to calculate the consistency of the index by deriving a matrix C, which is the product of matrix A and matrix E. Based on matrix C, it is determined whether matrix A is consistent. If it is consistent then the weight of the matrix A is calculated by calling the Algorithm for Calculating the Weighted Normalization Value.

4) The output of this algorithm is the Criteria Weight (Wk) for each DM.

Calculating Weight Criteria Algorithm

Input:
DM_i = {DM $_i$ | i = 1,2, 3..., n};
C_{ij} = {Cij | j= 1,2, 3...};
Output: W_k;
Process:
Begin
1. A[i,j] ;
2. A^2[i,j] = A[i,j] * A[i,j];
3. for k = 1 to i
B(k,1) = $\sum_{i=1}^{i=j} A2(k,l)$;
End
4. for k = 1 to i
E(k,1) = B(K,1)/$\sum B(:,1)$
End
5. C [k,1] = A[i,j] *E[k,1] ;
6. λ max = 1/k * ($\sum_{k=1}^{k=i} \frac{E[k,:]}{C[K,:]}$)
7 C1 = $\frac{\partial - k}{k-1}$
8. if C1 < 10%
W = Bobot (A)
Else
Comparison value evaluation
pairwise comparison on A[i,j]
End
End

Fig. 4. Calculating Weight Criteria Algorithm.

C. Algorithm for Calculating Normalization Value

This algorithm explains the calculation of the normalized value of all alternatives for each criterion and calculates the normalized value of its weight as shown in Fig. 5.

1) The Decision Maker scores all ICT alternatives (projects) based on the assessment criteria it has so that after being converted based on the assessment rating, the results are in the form of a Scoring Matrix (SC).

2) Process 2 calculates the normalized value of all alternatives for each criterion (Matrix X) and process 3 calculates the weighted normalized value (Matrix Y).

Algorithm for Calculating Normalization Value

Input: Matriks Scoring (SC)
W = weight matrix of each criterion
Output: Normalized Scoring Matrix (R)
Normalized Scoring Matrix_Weight (Y)
Process:
Begin
1. i = size (SC,1); j=size (SC,2) // i,j is the dimension of the DM scoring matrix, i is the row representing the number of alternatives (projects) and j is the column representing the number of criteria
2. for k = 1 to i
for l = 1to j
R (k,l) = SC (k,l)/($\sum SC(k,l;)$)
End
End
3. for k = 1 to i
for l = 1to j
Y [k,l]= R [k,l] * E[1,l]
End
End
End

Fig. 5. Algorithm for Calculating Normalization Value.

IV. Design and Implementation System

A. Design

This study continues the research conducted by [21] who has not yet carried out the process of making web-based applications, so this study designs a Web-based GDSS Model Prototype [22] for Project Evaluation in Local Governments. The prototype is built based on a review of research results as described in the Literature Review, which relates to MCDM methods that can be applied in the scope of evaluating a project.

B. Implementation

This system is designed with several stages of implementation to produce an assessment that is used as the basis for making decisions on a project. The results of the application implementation are shown in Fig. 6 to 12.

Fig. 6 is a representation of the login page on the government project evaluation. On the page, the user and password fields are displayed, then the submit button. The home page contains display information regarding login access for users. In this case the user is an administrator who works in the government of the Province of South Sumatra. Users can login with User Name: admin and Password: admin, then press the login button, if the login is successful, it will go directly to the main menu.

When the user successfully logs in, they will immediately go to the main page (Dashboard) of the government project evaluation application. On the main page of the application, there are several menus and a comprehensive list of government projects as shown in the following image.

On the dashboard display from the Fig. 7, there are two menu sections located on the left side of the dashboard display and the upper right corner of the dashboard display. In the menu display on the left side of the dashboard, there are several menus, namely, Dashboard, Activities, Criteria, SKPD, Region, DM, Menu and Users.

From the Fig. 8 shown the activity page, all project activities that are being carried out will be displayed, to add activities, it can be done by selecting the add menu which is in the upper right corner of the activity page display. On the Add Activities page, the Add Activities form will appear, while the data that must be filled in this form include the name of the activity, the activity ceiling, volume and budget year.

Fig. 6. Login Page.

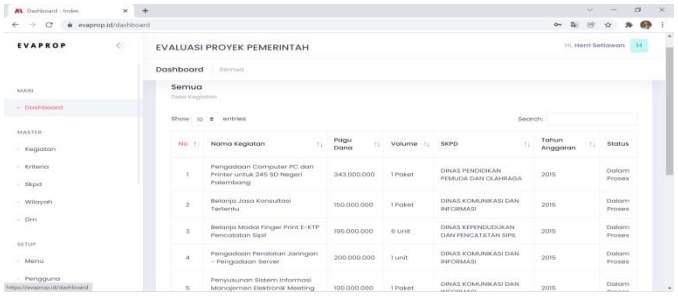


Fig. 7. Dashboard.

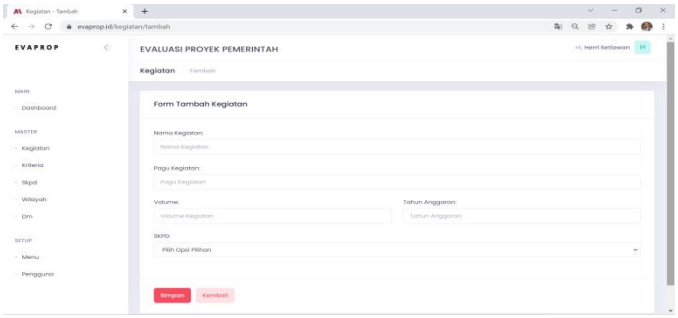


Fig. 8. Activity Setting.

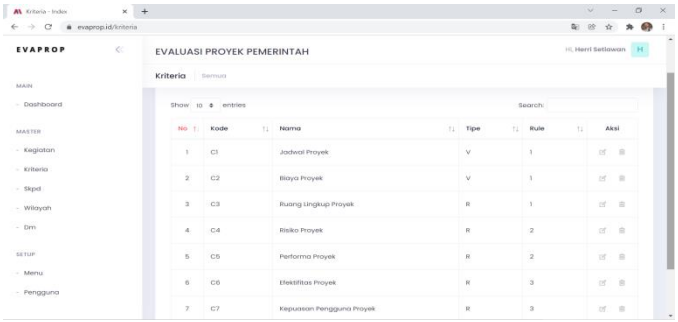


Fig. 9. Criterion Setting.

On the criteria page from Fig. 9, all the assessment criteria used will be displayed, adding criteria can be done by selecting the add menu which is in the upper right corner of the criteria display. On the Add Criteria page, the Add Criteria form will appear, which must be filled in, among others, criteria code, criteria name, criteria type and criteria rule. To save the newly added criteria can be done by selecting the save menu on the added criteria form.

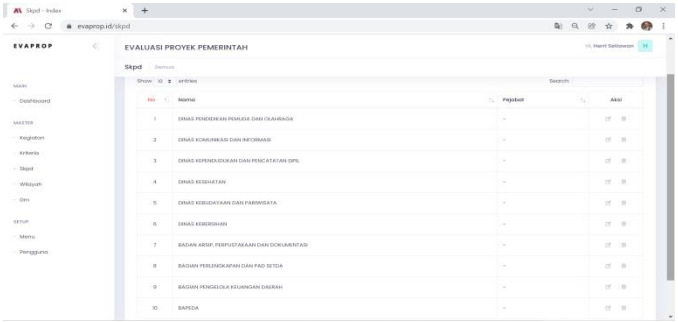


Fig. 10. Setting SKPD.

To fill in the SKPD on the added activity form, choose according to the list of available SKPD, to save new activities, select the save menu on the added activity form. On the Add SKPD page, the Add SKPD form will appear. To add a new SKPD, you must fill in the names of the SKPD and SKPD officials on the Add SKPD form. After all data is filled in, it can be saved by selecting the Save menu on the Add SKPD form.

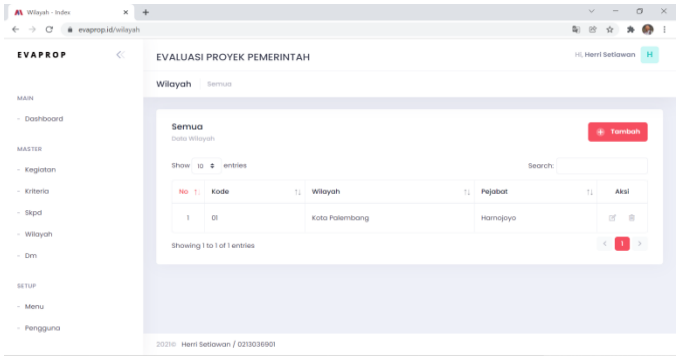


Fig. 11. Area Configuration.

In the regional menu, a regional data page will be displayed, to add regional data, it can be done by selecting the add menu in the upper right corner of the regional page. The Add Region menu will display the Add Region form. To add new area data, you must fill in all the required data on the Add Region form, including area code, area name and official name. After all data is filled in correctly, the data can be saved by selecting the save menu in the Add Region form.

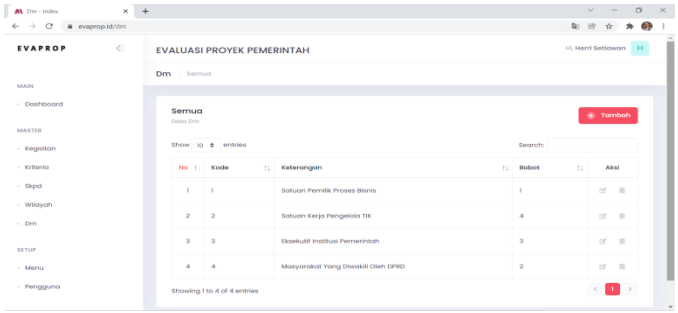


Fig. 12. Decision Maker.

Fig. 12 shows the Decision Maker (DM) page which is the result of weighting for codes 1, 2, 3, and 4 based on the criteria page, each weighting being measured is the Business Process Owner Unit, ICT Management Work Unit, Government Institution Executive, and Communities Represented By the Council. The weighting in each of these categories will be measured based on the criteria that have been determined on the criteria page.

V. Result and Analysis

In the implementation, DM1 will evaluate each alternative (project) based on 3 (three) criteria C= {C1, C2, C3), DM2 will evaluate each alternative based on 2 (two) criteria C= {C4, C5}, DM3 and DM4 will evaluate each alternative based on 2 (two) criteria C= {C6, C7}.

DM1 determines the relative importance between the criteria of Project Schedule (C1), Project Cost (C2), Project Scope (C3). The assessment uses the weighting standard with a scale ranging from 1 to 9 and vice versa, the values of these criteria, according to a DM1 pairwise comparison matrix, are as follows:

	C1	C2	C3
C1	1	0.5	0.3
C2	2	1	0.5
C3	3	2	1

(1)

This stage calculates the priority weighting by finding the eigenvector value of the A matrix through the following process. Squaring the A matrix.

$$\begin{vmatrix} 1 & 0.5 & 0.3 \\ 2 & 1 & 0.5 \\ 3 & 2 & 1 \end{vmatrix} \text{ X } \begin{vmatrix} 1 & 0.5 & 0.3 \\ 2 & 1 & 0.5 \\ 3 & 2 & 1 \end{vmatrix} = \begin{vmatrix} 2.9 & 1.6 & 0.85 \\ 5.5 & 3 & 1.6 \\ 10 & 5.5 & 2.9 \end{vmatrix}$$

The next process is to add up the elements of each row of the A2 matrix so that a matrix is obtained. Then arrange matrix B, and add up all elements of matrix B with the following values below.

$$B = \begin{bmatrix} 5.3500 \\ 10.1000 \\ 18.4000 \end{bmatrix} \quad (2)$$

From the B matrix that has been obtained in the above step, then normalization is carried out on the B matrix to obtain the eigenvector value of the B matrix.

$$E = \begin{bmatrix} e1 \\ e2 \\ e3 \end{bmatrix} \quad (3)$$

e_1=5.3500/(5.3500+10.1000+ 18.4000)=0.1581.

e_2=10.1000/(5.3500+10.1000+ 18.4000)=0.2983.

e_3=18.4000/(5.3500+10.1000+ 18.4000)=0.5436.

$$E = \begin{bmatrix} 0.1581 \\ 0.2983 \\ 0.5436 \end{bmatrix} \quad (4)$$

The three processes above are repeated and at the end of each iteration, the difference between the eigenvector matrix E values obtained and the previous eigenvector matrix E values is sought until a number close to zero is obtained. The matrix E obtained in the last step shows the priority of the criteria indicated by the coefficient of the eigenvector value so that the eigenvector matrix E obtained is:

$$E = \begin{bmatrix} 0.1638 \\ 0.2972 \\ 0.5390 \end{bmatrix} \quad (5)$$

To measure the consistency of the matrix, the first thing to do is to calculate the Consistency Index (CI) by calculating the weighted vector number. The product of the matrix A on the Eigenvalue (matrix E).

$$\begin{bmatrix} 1 & 0.5 & 0.3 \\ 2 & 1 & 0.5 \\ 3 & 2 & 1 \end{bmatrix} X \begin{bmatrix} 0.1638 \\ 0.2972 \\ 0.5390 \end{bmatrix} = \begin{bmatrix} 0.4921 \\ 0.8941 \\ 1.6248 \end{bmatrix}$$

quotient above by the number of elements present, the result is called max (λmax)

λmax = 1/3 * (0.4972/0.1638) + (0.8943/0.2972) + (1.6248/0.5390)

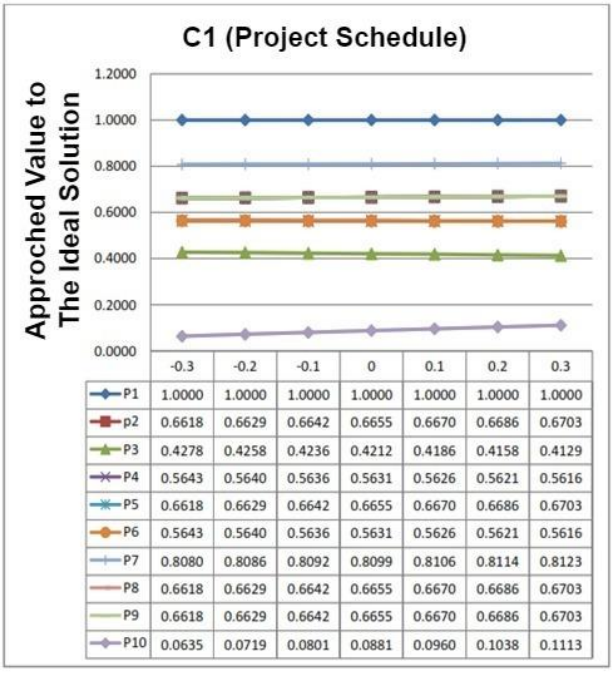
λmax = 3.0092 (6)

So that the value of the Consistency Index

(CI) = (3.0092-3) / (3-1) = 0.0066 (7)

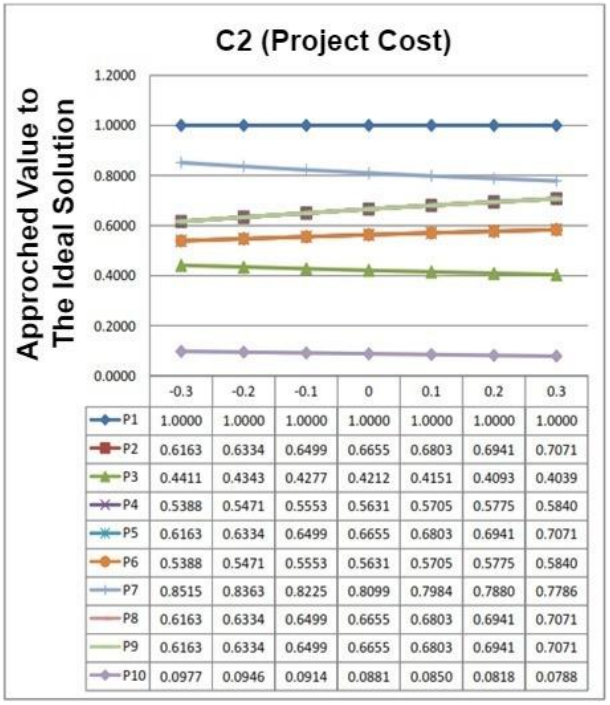
The matrix has 3 (three) elements, so the IR value is 0.58. So the value of CR is 0.0066/0.58=.0.0079. Because the CR value is less than 10%, the data judgment is correct. It can be concluded that matrix A is quite consistent.

From the values that have been tested in this study, sensitivity analysis was carried out by changing the weight of the criteria. Changes in the weight value of each criterion are carried out by lowering or increasing the weight to see whether the alternative ranking results (projects) tend to change or not. The trial weight changes were increased or decreased from the initial values of 10%, 20%, and 30%. Sensitivity analysis was performed against all criteria. Sensitivity analysis on criteria C1 (Project Schedule) Based on the graph of the results of sensitivity analysis against criteria C1 (Project Schedule) as shown in Fig. 13, test for changes in the weight of the criteria there is a change in the value of the alternative (project schedule).



	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
P1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
p2	0.6618	0.6629	0.6642	0.6655	0.6670	0.6686	0.6703
P3	0.4278	0.4258	0.4236	0.4212	0.4186	0.4158	0.4129
P4	0.5643	0.5640	0.5636	0.5631	0.5626	0.5621	0.5616
P5	0.6618	0.6629	0.6642	0.6655	0.6670	0.6686	0.6703
P6	0.5643	0.5640	0.5636	0.5631	0.5626	0.5621	0.5616
P7	0.8080	0.8086	0.8092	0.8099	0.8106	0.8114	0.8123
P8	0.6618	0.6629	0.6642	0.6655	0.6670	0.6686	0.6703
P9	0.6618	0.6629	0.6642	0.6655	0.6670	0.6686	0.6703
P10	0.0635	0.0719	0.0801	0.0881	0.0960	0.1038	0.1113

Fig. 13. Sensitivity Analysis on Criteria C1 (Project Schedule).

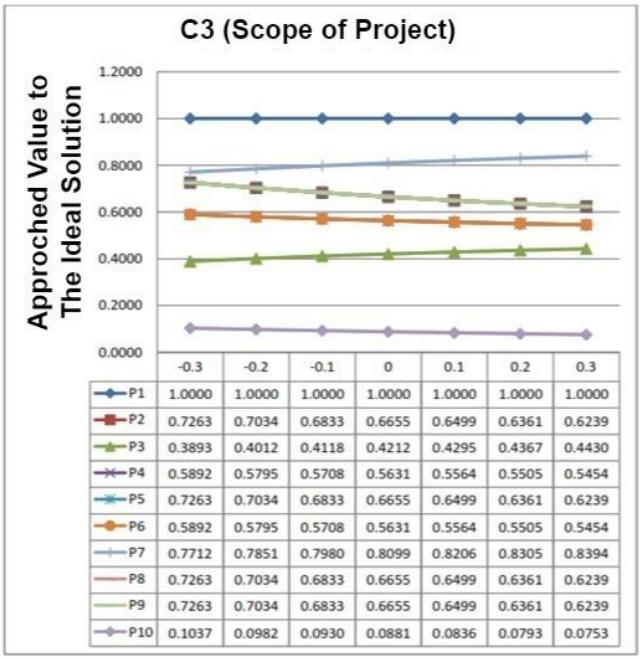


	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
P1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
P2	0.6163	0.6334	0.6499	0.6655	0.6803	0.6941	0.7071
P3	0.4411	0.4343	0.4277	0.4212	0.4151	0.4093	0.4039
P4	0.5388	0.5471	0.5553	0.5631	0.5705	0.5775	0.5840
P5	0.6163	0.6334	0.6499	0.6655	0.6803	0.6941	0.7071
P6	0.5388	0.5471	0.5553	0.5631	0.5705	0.5775	0.5840
P7	0.8515	0.8363	0.8225	0.8099	0.7984	0.7880	0.7786
P8	0.6163	0.6334	0.6499	0.6655	0.6803	0.6941	0.7071
P9	0.6163	0.6334	0.6499	0.6655	0.6803	0.6941	0.7071
P10	0.0977	0.0946	0.0914	0.0881	0.0850	0.0818	0.0788

Fig. 14. Sensitivity Analysis on Criteria C2 (Project Cost).

Based on Fig. 14, the graph below shows the results of the sensitivity analysis against the C2 (Project Cost) criteria. The test to changes in the weight of the criteria results in a change in the value of the alternative (project cost).

The results of the sensitivity analysis against the C3 criteria (Scope of the Project) as shown in Fig. 15, the test of changes in the weight of the criteria changes in the value of the alternative (project).



	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
P1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
P2	0.7263	0.7034	0.6833	0.6655	0.6499	0.6361	0.6239
P3	0.3893	0.4012	0.4118	0.4212	0.4295	0.4367	0.4430
P4	0.5892	0.5795	0.5708	0.5631	0.5564	0.5505	0.5454
P5	0.7263	0.7034	0.6833	0.6655	0.6499	0.6361	0.6239
P6	0.5892	0.5795	0.5708	0.5631	0.5564	0.5505	0.5454
P7	0.7712	0.7851	0.7980	0.8099	0.8206	0.8305	0.8394
P8	0.7263	0.7034	0.6833	0.6655	0.6499	0.6361	0.6239
P9	0.7263	0.7034	0.6833	0.6655	0.6499	0.6361	0.6239
P10	0.1037	0.0982	0.0930	0.0881	0.0836	0.0793	0.0753

Fig. 15. Sensitivity Analysis on Criteria C3 (Scope of the Project).

VI. Conclusion

This study resulted in a GDM model for evaluating ICT activities in local government agencies in Indonesia using the GDSS concept. DM involved consists of 1) Executive government institutions, 2) ICT Management Unit, and 3) Business Process Owner Unit, and the community represented by DPRD. From the values that have been tested in this study, sensitivity analysis was carried out by changing the weight of the criteria. Changes in the weight value of each criterion are carried out by lowering or increasing the weight to see whether the alternative ranking results (projects) tend to change or not. The trial weight changes were increased or decreased from the initial values of 10%, 20%, and 30%. Sensitivity analysis was performed against all criteria.

Determination of the best ICT project from several alternatives using several determining qualitative and quantitative parameters and criteria. In the project output parameters, the criteria used in the assessment are project schedule, project cost, project scope, project risk, and project performance. While the outcome parameters are projected effectiveness criteria and project user satisfaction. With the guidance of the Project Management Body of Knowledge (PMBOK), the provision of qualitative data scoring projects has an assessment basis to reduce the subjectivity of DM.

At the main stage of this study, we concluded that the method succeeded in parsing the draw voting process by developing it by adding a winning gap (distance) between alternatives during the pairwise contest and then multiplying the existing gap by the weight or population of DM. the results of the model trial case study, the voting results of all DMs on ten (10) ICT projects in the Palembang city government, it was found that and also the results of the sensitivity test against the three criteria affected project evaluation. In the future, we will try a ranking system to get results that can determine transparent winners in a project in government.

Acknowledgment

Thanks to Indo Global Mandiri Foundation (Yayasan Indo Global Mandiri, IGM) and Informatics Engineering Faculty of Computer Science, Indo Global Mandiri University.

References

[1] Lembaga Administrasi Negara Republik Indonesia, "Pedoman Penyusunan Pelaporan Akuntabilias Kinerja Instansi Pemerintah," 2003.

[2] H. Setiawan, J. E. Istiyanto, R. Wardoyo, and P. Santoso, "The use of KPI in group decision support model of ICT projects performance evaluation," Int. Conf. Electr. Eng. Comput. Sci. Informatics, vol. 2, no. August, pp. 233–237, 2015.

[3] Urena, Raquel, et al. A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. Information Sciences, 2019, 478: 461-475.

[4] Koksalmis, Emrah; KABAK, Özgür. Deriving decision makers' weights in group decision making: An overview of objective methods. Information Fusion, 2019, 49: 146-160.

[5] C. L. Hwang and M.-J. Lin, Group Decision Making under Multiple Criteria: Method and Applications, vol. 281, no. 0. Springer-Verlag, 1987.

[6] D. Turban, E; Sharda, R; Delen, Decision Support Systems, and Intelligent Systems. Boston: Prentice Hall, 2011.

[7] T. Bakshi, A. Sinharay, and B. Sarkar, "Exploratory Analysis of Project Selection through MCDM," in ICOQM-10, 2011, pp. 128–133.

[8] S. M. Kazemi, S. M. M. Kazemi, and M. Bahri, "Six Sigma project selections by using a Multi Criteria Decision making approach: a Case study in Poly Acryl Corp.," in Proceedings of the 41st International Conference on Computers & Industrial Engineering, 2011, pp. 502–507.

[9] H. Ismaili, "Multi-Criteria Decision Support for Strategic Program Prioritization at Defence Research and Development Canada," University of Ottawa, 2013.

[10] E. W. N. Bernroider, N. Obwegeser, and V. Stix, "Dissemination and impact of multi-criteria decision support methods for IT project evaluation," Proc. Annu. Hawaii Int. Conf. Syst. Sci., pp. 1103–1112, 2014.

[11] J. Żak and M. Kruszyński, "Application of AHP and ELECTRE III/IV Methods to Multiple Level, Multiple Criteria Evaluation of Urban Transportation Projects," Transp. Res. Procedia, vol. 10, no. July, pp. 820–830, 2015.

[12] A. Rabbani, M. Zamani, A. Yazdani-Chamzini, and E. K. Zavadskas, "Proposing a new integrated model based on sustainability balanced scorecard (SBSC) and MCDM approaches by using linguistic variables for the performance evaluation of oil producing companies," Expert Syst. Appl., vol. 41, no. 16, pp. 7316–7327, 2014.

[13] H. S. Shih, H. J. Shyur, and E. S. Lee, "An extension of TOPSIS for group decision making," Math. Comput. Model, vol. 45, no. 7–8, pp. 801–813, 2007.

[14] R. McLeod and G. P. Schell, Management Information System, 10rd ed. New Jersey: Pearson Prentice Hall, 2007.

[15] R. Linzalone and G. Schiuma, "A review of program and project evaluation models," Meas. Bus. Excell., vol. 19, no. 3, pp. 90–99, 2015.

[16] T. Bakshi, A. Sinharay, B. Sarkar, and S. K. Sanyal, "MCDM based project selection by F-AHP & VIKOR," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7076 LNCS, no. PART 1, pp. 381–388, 2011.

[17] G. Büyüközkan and A. Görener, "Evaluation of product development partners using an integrated AHP-VIKOR model," Kybernetes, vol. 44, no. 2, pp. 220–237, 2015.

[18] Project Management Institute, Project Management Institute, 2008. Guide To The Project Management Body Of Knowledge (PMBOK ® GUIDE) Fourth. 1384.

[19] R. E. Indrajit, "PMBOK sebagai Konsep Best Practice," vol. 37, no. C, pp. 1–37, 2013.

[20] B. Gavish and J. H. Gerdes, "Voting mechanisms and their implications in a GDSS environment," Ann. Oper. Res., vol. 71, pp. 41–74, 1997.

[21] H. Setiawan, J. Eko, R. Wardoyo, and P. Santoso, "The Group Decision Support System to Evaluate the ICT Project Performance Using the Hybrid Method of AHP, TOPSIS and Copeland Score," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 4, pp. 334–341, 2016.

[22] Azmi, Meri, et al. GDSS Prototype Model for Supplier Selection at MDM Cooperative. JOIV: International Journal on Informatics Visualization, 2021, 5.1: 16-21.

M-SVR Model for a Serious Game Evaluation Tool

Kamal Omari[1], El Houssine Labriji[4], Ali Labriji[5]
Department of Mathematics & Computer Science, Faculty of Sciences Ben M'sik University Hassan II, Casablanca, Morocco

Said Harchi[2]
Laboratory of Innovation in Management and Engineering for the Enterprise Higher Institute of Engineering and Business. Rabat, Morocco

Mohamed Moussetad[3]
Department of Physics Faculty of Sciences Ben M'sik University, Hassan II Casablanca, Morocco

***Abstract*— Today, due to their interactive, participatory and entertaining nature, the Serious Games set themselves apart from other learning methods used in teaching. Much progress has been made in the design techniques and methods of Serious Games, but little in their evaluation. In order to fill this gap, we had proposed in our previous work an evaluation tool capable of helping practitioners to evaluate Serious Games in different training contexts. This evaluation tool for Serious Games is designed around four dimensions, namely the pedagogical, technological, ludic and behavioral dimensions, which are measured by clearly defined criteria. During this process, it was highlighted that the human factor (evaluator) influences considerably the result of the weightings through the choice to weight the evaluation dimensions of the Serious Games. In order to reduce this influence during the evaluation process and to keep the correlation between the variables of our evaluation system, we present in this paper, an improvement of our evaluation tool by equipping it with an intelligent supervised self-learning algorithm allowing self-regulation of the weights according to the context of use of the Serious Game to be evaluated. Thanks to the experimental verification of the optimization results, the root mean square error and the coefficient of determination are 0.016 and 98.59 percent respectively, indicating that the model has high precision which guaranteed better predictive performance. A comparison was made between this intelligent model and the models presented in our previous work; the results obtained indicated the same order of the four dimensions, and this by reducing the influence of the human factor during the Multi-Output Support Vector Regression weighting process.**

Keywords*—Serious game; evaluation tool; multi-output support vector regression

I. INTRODUCTION

Serious Games are increasingly present in any innovative educational strategy aimed at effective and motivating learning [1]. However, before endorsing a serious game in any training, it is essential to evaluate it in addition to evaluating its impact [2], [3].

In [4], we have presented a serious game evaluation tool based on four dimensions, {Pedagogical (P), Technological (T), Ludic (L) and Behavioural (B)}, that a serious game must satisfy in order to perform the task for which it was designed. In addition, given that the importance of one dimension compared to another depends on the context in which the serious game is used, the fuzzy multi-criteria decision making methods fuzzy AHP, fuzzy TOPSIS, and fuzzy ELECTRE have been used [5], to validate the choice of the weighting of these dimensions.

In conclusion of this work, it is demonstrated that the human factor has a significant influence on the result of weightings when choosing weights for the serious game evaluation dimensions. This is done by favouring, during the weighting process, a higher value of one dimension over another.

Indeed, depending on the context of use of the serious game to be evaluated, the evaluator favours one or more dimensions deemed to be more important. For example, in a Context where the order of the dimensions is as follows P>T>L>B, the evaluator can assign, to the dimensions, all possible percentages satisfying the order of the chosen context of use and this while ensuring that their sum equals 1.

And so, to reduce this human influence in the serious game evaluation process through the choice of weightings of the evaluation dimensions, we present in this paper an improvement of our evaluation system, by endowing it with an automatic algorithm intelligent supervised learning, allowing self-regulation of the weightings according to the context of use of the serious game to be evaluated.

Thus, the Multi-Output Support Vector Regression (M-SVR) will analyze the context data of use of the serious game to allow the evaluation system to build its reasoning system without having to impose a program beforehand. In this learning phase, the algorithm is based on several examples of data to find the existing patterns in the data allowing it to build a model that will be evaluated subsequently in order to estimate its general predictive accuracy for future data.

In this paper, we advocate the use of an intelligent supervised machine learning algorithm to self-regulate weights according to the serious game context of use, to minimize this human influence in the serious game evaluation process.

This paper is divided into four sections. In Section 1, the problem, the formalization and the postulated hypotheses are presented. The description of the proposed model is illustrated in Section 2. Section 3 presents the modeling process steps carried out, together with the obtained results. In Section 4, a general conclusion is provided.

II. Problem - Formulation and Hypotheses

As mentioned above, this work is based on the findings made in our previous work that can be summarized in the following points.

A. Influence of the Human Factor (the Evaluator) in the Choice of Weightings of the Serious Game Evaluation Dimensions

As shown in Fig. 1, the evaluator participates in the serious game evaluation process by choosing the ordering of the evaluation dimensions according to the context of use of the serious game to be evaluated.

For example, if the serious game is used in a purely formative context, the evaluator will consider the pedagogical dimension as dominant over the other dimensions.

However, if the evaluator errs in his judgment of the adequacy between these choices of ordering of evaluation dimensions and the context of use of the serious game, then the entire evaluation system will be biased.

B. Existence of Correlation between the Variables of our Evaluation System

By analysing our evaluation system, we note the existence of dependency relations between its variables. Thus, we note that there is a relationship between the weighting variables expressed by their sum, which must be equal to 1. Likewise, there is a direct relationship between the evaluation dimensions' variables (P, T, L, B) and the serious game context of use. Fig. 2 presents four main cases of serious game context of use classified according to their main dominant factor: Pedagogy (Fig. 2(a)), Technology (Fig. 2(b)), Ludic (Fig. 2(c)), and Behaviour (Fig. 2(d)).

In each major type of the serious game context of use and according to the ordering of the other dominated factors, we obtain a finite number of serious game contexts of use (branches of the tree).

We also note that in each tree branch, we have an infinite number of possible serious game contexts of use. For example, take the serious game context of use (P > T> L> B) (Fig. 2(a)) where each branch (P > T) or (T > L) or (L > B) will be interpreted, as a weighting value (y_1(P) > y_2(T)), (y_2(T) > y_3(L)) and (y_3(L) > y_4(B)) respectively. With (y_1, y_2, y_3, y_4) varying between 0 and 1 that generates an infinite number of possibilities.

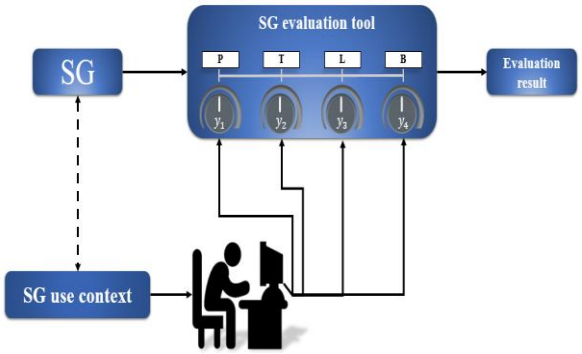


Fig. 1. Serious Game Evaluation Process.

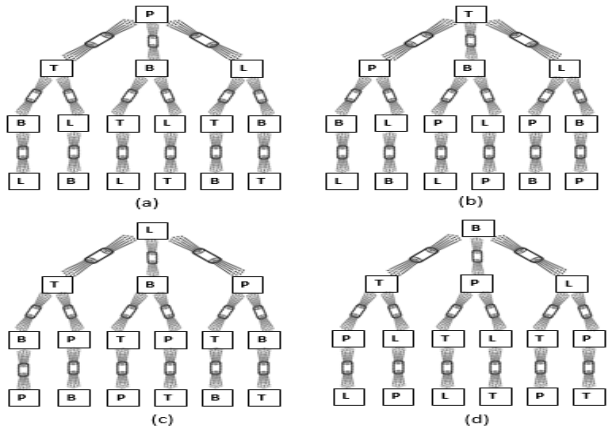


Fig. 2. Possible Serious Game use Contexts.

The Serious Game context of use is represented by the vector $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$. It is assumed that the serious game type to be evaluated corresponds to its context of use. The number of elements of this vector is deduced from the pair-to-pair comparison of the four dimensions of the evaluations with $(x_i) \in]0, 9]$ Table I.

TABLE I. Example of Representation of Context of Serious Game use

Dimensions Pairwise comparison	Vector X
(P, T)	x_1 =5
(P, B)	x_2 =7
(P, L)	x_3 =8
(T, B)	x_4 =3
(T, L)	x_5 =4
(L, B)	x_6 =0.5

Likewise, the weighting vector represented by $Y = \{y_1, y_2, y_3, y_4\}$, respectively determines the weighting of each dimension {P, T, L, B}. With $(y_i) \in]0, 1]$ is considered as a continuous dependent variable.

III. State of the Art

In order to minimize the evaluator influence during the choice of dimension weights in serious game evaluation process and keep the correlation between the variables of our evaluation system, we believe that it is wise to use the supervised machine learning algorithms power to self-regulate the weightings according to the serious game context of use to be evaluated. Among the machine learning algorithms that can meet our need, there are multi-output regression algorithms [6]. These algorithms, using a single model, aim to simultaneously predict several continuous variables when a common collection of input variables is given [7]. This takes into account not only the underlying relationships between the input variables and the corresponding targets, but also the relationships between the correlated targets [8]. This guarantees better predictive performance [16].

This algorithm type has received a lot of attention from the machine learning science community. It has already proven itself in a wide variety of real life applications [9] such as

health [10], [13], wind speed [11], heating load in buildings. Energy efficiency [12], natural language processing [14] and bioinformatics [15].

Among the multi-output regression algorithms, we opted for the multi-output support vector regression (M-SVR) algorithm proposed by Pérez-Cruz et al. [19]. This choice was dictated by:

- The infinite number of input vectors of our system, which represent the evaluation of serious game context of use and the number of output vectors, which represent the weighting, values of the dimensions of evaluation.
- Its ability to predict with high certainty multiple correlated outputs as shown in [24], [25].

IV. Proposed Model

The model proposed in this paper is an intelligent evaluator of serious game that can be adapted to any type of serious game depending on its context of use. As shown in Fig. 3, our evaluator system will be endowed with an intelligent supervised self-learning algorithm.

Therefore, the M-SVR will make it possible to simultaneously self-regulate the weightings $\{y_1, y_2, y_3, y_4\}$ of each dimension according to the serious game context of use to be evaluated $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, by capturing all existing dependencies and internal relationships, in order to give better performance (Fig. 4).

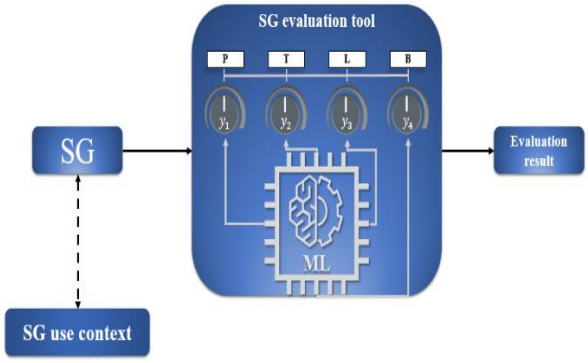


Fig. 3. Intelligent System of Weightings of the Serious Game Evaluation Dimensions.

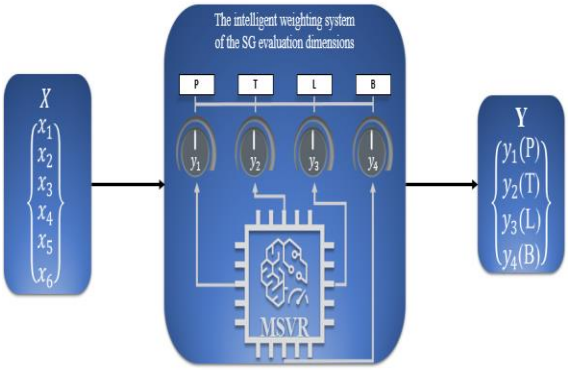


Fig. 4. M-SVR as Intelligent System Implemented in an Evaluator Tool.

A. Algorithm Description

As presented in [19], the M-SVR goal is to find the regressor w^j and b^j (j = 1,…., m) for each output which minimizes the following function (1):

$$\min_{W,b} L_p = \frac{1}{2}\sum_{j=1}^{m}\left\|w^j\right\|^2 + C\sum_{i=1}^{N} L(u_i) \quad (1)$$

Where:

$$u_i = \|e_i\| = \sqrt{e_i^T e_i}\,, e_i^T = y_i^T - \varphi(x_i)^T W - b^T$$

$W = [w^1, \dots, w^m]$: Vector of coefficients of the multiple outputs,

$b = [b^1, \dots, b^m]^T$: The constant vector representing the bias of each output,

C : The regularization parameter that balances the complexity of the model and the approximation precision,

$\varphi(.)$: Denotes a nonlinear mapping of n-dimensional input space to m-dimensional feature space, $\mathbb{R}^n \to \mathbb{R}^m$.

$L(u)$ is an ε-insensitive quadratic cost function, defined by the following equation:

$$L(u) = \begin{cases} 0, u < \varepsilon \\ u^2 - 2u\varepsilon + \varepsilon^2\,, u \geq \varepsilon \end{cases} \quad (2)$$

When ε = 0, in equation (2), this problem boils down to an independent regularized kernel least squares regression for each component.

For ε ≠ 0, it will take into account all the outputs to build the regressors for each individual, to then produce a single support vector for all the dimensions, in order to obtain more robust predictions. To solve equation (1), an iterative method called iteratively re-weighted least squares (IRWLS) [21] was used in [20], [22].

V. Steps Carried Out in the Modelling Process

The machine learning goal, implemented in our evaluation system, is to allow the multi-output regression algorithm (M-SVR) to learn a correspondence between the input vector (X) $X^{(i)} = \left(x_1^{(i)}, \dots., x_n^{(i)}\right)$ and the output vector (Y) $Y^{(i)} = \left(y_1^{(i)}, \dots, y_m^{(i)}\right)$ from the training data set (D) $D = \left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{i=1}^{N} \subset \mathbb{R}^n \times \mathbb{R}^m$ for N samples. Therefore, find a function h, which relates the input vector X to the output vector Y, $h(X) = Y$.

And so, for a given new input vector $\hat{X}$, the model will be able to predict an output vector $\hat{Y}$. $\hat{Y}$ = h ($\hat{X}$) which best approximates the real output vector Y.

Fig. 5 shows the process of machine learning algorithm like M-SVR, which consists of three main steps:

In our experiment, we used the implementation of the M-SVR algorithm with the programming language Python, using the machine-learning library Scikit-Learn [23]. The initial setup was done with the kernel = 'rbf', γ = 0.001, ε = 0.001 and C = 40.

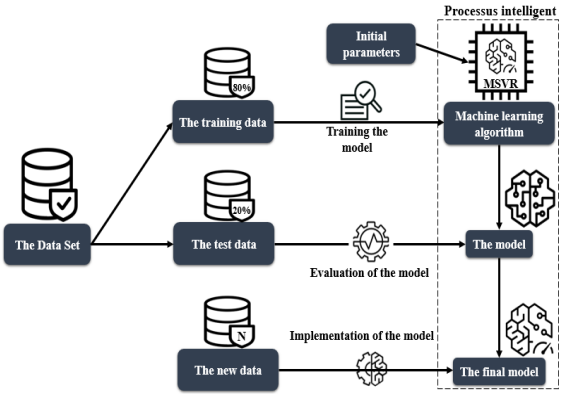


Fig. 5. Steps carried out in the Modelling Process.

A. Model Training

In the training phase of the model, we used the training data from the set (D). $D = \{ (X^{(i)}, Y^{(i)}) \}_{i=1}^{N}$. With N = 2500 knowing that the complexity of the adjustment time is more than quadratic, which makes it difficult to adapt the M-SVR to data sets of more than 10000 samples. Table II shows an example of the pair of training vectors (X, Y) used in this step.

The input vector (X) $X^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$ and the output vector (Y) $Y^{(i)} = (y_1^{(i)}, \dots, y_m^{(i)})$ with n = 6 and m = 4.

We followed the recommendations mentioned in the scientific literature of Machines Learning, taking 80% of the learning data as model training data and 20% of the training data as model evaluation data.

TABLE II. TRAINING DATA

Input vectors: serious game contexts of use's						**Output vectors: evaluation dimensions weightings'**			
x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3	y_4
0.5	3	0.2	0.2	0.5	0.5	0.135	0.133	0.154	0.578
7	1	7	3	5	3	0.508	0.234	0.196	0.062
5	1	3	7	9	5	0.387	0.371	0.182	0.06
3	5	5	0.142	7	7	0.514	0.134	0.307	0.045
7	3	0.125	5	3	5	0.307	0.293	0.189	0.212
7	3	3	5	7	0.333	0.527	0.279	0.076	0.118
0.333	5	0.125	3	0.142	1	0.138	0.214	0.104	0.544
0.2	1	0.2	1	0.2	3	0.100	0.223	0.290	0.386
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.5	3	0.2	5	0.2	0.5	0.128	0.247	0.076	0.549
3	5	7	5	5	7	0.540	0.293	0.124	0.043

B. Model Evaluation

After the training phase, the evaluation data is used to evaluate the performance of the model based on the statistical measure R^2.

This statistical measure represents the quality of the regression model adjustment. The closer the value of R^2 is to 1, the more accurate the regression model [18].

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}} \quad (3)$$

Where SS_{reg} the sum of squares is explained by the regression, SS_{tot} refers to the total sum of squares and SS_{err} is the sum of the squared error.

In addition, the root mean square error (RMSE) [17] is calculated to have the standard deviation of the errors that occur when a prediction is made on a data set. The closer the value is to 0, the less error the model produces.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \ddot{y}(i)\|^2}{N}} \quad (4)$$

Where N is the number of data items, $y(i)$ is the ith measure, and $\ddot{y}(i)$ is its corresponding prediction.

The results in Table III give a value of RMSE = 0.016 and R^2 = 98.59%. From these values, we can deduce that the accuracy of the prediction model is acceptable and, therefore, we can proceed to the exploitation of the model.

C. Model Exploitation

Once the model is trained, tested and validated, we were able to exploit it by introducing new input values $\hat{X}_i$ characterizing contexts of use of possible serious games for the prediction of the weighting values of the weighting dimensions evaluation of serious game $\hat{Y}_i$. Table IV shows an example of the operating values of the model.

TABLE III. MODEL TEST

Test input vectors						**Real output vectors**				**Model output vectors**			
x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3	y_4	$\ddot{y}_1$	$\ddot{y}_2$	$\ddot{y}_3$	$\ddot{y}_4$
7	3	0.125	5	0.125	3	0.271	0.117	0.147	0.465	0.270	0.126	0.140	0.462
7	1	7	3	5	1	0.524	0.241	0.153	0.082	0.519	0.237	0.146	0.094
3	3	7	7	7	0.333	0.492	0.357	0.065	0.087	0.444	0.409	0.070	0.077
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	3	7	9	9	0.242	0.515	0.195	0.049	0.249	0.497	0.190	0.060
9	5	0.125	3	0.142	3	0.317	0.098	0.144	0.441	0.304	0.102	0.138	0.454

TABLE IV. PREDICTION OF DIMENSION WEIGHTS

New Input Vectors						Model output vectors			
$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$	$\hat{x}_5$	$\hat{x}_6$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
3.3	3	5	3	5	3	0.500	0.283	0.163	0.056
5	3	5	3	2	3	0.543	0.181	0.160	0.119
1	5	3	5	7	3	0.350	0.483	0.107	0.060
5	3	1	3	2	3	0.351	0.180	0.179	0.291
1	3	3	7	7	3	0.295	0.527	0.116	0.060
2	1	2	7	7	6	0.271	0.471	0.192	0.063
5	3	1	3	2	7	0.345	0.198	0.242	0.217

We also tested our evaluator model with the M-SVR model using the same evaluation context proposed in [4], where the context of use of the serious game is purely educational, with an academic and scientific target audience.

The vector X = {3,5,7,3,5,3} represents the context of use and the corresponding output vector generated by the M-SVR model is Y = {0.557,0.270,0.123,0.052}. The serious game to be evaluated is "Leuco'war", Fig. 6 illustrates the results obtained which confirm those obtained in [4].

According to the results obtained, we note that the use of the M-SVR algorithm at the centre of the weighting process of the chosen dimensions has provided our serious game evaluation tool the ability to self-regulate with an acceptable precision the weights of the dimensions in different Serious Game evaluation contexts. This allowed us to respond to the observations noted during the experimental studies conducted in [4], [5]. Indeed, M-SVR, using a unique predictive model of several continuous variables, guaranteed a better predictive performance confirmed by the values of RMSE and R^2 respectively equal to 0.016 and 98.59 per cent. This by taking into consideration the underlying relationships between the context variables of use of the Serious Game and the weights of the corresponding dimensions, and their correlated relationships. Likewise, the comparison of the results obtained by this evaluation model confirms, with a notable reduction in the influence of the human factor in the process of weighting of dimensions, the same order of its last obtained in [4], [5].

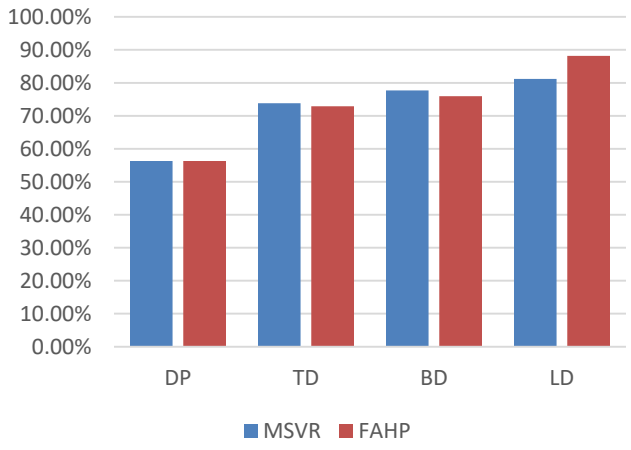


Fig. 6. Serious Game "Leuco'war" Evaluation Results.

Finally, we note that the most important added value of the use of the M-SVR multi-output supervised machine learning algorithm is its ability to adapt the weightings of the serious game evaluation dimensions to the very large number of possible contexts of use.

VI. CONCLUSION

By introducing a smart process such as M-SVR into our serious game evaluator system, we were able to reduce the subjective evaluation introduced by the human factor by having automatic weighting values according to the context of use of the chosen serious game. In addition, the values of the test parameters RMSE and R^2 equal to 0.016 and 98.59% respectively testify to an acceptable performance of the algorithm used. However, we noticed an instability of the algorithm when we used a very large dataset volume. Thus, we plan to compare the results obtained with M-SVR with another other algorithms of the same type that is more stable with respect to the volume of the data set of more than 10000 samples, because the complexity of the adjustment time is more than quadratic. Likewise, we plan to place an intelligent process in our system linking the serious game context of use and the corresponding serious game.

REFERENCES

[1] Vaz de Carvalho, Carlos & Cerar, Špela & Rugelj, Jože & Tsalapatas, Hariklia & Heidmann, Olivier. (2020). Addressing the Gender Gap in Computer Programming Through the Design and Development of Serious Games. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje. PP. 1-1. 10.1109/RITA.2020.3008127.

[2] Petri, Giani & Gresse von Wangenheim, Christiane. (2016). How to Evaluate Educational Games: a Systematic Literature Review. Journal of Universal Computer Science. 22. 992.

[3] Liu S., Ding W. (2009) An Approach to Evaluation Component Design in Building Serious Game. In: Chang M., Kuo R., Kinshuk, Chen GD., Hirose M. (eds) Learning by Playing. Game-based Education System Design and Development. Edutainment 2009. Lecture Notes in Computer Science, vol 5670. Springer, Berlin, Heidelberg.

[4] Omari, K., Moussetad, M., Labriji, E., & Harchi, S. (2020). Proposal a New Tool to Evaluate a Serious Game. *International Journal Of Emerging Technologies In Learning (IJET)*, 15(17), pp. 238-251. doi:http://dx.doi.org/10.3991/ijet.v15i17.15253.

[5] Omari, K., Harchi, S., Ouchaouka, L., Rachik, Z., Moussetad, M., & Labriji, E. (2021). Application the fuzzy topsis and fuzzy electre in the serious games evaluation tool. *Journal of Theoretical and Applied Information Technology* (JATIT), Vol. 99, No.09.

[6] Borchani, Hanen & Varando, Gherardo & Bielza, Concha & Larranaga, Pedro. (2015). A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 5. 10.1002/widm.1157.

[7] Li, Ximing & Wang, Yang & Zhang, Zhao & Hong, Richang & Zhuo, Li & Wang, Meng. (2020). RMoR-AION: Robust Multi-output Regression by Simultaneously Alleviating Input and Output Noises. IEEE Transactions on Neural Networks and Learning Systems. 10.1109/TNNLS.2020.2984635.

[8] Tuia D, Verrelst J, Alonso L, P_erez-Cruz F, Camps-Valls G. Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE Geosci. Remote Sens. Lett. 2011, 8(4):804-808.

[9] L. Lin, E. Liu, L. Wang, and M. Zhang, "Fingerprint orientation field regularisation via multi-target regression", Electronics Letters, vol. 52 no. 13, pp. 1118–1120, 2016.

[10] X. Wang, X. Zhen, Q. L, D. Shen and H. Huang, "Cognitive Assessment Prediction in alzheimer's Disease by Multi-Layer Multi Target Regression", Neuroinformatics, vol. 16, pp. 285-294, 2018.

[11] A. Appice, A. Lanza and D. Malerba, “Handling Multi-scale Data via Multi-target Learning for Wind Speed Forecasting”. In: Ceci M., Japkowicz N., Liu J., Papadopoulos G., Raś Z. (eds) Foundations of Intelligent Systems. ISMIS 2018. Lecture Notes in Computer Science, vol 11177. Springer, Cham.

[12] Moayedi, Hossein & Bui, Dieu & Dounis, Anastasios & Lyu, Zongjie & Foong, Loke. (2019). Predicting Heating Load in Energy-Efficient Buildings Through Machine Learning Techniques. Applied Sciences. 9. 4338. 10.3390/app9204338.

[13] Gerdes, Henry & Casado, Pedro & Dokal, Arran & Hijazi, Maruan & Akhtar, Nosheen & Osuntola, Ruth & Rajeeve, Vinothini & Fitzgibbon, Jude & Travers, Jon & Britton, David & Khorsandi, Shirin & R. Cutillas, Pedro. (2021). Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. Nature Communications. 12. 10.1038/s41467-021-22170-8.

[14] Garg, Ravi & Oh, Elissa & Naidech, Andrew & Kording, Konrad & Prabhakaran, Shyam. (2019). Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. Journal of Stroke and Cerebrovascular Diseases. 28. 10.1016/j.jstrokecerebrovasdis.2019.02.004.

[15] Bhargava, Harshita & Sharma, Amita & Valadi, Jayaraman. (2021). Machine Learning for Bioinformatics. 10.1007/978-981-15-9544-8_11.

[16] Kocev, Dragi & Džeroski, Sašo & White, Matt & Newell, Graeme & Griffioen, Peter. (2009). Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling - ECOL MODEL. 220. 1159-1168. 10.1016/j.ecolmodel.2009.01.037.

[17] Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE) arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250.

[18] Windmeijer, Frank & Cameron, A.. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics. 77. 329-342. 10.1016/S0304-4076(96)01818-0.

[19] Pérez-Cruz, Fernando & Camps-Valls, Gustau & Olivas, Emilio & Perez-Ruixo, Juan & Figueiras-Vidal, Aníbal & Artés Rodríguez, Antonio. (2002). Multi-dimensional Function Approximation and Regression Estimation. Lecture Notes in Computer Science - LNCS. 757-762. 10.1007/3-540-46084-5_123.

[20] Mao WT et al (2014a) A fast and robust model selection algorithm for multi-input multi-output support vector machine. Neurocomputing 130:10–19.

[21] F. Pérez-Cruz, A. Navia-Vázquez, P. L. Alarcón-Diana, and A. Artés-Rodríguez, “An IRWLS procedure for SVR,” in Proc. EUSIPCO, Tampere, Finland, Sept. 2000.

[22] Sánchez-Fernández, Matilde & De-Prado-Cumplido, Mario & Arenas-Garcia, Jerónimo & Pérez-Cruz, Fernando. (2004). SVM Multiregression for Nonlinear Channel Estimation in Multiple-Input Multiple-Output Systems. Signal Processing, IEEE Transactions on. 52. 2298 - 2307. 10.1109/TSP.2004.831028.

[23] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

[24] Mao, Wentao & Tian, M & Yan, G. (2012). Research of load identification based on multiple-input multiple-output SVM model selection. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 226. 1395-1409. 10.1177/0954406211423454.

[25] Zhao, Wei & Liu, J.K. & Chen, Y.Y.. (2015). Material Behavior Modeling with Multi-Output Support Vector Regression. Applied Mathematical Modelling. 39. 10.1016/j.apm.2015.03.036.

Modified Method of Traffic Engineering in DCN with a Ramified Topology

As'ad Mahmoud As'ad Alnaser[1]
Department of Applied Science
Ajloun University College
Al-Balqa Applied University, Ajloun, Jordan

Yurii Kulakov[2], Dmytro Korenko[3]
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," 37 Peremohy ave.
03056 Kyiv, Ukraine

***Abstract*—This article consider two main local network topologies. Based on the basic DFS protocol, a mathematical model has been developed for a new method of multipath routing and traffic engineering in data centers with a ramified topology. This method was developed with the features and benefits of SDN in mind. Also, the simulation of the developed method was carried out on two local topologies considered earlier.**

***Keywords*—Local networks; traffic engineering; SDN; DCN; DFS; Mininet**

I. Introduction

The successful operation of the enterprise in a dynamic market and competition is largely determined by the ability to make quick decisions. Timely and high-quality decision-making is possible only if the reliable and productive operation of the company's IT infrastructure. In turn, modern IT systems are evolving dynamically: there are more and more different business applications and automated workplaces, increasing the amount of information, requires continuous provision of IT services.

The basis of building the company's IT infrastructure is the Data Center (DC). The main task of the DC is to ensure guaranteed trouble-free operation of the enterprise's IT infrastructure. In this case, we are talking not only about the automation of business processes, but also about the most reliable data storage and guaranteed constant access to them. Databases can be upgraded by adding computing resources when implementing new business applications, as well as increase the amount of data storage, which means the ability to quickly adapt to changing business requirements.

First of all, data centers are in demand by large organizations, such as banks, insurance and trading corporations, mining companies, telecommunications companies (billing systems, hosting, various Web-services and social services). They all use complex business applications, and their activities depend on the reliability of the IT infrastructure.

The most important advantages of creating a DC are the consolidation of computing power and storage systems. As shown in [1], It is known that centralized management of IT infrastructure and information systems is more efficient than in the case of a distributed heterogeneous solution. In addition, it is easier to provide surveillance of a single complex and protection against possible failures.

At the same time, the growing needs of business must be met in conditions of limited resources: more data must be stored in the allocated space, this will prevent consumption or allocation of too much energy so calculations will be faster, to transmit more information through existing channels of communication, to ensure maximum readiness operated IT systems. And all this with limited funding. In such conditions, the competent design of the DC is a key link to achieve efficient operation of the enterprise, and these limitations determine the choice of technologies and equipment used.

It is known that in the field of high-availability IT systems, the situation is constantly and rapidly changing. The lack of a unified approach to the organization of DC, standards for the design and operation of various processing centers and server rooms pose the problem of developing a systematic approach to infrastructure implementation, as well as developing methods and models of DC to a new level. The emergence of experience in the form of standards, models and methods will unify the implementation and simplify changes in its infrastructure, thereby contributing to the replication and scalability of solutions.

In the absence of such experience, it is easy to be tempted and prefer the latest developments. Obviously, it is a rational choice, justified from the economic and technological point of view of the infrastructure that would provide long-term investment protection and allow the company to perform current tasks and develop. The amount of information in the future will only grow. Further growth of such sectors of the Internet related to business applications as e-commerce, payments, communications require an appropriate infrastructure. It is necessary to make sure that in the future there is an opportunity to ensure cost-effective growth and expansion of diabetes, because disruptions in electronic services can have significant economic consequences, both for individual enterprises and for government agencies and sectors of the economy.

The above indicates the relevance of the study of ways to effectively design traffic in networked data centers with a branched topology, taking into account open standards, which helps to minimize the problems of interaction when scaling the DC [2].

Improvement of quality of service (QoS) of traffic and reduction of it's design time can be hold by using multipath routing in SDN for centralized formation of multiple paths [3],

unlike known methods that have high time complexity. So a modified method was proposed in which formation of multiple paths has less complexity compared to known methods.

This paper considers the modeling proposed in article [4] mathematical model of a modernized method of constructing traffic in networked data centers with a branched topology, focused on software configuration of the network. This method, by taking into account the peculiarities of the organization of SDN, in particular due to the presence in the network of a central controller, reduces the time of formation of many routes of access to network resources [5], [6], [7], [8] and [9].

II. Algorithm Modification

In terms of routing, we are usually only interested in optimal routes. This applies to both one-path routing (search for the shortest route) and multi-path (search for a set of non-ordinary, partially ordinary routes). However, applying this method to this model is ineffective, because the controller requires a rapid response to changes in the state of communication channels, which leads to changes in the metric. Therefore, it is unwise to list the paths to find the shortest path.

For example, consider dynamic routing protocols such as OSPF and IS-IS [10]. These protocols are similar and designed for dynamic routing and take into account the state of the channels when constructing routes. These protocols use the Dijkstra algorithm to find the shortest route. In time, the complexity of the algorithm is $O(n^2)$. From this we can conclude that as the number of nodes increases, the time of finding the shortest path increases quadratically. With frequent updating of the status of communication channels, this approach is ineffective.

In the proposed method, the network topology rarely changes, therefore, it is advisable to consider a one-time calculation of all possible paths and the calculation of the next route to form new connections. Therefore, if it is necessary to find the optimal route, it is not necessary to calculate the route using the adjacency matrix; it is enough to list metrics and to choose an optimum line on the basis of the available information on a condition of knots and channels. This method can be used for different types of metrics. In the model used, the speed of information transmission along the route is used as a metric, so the controller will choose the path at high speed and send it to the first vertex along the route to the router [11].

Using the algorithm of passage of the graph in width makes it possible to identify all possible paths between the vertices. Ideally, this method will be applied once for each pair of vertices, while lists with calculated routes will be stored in the controller.

The method of constructing a route is based on the use of recursive steps. As a result, a set of all possible routes from the final to the initial vertex is formed. To do this, the extreme vertex is selected in each path and, starting from it, new routes are formed, which includes the already formed sub-route from the final vertex to the current one with the inclusion of one of the neighboring vertices. The algorithm terminates for this subroute if a new neighboring vertex is already in the route or the current vertex is the last. The procedure is complete if all routes have reached the initial peak.

However, a new route will not be formed if the adjacent vertex is already on the desired path. As mentioned, this step is recursively repeated for all routes until the next peak on the way is final. In the case of the formation of new links in the network, the already formed routes are supplemented as follows:

1) Select all paths that include vertices connected by a formed channel;

2) For each selected path, a sub-route is created from the vertex that participates in the new connection;

3) The vertex, which is located on the other side of the new connection, is added to the received sub-routes.

4) Recursive steps described for the formation of all routes, starting with the corresponding sub-route, are repeated.

Consider forwarding a packet to the next node on the way to its destination.

In high-load networks, the state of communication lines is constantly changing. These changes are difficult to predict because there are a number of reasons that can cause an unexpected load on a connection that was lightly loaded and guaranteed data transfer while maintaining a high level of service quality.

The packet forwarding algorithm used (MPLS VC) calculates the path during the traffic generation phase and stores the intermediate node labels in the packet headers. This algorithm using a virtual data channel does not allow timely response to changes in the state of communication channels in the network, because the intermediate nodes on the data route are stored as labels in the packet. Therefore, it is possible that the intermediate nodes stored in the packet header will not create the optimal path for the current time. This situation is likely to lead to some delays in data transmission for both the destination node and those nodes along which the routes intersect with the considered.

The proposed modernized method is based on the traditional approach to the routing table. These tables are updated after a certain period of time, each of the paths is updated independently of the others, just when you want to use it, or the time until the next update of the current route. After a request from the router, the controller calculates indicators if the paths for the destination node are already formed, generates these paths, and then sends them to the router, which sent a request to update the path with the number of the next node on the optimal path now.

Consider an unmodified width search algorithm (DFS) [12],[13]. It is based on a recursive stroke of the graph's vertices, through the tour, visited and non-visited nodes were marked. In the start point, some nodes are visited. The visited node is the one that receives the control packet and analyze information about neighboring nodes. After that the packet is transmitted to one of the neighboring unlabeled nodes. Otherwise, the control packet will pass back, if there is no unmarked neighboring nodes related to this node.

The disadvantages of main algorithm come from the ignorance of:

1) Channels communication capacity.
2) Network features topology.

Strategies for channel selection:

- Worst .Fit: In which the highest available bandwidth channel is selected.
- First Fit: In which any available bandwidth Chanel fits the requirements is selected.
- Best Fit: In which the available bandwidth Chanel that best fit the requirements

Taking in account, when DFS method is used to create multiple paths in DCN, the most effective strategy is Worst Fit. Moreover, the efficiency of aeration of many paths using DFS method is increased in association with the peculiarities of the DCN hierarchical organization and the communication channels bandwidth between the nodes of the network [14].

The DFS method is dependent on the centralized routing method. In this method all the information that are required to generate routes are found in the central controller, which contains information about network topology and capacity of communication channels, which is used to select the communication channel at a higher level or return to a lower level in the tree. Following the next path formation the communication channel with the allowable bandwidth for the selected path is reduced by a given value.

Based on that the network topology information is found in the central controller, the first step that is determined by the algorithm is the sender and recipient topology level where the nodes are unified. The modes are connected directly, if the path have many switches to pass through, their number is determined by the nodes relative location.

If the connection level of nodes is determined, routing will be facilitated, there for eliminating the unnecessary transitions between levels.

Fig. 1 shows the sequence of operations for modified DFS algorithm:

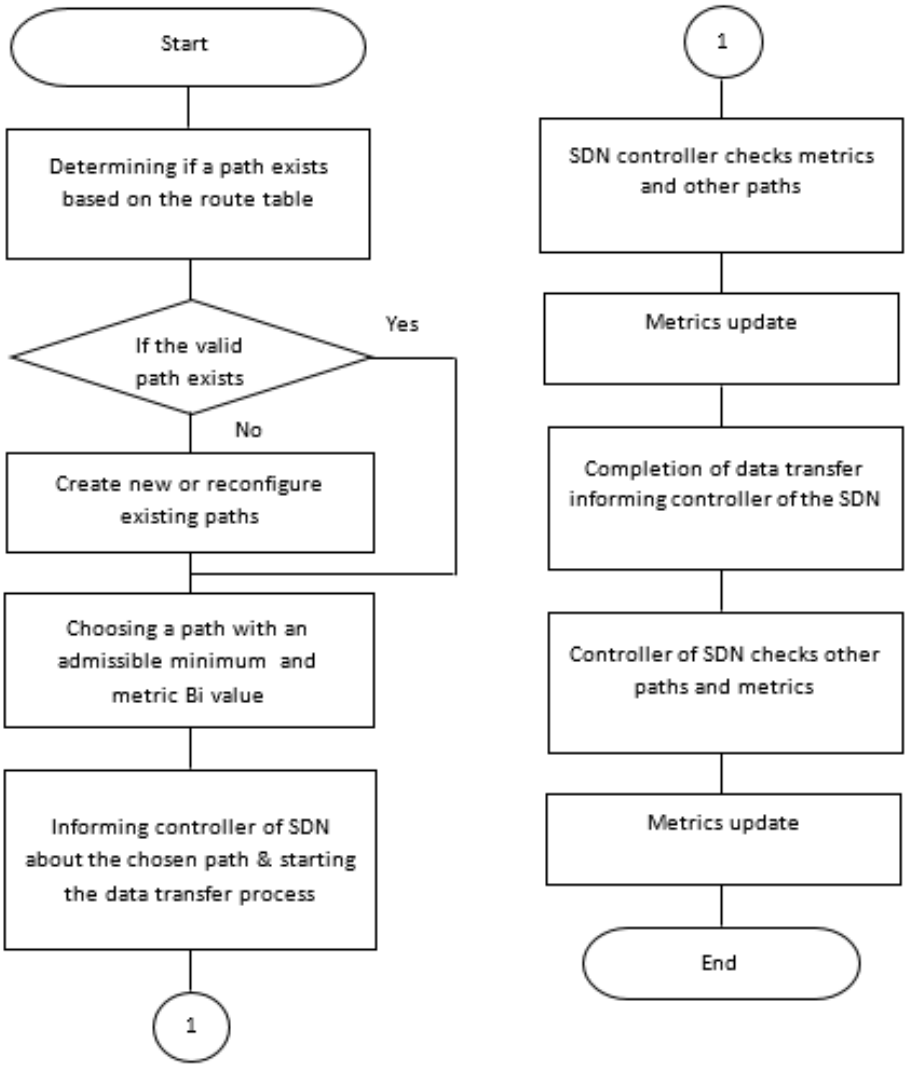


Fig. 1. Modified DFS Algorithm.

An advantage of SDN is that the network organization and management at the software level is carried out by virtual switches [15].

The Open Flow standard, [16], [17] and [18] in which the SDN concept is based, defines the network traffic management flow principles. The implementation of multi-threaded traffic routing in a software - configurable network is proposed due to the ability to configure the route of each individual traffic at the level of software-configurable SDN switches.

Streaming algorithms for multibeam routing have a minimal time complexity with known algorithms [19].

A set of continuous paths can be formed in the presence of centralized SDN control based on a network controller, by organizing counter - flows between sender and recipient nodes. Finding paths task is to find the tree's points of connections from the start and end vertices. As in [20], for a route similar to the modified wave algorithm, as the next for a certain path, a vertex with a smaller external degree is selected. So, a divergent paths formation is ensured.

Path trees are built until all the intersecting paths between the start and end vertices are built. Forming a set of disparate paths algorithm is illustrated below.

Define nodes set $X_1 = \{U_n\}$;
B_i=0;
J=0;
for $j = j + 1$ step 1, create a nodes set X_j+1 ={U_i| i=1,…n} adjacent to the set nodes of Y_i {U_n, U_l, N_i,c, b_i}, where n is the power's sum of the set of node X_{j1} ={U_i| i=1,…n};
if $X_{j+1} = \varnothing$ then go to 10 do
for i=1 step 1 to n calculate
Y_i {U_n, U_l, N_i,c, b_i}
if $b_j > B_i$ then $B_i = b_j$
end;
go to 4

end.

In the centralized formation of a set of independent paths, complete information about the formed paths and the trees that generated them is found in the central controller SDN. This gives us the chance to optimize them in the process of forming paths in accordance with the specified indicators.

III. Modeling

The Mininet environment was used to demonstrate the operation of the algorithm. Mininet is a computer network emulator. It allows you to quickly set up a network on a personal computer. This network will be almost indistinguishable from the real one, it will just not be able to send pings to external IP addresses.

To model the proposed method of traffic design, you need to build and configure the topologies considered in article [4], namely the double extended star and the double ring.

Two (2) controllers, 6 switches and 12 hosts were used to build the Double Extended Star topology (Fig. 2).

Before the simulation begins, the controllers form a model of the constructed topology. To do this, they build an adjacency matrix and store nodes and connections. As a result of the correct start of modeling of controllers on the console the list of all nodes and their connections will be deduced. The console displays basic information about the nodes and the connections with other nodes, namely the actual node name, bandwidth and channel load (delay). In this step, each connection has a bandwidth of 100Mbit, i.e. it does not contain any load. This message is displayed only once during the initial initialization, then when the topology changes, the network model is updated by the controller without messages of this type.

To simulate the operation of the algorithm, it is necessary to initiate traffic between hosts h7 and h14. To demonstrate, execute the command to compile the dump and start the traffic between hosts h7 and h14 (Fig. 3).

As you can see from the results, the routing table that was created at network startup and updated periodically, controllers immediately know the optimal route for traffic. In our case, we can see in the logs, which are collected during the transmission of the packet, a complete description of the route, and the time of passage of the packet between nodes. We also obtain the IP and MAC addresses of the sending and receiving nodes, which simplifies the analysis of the collected data and allows easy packet routing.

To demonstrate the effective operation of the developed method of traffic design for this topology, a original DFS algorithm was simulated.

As you can see in Fig. 4 the execution time of the first packet transmission takes 0.11 ms, which is 0.058 ms more than the modified one. This is because before the transmission begins, complete information is collected about the state of the nodes and communication channels in the topology, which takes some time. If the topology was static (nodes are always working, communication channels have a constant non-variable load) then a slightly longer time to send the first packet did not play an important role, but because it is not possible and the topology is dynamic (nodes fail or topology is supplemented by new, in communication channels are constantly changing the load or they break off altogether) then each time to collect new information about the state of the topology before sending the packet is quite time-consuming operation.

To obtain more detailed information about the operation of the modified method of traffic design, simulations were performed with different numbers of packets (Graph 1 and Table I).

Four (4) controllers, 8 switches and 12 hosts were used to build the Double Ring topology (Fig. 5).

In this simulation, the situation is the same as described earlier. A topology model is formed due to the adjacency matrix and information about nodes and connections between them.

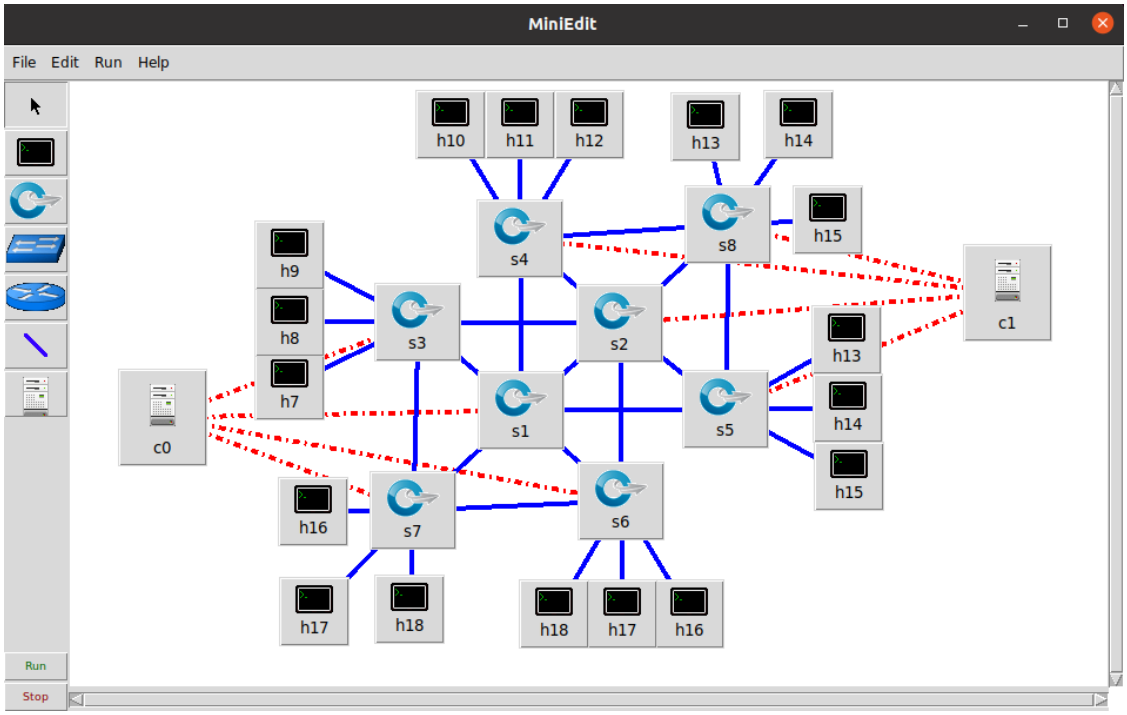


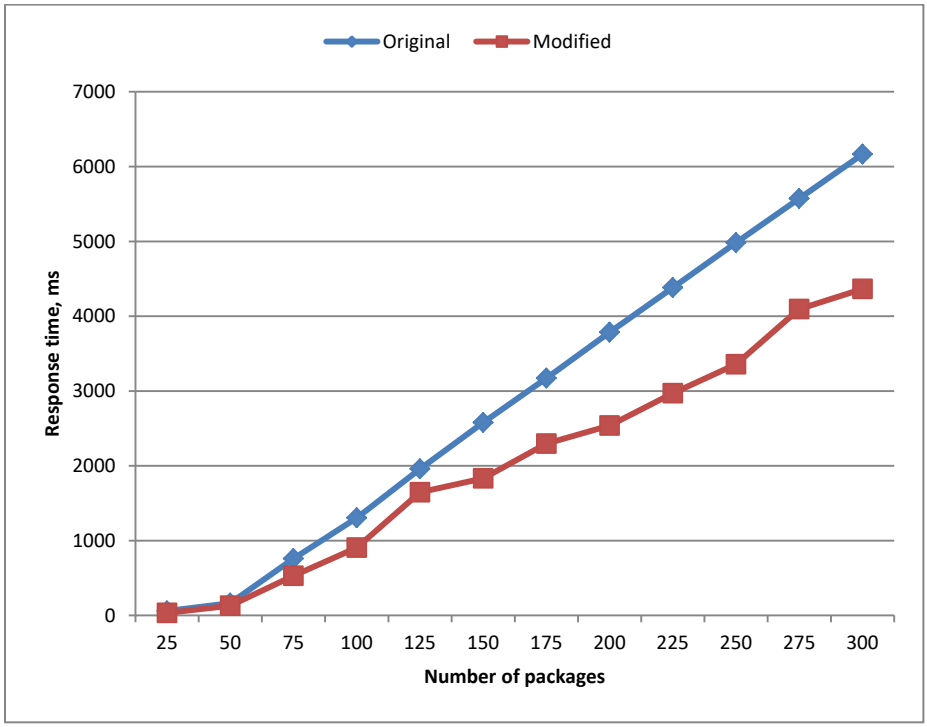
Fig. 2. Topology Double Extended Star in Mininet.

```
mininet> dpctl dump-flows
*** s1 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s2 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s3 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s4 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s5 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s6 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
mininet> h7 ping -c 2 h14
PING 127.0.3.7 (127.0.3.7) 56(84) bytes of data.
64 bytes from 127.0.3.7: icmp_seq=1 ttl=64 time=0.052 ms
64 bytes from 127.0.3.7: icmp_seq=2 ttl=64 time=0.021 ms

--- 127.0.3.7 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.021/0.026/0.032/0.005 ms
```

Fig. 3. Start Dump Collection and Start Traffic.

```
mininet> h7 ping -c 2 h14
PING 127.0.3.7 (127.0.3.7) 56(84) bytes of data.
64 bytes from 127.0.3.7: icmp_seq=1 ttl=64 time=0.11 ms
64 bytes from 127.0.3.7: icmp_seq=2 ttl=64 time=0.035 ms

--- 127.0.3.7 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.021/0.026/0.032/0.005 ms
mininet>
```

Fig. 4. DFS Modeling.



Graph 1.The Dependence of the Average Response Time on the Number of Packets.

TABLE I. SIMULATION RESULT FOR A LARGE NUMBER OF PACKAGES

Number of packages		**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**
Response time, ms	**Original algorithm**	60	165	760	1305	1958	2580	3171	3785	4382	4982	5573	6166
	Modified algorithm	35	131	531	907	1647	1833	2297	2539	2970	3358	4095	4364

Next, consider modeling a double ring topology:

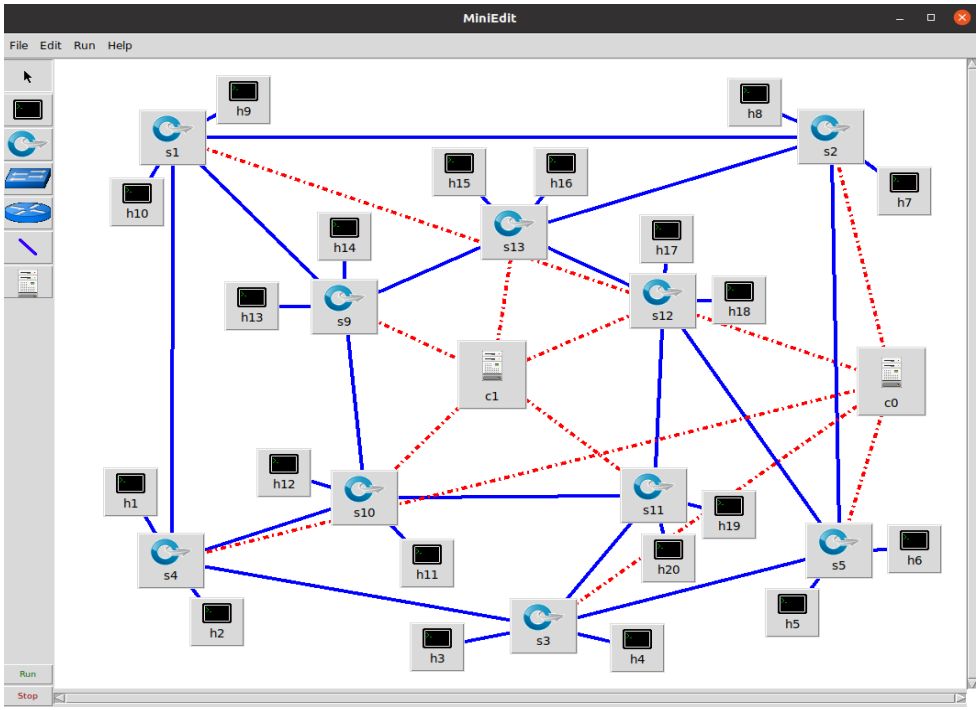


Fig. 5. Topology of a Double Ring in a Mininet.

The simulation follows the same scenario as in the previous section. We initiate traffic between nodes h7 and h14 and collect logs based on simulation results (Fig. 6).

To demonstrate the effective operation of the developed method of traffic design for this topology, an original DFS algorithm was simulated.

As you can see in Fig. 7 the execution time of the first packet transmission takes 0.087 ms, which is 0.057 ms more than the modified one.

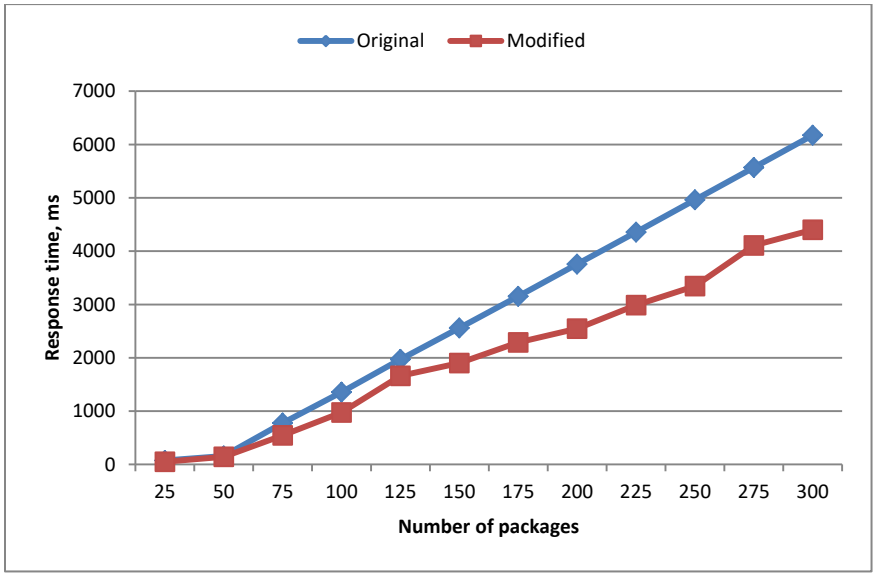
Graph 2 and Table II show the simulation results for the Double Ring topology of the original and modified method of constructing traffic for a large number of packets.

```
mininet> dpctl dump-flows
*** s9 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s12 -----------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s11 -----------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s10 -----------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s1 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s4 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s3 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
*** s2 ------------------------------------------------------------------------
NXST_FLOW reply (xid=0x4):
mininet> h7 ping -c 2 h14
PING 127.0.2.6 (127.0.2.6) 56(84) bytes of data.
64 bytes from 127.0.2.6: icmp_seq=1 ttl=64 time=0.028 ms
64 bytes from 127.0.2.6: icmp_seq=2 ttl=64 time=0.038 ms

--- 127.0.2.6 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1003ms
rtt min/avg/max/mdev = 0.028/0.033/0.038/0.005 ms
```

Fig. 6. Start Dump Collection and Start Traffic.

```
mininet> h7 ping -c 2 h14
PING 127.0.3.7 (127.0.3.7) 56(84) bytes of data.
64 bytes from 127.0.3.7: icmp_seq=1 ttl=64 time=0.087 ms
64 bytes from 127.0.3.7: icmp_seq=2 ttl=64 time=0.041 ms

--- 127.0.3.7 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1003ms
rtt min/avg/max/mdev = 0.021/0.026/0.032/0.005 ms
mininet>
```

Fig. 7. DFS Modeling.



Graph 2.The Dependence of the Average Response Time on the Number of Packets.

TABLE II. SIMULATION RESULT FOR A LARGE NUMBER OF PACKAGES

Number of packages		25	50	75	100	125	150	175	200	225	250	275	300
Response time, ms	Original algorithm	76	158	776	1357	1972	2559	3153	3757	4357	4962	5570	6175
	Modified algorithm	51	143	547	974	1663	1901	2289	2547	2989	3346	4107	4401

IV. CONCLUSION

This paper simulates the operation of a modified method of constructing traffic in networked data centers with a branched topology, which, given the peculiarities of the organization of SDN, reduces the time of forming a set of routes to access network resources and simplify the procedure for changing the route.

Topology data using this method [6] allow to virtually eliminate the delay or loss of packets in the process of traffic reconstruction. At the same time, the more paths formed in the topologies, the less likely it is that packets will be delayed or lost.

The results, based on the two most popular topologies for large businesses, showed a 30-50 percent acceleration of packet transmission between nodes due to the collection of information about communication channels by controllers and the construction of a contiguity matrix. As'ad et al [21], [22], [23] and [24] enter new concept on networks and computer science which it bipolar intuitionistic fuzzy sets and he used it in many of his articles.

REFERENCES

[1] Aguado, M. Davis, S. Peng, M.V. Ãlvarez, V. LÃpez, T. Szyrkowiec, A. Autenrieth, R. Vilalta, A. Mayoral, R. Muoz, R. Casellas, R. Martnez, N. Yoshikane, T. Tsuritani, R. Nejabati, D. Simeonidou, Dynamic virtual network reconfiguration over sdn orchestrated multitechnology optical transport domains, J. Lightwave Technol. 34 (8) (2016).

[2] Y. Han, S. Seo, J. Li, J. Hyun, J. Yoo, J. W. Hong, Software Defined Networking-based Traffic Engineering for Data Center Networks: In Asia-Pacific Network Operations and Management Symposium (2014), https://ieeexplore.ieee.org/document/6996601.

[3] E. Chemerinsky, R. Smeliansky, On QoS Management in SDN by Multipath Routing. In: Proceedings International Science and Technology Conference (Modern Networking Technologies) (MoNeTeC) (2014) https://ieeexplore.ieee.org/document/6995581.

[4] Y. Kulakov, D. Korenko, Traffic engineering in DCN with a ramified topology. 2020 6th High Performance Computing Conference, 2020.

[5] ZHAOGANG SHU, JIAFU WAN, JIAXIANG LIN, Traffic Engineering in Software-Defined Networking: Measurement and Management, IEEE Access. – 2016. – №4. – С. 3246–3256.

[6] Y. Kulakov, S. Kopychko, V. Gromova, Organization of Network Data Centres Based on Software-Defined Networking . In Proceedings International Conference on Computer Science, Engineering and Education Applications ICCSEEA 2018: pp.447-455 / https://link.springer.com/book/10.1007/978-3-319-91008-6.

[7] B. Isong, T. Kgogo, F. Lugayizi , Trust establishment in SDN: controller and applications. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) 9(7), 20–28 (2017) . http://dx.doi.org/10.5815/ijcnis.2017.07.03.

[8] K.S. Sahoo, S.K. Mishra, S. Sahoo, B. Sahoo, Software defined network: the next generation internet technology. Int. J. Wirel. Microwave Technol. (IJWMT) 7(2), 13–24 (2017). https://doi.org/10.5815/ijwmt.2017.02.02.

[9] A.M.A. Alnaser, A method of multipath routing in SDN networks , Advances in Computer Science and Engineering Volume 17, Number 1, 2018, pp 11-17.

[10] Katz, Dave (2000). OSPF and IS-IS, A Comparative Anatomy. North American Network Operators Group NANOG 19. Albuquerque. Archived from the original on June 20, 2018.

[11] P. Kumar, R. Dutta, R. Dagdi, K. Sooda, A. Naik, A programmable and managed softwaredefined network. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) 12, 11–17 (2017). https://doi.org/10.5815/ijcnis.2017.12.02. In MECS http://www.mecs-press.org/. Accessed Dec 2017.

[12] Kleinberg, Jon; Tardos, Éva (2006), Algorithm Design, Addison Wesley, pp. 92–94.

[13] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein., Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Section 22.3: Depth-first search, pp. 540–549.

[14] B. Isong, T. Kgogo, F Lugayizi, Trust establishment in SDN: controller and applications. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) 9(7), 20–28 (2017). https://doi.org/10.5815/ijcnis. 2017.07.03.

[15] M. Luo, Y. Zeng, J. Li, W. Chou, An adaptive multi-path computation framework for centrally controlled networks. Comput. Netw. 83, 30–44 (2015).

[16] M. Moza, S. Kumar, Analyzing multiple routing configuration. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) 5, 48–54 (2016). https://doi.org/10.5815/ijcnis.2016.05.07. In MECS http://www.mecs-press.org /. Accessed may 2016.

[17] Z. Shu, J. Wan, J. Lin, S. Wang, D. Li, S. Rho, C. Yang, Traffic engineering in softwaredefined networking: measurement and management. IEEE Access 4, 3246–3256 (2016). http://www.ieee.org/publications_standards/publications/rights/index.html.

[18] A. M. A. Alnaser, Certain Contributions in Traffic Engineering Based on Software-Defined Networking Technology , Journal of Computer Science 15 (7), 2019, pp 944.953, DOI: 10.3844/jcssp.2019.944.953.

[19] A. M. A. Alnaser, Streaming algorithm for multipath secure routing in mobile networks, IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 4, No 1, July 2014 pp 112 – 114.

[20] M. R. Abbasi, A. Guleria, M.S.Devi, Traffic engineering in software defined networks: a survey. J. Telecommun. Inf. Technol. 4, 3–13 (2016).

[21] A. M. A. Alnaser, Novel Properties for Total Strong - Weak Domination Over Bipolar Intuitionistic Fuzzy Graphs.

[22] Ahlam Fallatah, As'ad Alnaser, Mourad Oqla Massa'deh , Bipolar Intuitionistic Fuzzy Graph Over Cayley Groups , J. Math. Comput. Sci. 11 (2021), No. 5, 6403-6419.

[23] Ahlam Fallatah, Mourad Oqla Massa'deh, As'ad Mahmoud As'ad Alnaser , Some Contributions on Operations and Connectivity Notations in Intuitionistic Fuzzy Soft GRAPHS , Advances and Applications in Discrete Mathematics , Volume 23, Number 2, 2020, Pages 117-138.

[24] A. M. A. Alnaser1, Wael A. AlZoubi1 and Mourad O. Massadeh, Bipolar Intuitionistic Fuzzy Graphs and it's Matrices, Applied Mathematics & Information Sciences 14(2)(2020), 205-214.

Analysis of Crime Pattern using Data Mining Techniques

Chikodili Helen Ugwuishiwu[1], Peter O. Ogbobe[2], Matthew Chukwuemeka Okoronkwo[3]
Computer Science Department, University of Nigeria, Nsukka, Enugu state, Nigeria[1, 3]
National Board for Technology Incubation, Abuja, Nigeria[2]

***Abstract*—The advancement in Information Technology permits high volume of data to be generated in databases of institutions, organizations, government, including Law Enforcement Agencies (LEAs). Technologies have also been developed to store and manipulate these data to enhance decision making. Crime remains a severe threat to humanity. Criminals currently, exploit highly sophisticated technologies to perform criminal activities. To effectively combat crime, LEAs must be adequately equipped with technological tools such as data mining technology to enable useful discoveries from databases. To achieve this, a Real-time Integrated Crime Information System (RICIS) was developed and mobile phones were used by informants (general public) to capture information about crimes being committed within Southern-East, Nigeria. Each crime information captured is being sent to the LEA responsible for the crime type and the information is stored in the agency database for data analysis. Thus, this study uses data mining algorithms to analyze crime trends and patterns in Southern-Eastern part of Nigeria between 2012 and 2013. The algorithms adopted were Classification and Rule Induction. The data set of 973 were collected from Eleme Police station, PortHarcourt (2012) and Nsukka Police station (2013). The analysis enables identifications of some trends of crimes and criminal activities from various LEAs databases, enhancing crime control and public safety.**

Keywords*—*Information technology; law enforcement agency; data mining; crime; classification and rule induction

I. Introduction

Technology advancements have made the world a better place including access and manipulation of huge volume of dataset in virtually all fields of life [1]. Criminology is an important area for applying data analysis and it is a practice aimed at discovering crime characteristics [2]. Crime is of immense concern in our world today. Crime ranges from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks) [3]. Crime has negatively impact both developed and developing nations because of its emotional, economic and social disruptive tendencies. It threatens the quality of life, human rights and poses severe challenges to any society [4][5]. The motivation of this study is to understand the trend, pattern and the prevalence of each crime type in different parts of the South-East, Nigeria. This properly advises the Law Enforcers (LEs) on the most effective approach on crime management. Criminals today, use sophisticated technological tools [6] not just to commit crimes but also to avoid being detected. Unfortunately, most of the Nigeria LEAs are still using paper-based information systems to capture and store crime data, leading to delay in crime information flow and inefficient crime data analysis. The work aims to analyze crime data generated from the implemented RICIS model to assist the LEAs perform their operations.

Crime control and prevention are fundamental to the welfare, stability and development of any society. Though, Government and communities have been making effort to improve on security standard of the public (through establishing different LEAs and community policing), more effort should be done to equip the LEAs with current technologies (e.g. crime analytic tools) for effective crime management. Timely access to relevant information is also of utmost necessity in day-to-day business of LEA, especially in crime investigation and detection of criminals [7].

Data mining is a tool that enables efficient extraction of useful information and patterns from complex and large datasets [6][8][9][10]. It is primarily aimed at uncovering hidden relationships in data warehouse with the aid of artificial intelligence method [11][12]. Data mining has been applied in the areas of crime analysis and prediction to assist the LEs in crime decision making [13].

Crime analysis is a task that includes exploring and detecting crimes and their relationships with criminals to assist security personnel in planning the deployment of available resources for the prevention and suppression of criminal activities [14].

A crime analysis should be able to identify crime patterns quickly in an efficient manner [7]. Some areas of data mining applications include supermarket, hospitals, banks, insurance companies, airline, governments and many other fields of life [15]. Data mining has been used to solve some challenges such as detection and prevention of fraudulent activities in telecommunication services, criminal's activity, etc.

Some data mining algorithm for knowledge discovery in databases include Classification, Rule induction, Association rules, Clustering, Forecasting, and Visualization, etc. Most times, researchers combine data mining techniques to achieve more precise and accurate extractions. Some software packages (built in mining tools) [16] [17] [18] designed for implementation of these techniques for data analysis include: Waikato Environment for Knowledge Analysis (WEKA), KNIME, ORANGE, etc. PHP and other programming tools were used to implement the algorithms (classification and Rule induction) for data analysis. The research will help LEAs and as well the government in improved decision-making and public safety.

The remaining sections are organized as follows: Section 2 discusses literature review; Section 3 deals with the analysis, design and method; Section 4 is on implementation, analysis and results, and finally Section 5 discusses conclusion.

II. Review of Literature

This section deals with the data mining basic concepts, the imperatives for crime analysis and review of some existing literatures in the related areas.

A. Data Mining and Imperatives for Crime Analysis

Three steps involved in data mining include: exploration, pattern identification and deployment [1]. Exploration means to clean and transform data into a new form, then important variables and nature of data are determined based on the problem; Pattern Identification means to identify and choose the patterns which make the best prediction. Deployment organizes the patterns for desired outcome [1]. According to [16], [19], [15], crime data analysis provides summary statistics, general and specific crime trends to LEs in timely manner to enable understanding on crime and criminal behaviour. This assists LEAs to be proactive in crime detection and prevention while managing their limited resources effectively.

B. Types of Crime Analysis and Data Mining Tasks

The crime analysis types include:

1) Tactical crime analysis generates information on where, when, and how crimes take place to assist officers and investigators in identifying and understanding specific and immediate crime problems.

2) Strategic analysis examines long term changes in crime, known as “crime trends”. Administrative Crime Analysis provides summary statistics, and general trend information.

3) Criminal investigative Analysis involves profiling suspect and victims for investigators based on analysis of available information. One may analyze to find out the type of person committing a particular crime series.

4) Intelligence analysis focuses on organized crime, terrorism, and supporting specific investigations with information analysis and presentation.

5) Operations analysis examines how LEA is using its resources. It focuses on such topics as deployment, use of grant funds, budget issues, etc. [15].

Two stages of data mining Tasks include:

a) Data collection phase: The researcher collects the training dataset from a defined source.

b) Data Pre-processing Phase: Processing the collected data to get it suitable for analysis through data cleaning, integration, transformation, reduction and discretization [16].

C. Types of Data Mining Techniques

Classification deals with discovering and sorting crime data into groups or predefined classes such as type, location, time, etc. based on certain attributes/criteria discovered from databases [20][21]. Classification is the process of learning a function f that maps each attribute of a set $X=\{x1, x2, \dots, xn\}$ to a predefined class label y ***[21]***. The goal is to build a model to predict future outcomes. Data classification helps to predict how new individuals or events will behave based on the classification criteria [17]. Classification has been used to detect email spamming and find authors who send out unsolicited emails [11], [18].

Rule induction is a machine learning technique in which through observations, rules are extracted from a dataset. These rules may possibly denote a scientific model of a dataset or signify a pattern in the dataset. A rule is conveyed with “if-then statements”. Rule-based algorithm takes training dataset as input and generates rules by dividing the table with cluster analysis.

Association rules are generating rules from crime dataset based on frequent occurrence of patterns [16]. Association Rule searches for relationship between data that exist together in a given record to uncover crucial information. In [18], association rule detected suspicious e-mails by identifying unusual and deceptive communication in e-mails.

Clustering techniques is used to automatically associate different objects (such as persons, organizations, vehicles) that are similar to one another and dissimilar to objects of another group in crime records. Clustering is based on finding relationships between different Crime and Criminal attributes having some previously unknown common characteristics [15]. In crime analysis, cluster analysis is used to identify areas with high concentration of a particular crime type. By identifying these crime "hot spots" i.e. where a similar crime has happened over a period of time helps to manage law enforcement resources more effectively [3].

Forecasting deals with discovering patterns and data that may lead to reasonable predictions. It estimates the future value based on a record's pattern. It deals with continuously valued outcome. Visualization enables miners to rapidly and efficiently locate vital information that is of interest within the data [17].

D. Review of Related Literature

Data mining technology has been adopted by organization improve on business strategies through performing different knowledge discoveries [1]. Authors in [11] reviewed data mining techniques and presented four case studies of their crime data mining project as follows: entity extraction for police narrative reports, detecting criminal identity deceptions, authorship analysis in cybercrime and criminal network analysis. According to [16], a model was proposed for crime and criminal data analyzes using clustering and association rules algorithm. The work tends to help LEs in discovering crime patterns and trends, making forecasts and as well identifying possible suspects. The intension was to assist the Libyan Government on strategic decisions to reduce the high increase on crime rate.

In [19], k-means clustering algorithm was used to perform data analysis to assist LEs in crime reduction. The goal was to extract useful information from crime dataset to enable LEs to identify and analyze crime patterns for effective crime control and prevention. WEKA and Microsoft Excel were used for analysis. Data mining and machine learning tools such as the

Additive Regression, Linear Regression were used to conduct a comparative study between the violent crime patterns from the Communities and Crime Unnormalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi provided by neighborhoodscout.com. The researchers implemented the three selected tools using the same finite set of features on the Communities and Crime Dataset and found that the linear regression algorithm performed best in predicting violent crime patterns [17].

A concise review on Data mining techniques and its applications on over 100 different areas on crime analysis were presented in chronological order [18]. In [22], clustering algorithm was used for crime data analysis to detect crime patterns and speed up the process of solving crime issues. The aim was to formulate crime pattern detection as machine learning task and using data mining to support LEs in crime struggle. The modeling technique was able to identify the crime patterns from the crime data. Research on [3], presented a detailed overview of clustering technique, its types and some of its application areas.

In [2], data clustering method and artificial neural networks were adopted to unsupervisely classify crimes and crime pattern recognition respectively. This research finally proposed a method for crime matching processes. In [23], binary clustering and classification techniques were used to analyze criminal data and was aimed to potentially discover a criminal based on the witness/clue at the crime spot. An auto correlation model was used to validate criminals on the process. Researchers in [24] focused on property crime and explore the best approach to reliably predict the location, time, and/or likelihood of future residential burglary. The work adopted classification model to perform the crime forecasting.

Authors in [13] studied data mining technologies used in criminal investigation, highlight the approach adopted in crime data analytics and summarizes some challenges arising during the analysis process**.** A machine-learning classification algorithms, such as boosted decision tree and K-Nearest Neighbour (KNN) were adopted to create a model that will perform crime prediction on Vancouver dataset collected for a period of 15 years (2003-2018). It is recorded that prediction accuracy between 39% and 44% was achieved [25].

III. Design and Methods

A. Methodology

Data collection and the research output adopted quantitative method. The 973 primary data were collected from Eleme police station PortHacourt (2012) and Nsukka police station (2013) in the South-East Nigeria. Subsequently, data for analysis will be collected from the RICIS developed. To get the desired outcome, the data collected passed through three stages of data mining: exploration, pattern identification and deployment. The data were pre-processed using different pre-processing techniques (cleaning, missing values and removing inconsistency). The pre-processed data were used to find different crimes, criminal trends and behaviour. The crimes and criminals were grouped into clusters according to their important attributes. The data mining algorithm adopted were classification and Rule Induction. The Psuedocode of the two techniques is presented in section 3.3.

B. Conceptual Framework of the Study

The framework of the proposed research presents the summary of the research in a block diagram as shown in Fig. 1. This gives details of the processes of the mining from data collection to the final report (result). The work aims to analyze crime data generated from the implemented RICIS model shown in Fig. 2, to assist the LEAs perform their day-to-day operations, though, data currently used for analysis in this study was collected from two different Nigerian Police stations as mentioned in Section 3.1.

C. Analysis of the RICIS System

The RICIS is a crime reporting system developed to capture crime information from informant as shown in Fig. 2. The components of the RICIS as shown in Fig. 2 include an informant who reports crime, a mobile phone for crime report, an application for directing crime information to the appropriate agency, a middleware to enable data exchange among the agencies, individual agency's databases for their data autonomy, a centralized database for data analysis and backup, internet to enable communication and finally the bidirectional arrows indicating the direction of information flow.

This system enables LEA to collaborate by sharing resources such as one agency searching for the existence of a criminal's information in any of the participating agency or LEAs redirecting a crime information wrongly sent to any of them to the appropriate agency for instant responses. The system also enables LEAs substitute the traditional paper-based crime reports systems with an electronic system. The system provides instant response to crime and generate huge amount of crime data for data analysis. The reported crime provides data such as location codes, crime type, where, when, and how crimes happens (short crime details), information about suspect (if identified) and the victim. These information enables the law enforcement officers to offer a just-in-time response to crime scene and have quick access to the right data.

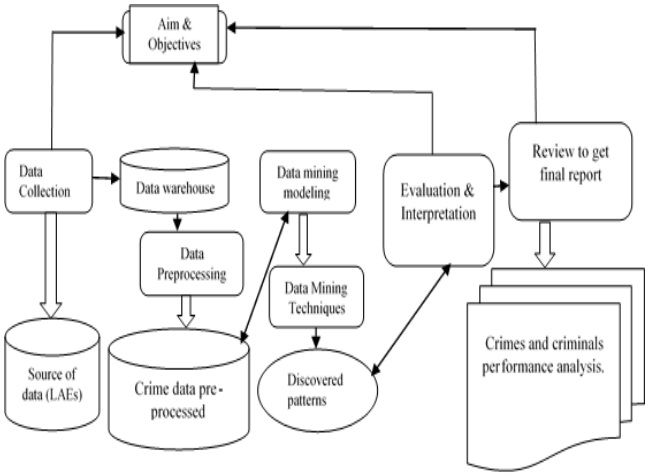


Fig. 1. Conceptual Framework of the Data Mining Process.

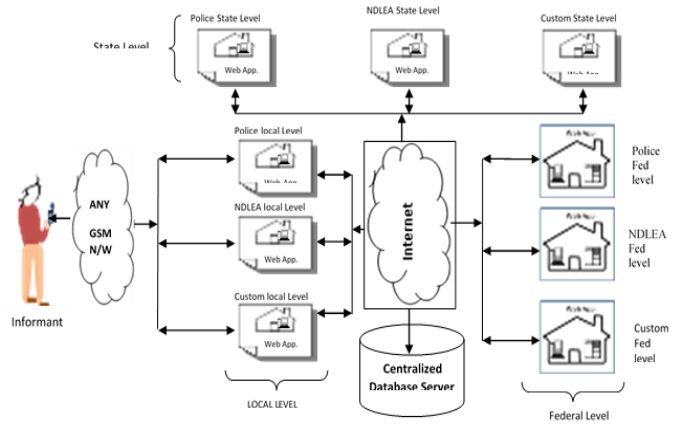


Fig. 2. The RICIS Model.

Mining the crime data on RICIC extract some crime pattern and relevant information such as: most prevalent crime in the country, state with highest number of criminals or crimes, age range of criminals with highest crime rate, location with highest crime rate, the class of people that commit most crimes (literate, illiterate, married, single, jobless etc.), time of the day crime usually occur in a particular place and which week of a month or month of the year or year with highest number of crime, etc. The results from this system can provide statistics of crimes handled in an agency in a specified time (e.g. number of crime incident investigated and handled in a year, etc.). This will to a great extent help LEs to manage their limited resources effectively and improve on the security standard of the nation.

D. Data Mining Algorithm

This section shows the algorithms (Pseudocode) of the data mining techniques (Classification and Association rule) used. The pseudo code shows the sequential steps taken in the implementation of these techniques. The analysis done is a modified algorithm that combines both classification and association rule. Association rule is a learning algorithm based on rule or adaptation [26][27]. The algorithm must be provided with TrainingSet already been classified in order to generate a list of classification rules.

Rule Induction algorithm

1.0 Input data
1.1 let the ClassificationRuleList be empty
2.0 Perform classification
2.1 repeat
2.1.1 let the ConditionalExpressionSet be empty and let the BestConditionExpression be nil
2.1.2 Repeat
2.1.2.1 Let the TrialConditionalExpressionSet be the set of conditional expressions, {x and y where x belongs to the ConditionalExpressionSet and y belongs to the SimpleConditionSet}.
2.1.2.2 Remove all formulae in the TrialConditionalExpressionSet that are either in the ConditionalExpressionSet (i.e., the unspecialized ones) or null (e.g., big = y and big = n)
2.1.2.3 for every expression, F, in the TrialConditionalExpressionSet
2.1.2.3.1 If F is statistically significant and F is better than the BestConditionExpression by user-defined criteria when tested on the TrainingSet then replace the current value of the BestConditionExpression by F
2.1.2.4 While the number of expressions in the TrialConditionalExpressionSet > user-defined maximum
2.1.2.5 Remove the worst expression from the TrialConditionalExpressionSet
2.1.2.6 Let the ConditionalExpressionSet be the TrialConditionalExpressionSet
2.1.3 until the ConditionalExpressionSet is empty
2.1.4 if the BestConditionExpression is not nil then let the TrainingSubset be the examples covered by the BestConditionExpression
2.1.5 Remove from the TrainingSet the examples in the TrainingSubset
2.1.6 Let the MostCommonClass be the most common class of examples in the TrainingSubset
2.1.7 Append to the ClassificationRuleList the rule 'if ' the BestConditionExpression ' then the class is ' the MostCommonClass
2.2 Until the TrainingSet is empty or the BestConditionExpression is nil

3.0 return the ClassificationRuleList [28].

The classification algorithm classifies data based on rules. The algorithm goes thus:

1.0 Input data
1.1 Convert continuous values to categorical values.
1.2 Set all classified to false (all classified is a variable that checks if all the data values are totally classified or not)
1.3 Set n = 0
1.4 Initialize classified dataset to empty
2.0 Repeat
2.1 pick the value of an attribute
2.2 generate rule for the combination (inserting value into a classified dataset)
2.3 classify the value of the attribute
2.4 set n=n+1
2.5 if n = total number of data in raw dataset then set all classified to true, go to 3.0
2.6 else repeat from step 2.1 until all classified = true

3.0 return classified data.

In the combined algorithm, there are key things to note:

1) Attributes – these are the field names.

2) Instances – these are the rows(records) in a dataset.

3) Class – this is the value of attribute (data stored in a field).

4) Subclass – this is the node of a class.

The task here is to loop through all the instances, pick the classes and classify them according to their respective attributes. In each classification, there is also a sub classification. Each classification has inner classifications which are based on the original criteria for mining the data. These are said to be sub attributes. The algorithm employs statistical method in forming a rule. Basically, the system used these techniques to display the crime type that is most committed, location with highest crime, age group with highest crime, state/LGA with highest crime, etc. Sorting rule is used to form this assertion. When an array is sorted in descending order, those elements at the beginning of the array are the ones with highest frequency, therefore to form the rules from the sorted array becomes easier. The data mining algorithm goes thus:

1.0 Input data
- *1.1 Input data mining criteria (select attribute)*
- *1.2 read data from the database and store in unclassified dataset*
- *1.3 initialize sub-attributes*

2.0 Manipulate data
- *2.1 convert continuous values to categorical values*
- *2.2 set classified dataset to empty*

3.0 LOOP through the unclassified dataset
- *3.1 Pick a class from an instance based on selected attribute*
- *3.2 Search through the Classified Dataset for the class*
  - *3.2.1 If class is not found then add a new node of the class to the classified dataset*
    - *3.2.1.1 Loop through the sub attributes*
    - *3.2.1.2 Pick a subclass from the instance*
    - *3.2.1.3 Add the subclass as a sub node to the class node*
    - *3.2.1.4 Repeat from 3.2.1.2 until all the sub attributes are treated*
  - *3.2.2 Else update the node where the class is found in the dataset*
    - *3.2.2.1 Loop through the sub attributes*
    - *3.2.2.2 Pick a subclass from the instance*
    - *3.2.2.3 Search for the subclass in the parent class node in the classified dataset*
    - *3.2.2.4 If subclass is not found then add the subclass as a sub node to the class node*
    - *3.2.2.5 Else update the subclass sub node in the class node*
    - *3.2.2.6 Repeat from 3.2.2.2 to 3.2.2.5 until all the sub attributes are treated*
- *3.3 Repeat from step 3.1 to 3.2.2.5 until all data are classified*

4.0. Generate Rules for the manipulated data
- *4.1 Sort the classified dataset in descending order*
- *4.2 Set n = 0*
- *4.3 Loop through the nodes of the classified dataset*
  - *4.3.1 Pick a class at node n*
  - *4.3.2 If n = 0 and the number of occurrence of class in node n is greater than the class in node n+1 then the rule follows that the class in node n is the best fit for the test. Go to step 4.4*
  - *4.3.3 Else if n > 0 and the number of occurrence of class in node n is greater than the class in node n+1 then the rule follows that all the classes from node 0 to n are the best fit for the test. Go to step 4.4*
  - *4.3.4 Else set n = n + 1*
  - *4.3.5 Repeat from step 4.3.1*
- *4.4 Return the results*

5.0 Print results
- *5.1 Display the results node by node*
- *5.2 End*

IV. System Implementation

These sections document the implementation and the results of the proposed system. Fig. 3 shows list of all reported cases that are yet to be investigated. To display this, simply login and select state and the LGA where the crime was committed.

Fig. 4 displays the investigated crimes (centralized database) where one can mine data. In mining crime data, the following 13 different parameters were used: Person, Sex, Age, Marital Status, Occupation, Educational Status, State, LGA, Crime Location, Crime Date, Crime Time, Crime Type and Tribe. If a parameter is selected, the program analyzes the crime data based on that parameter. Fig. 5 shows mining by age range. Data mining results on classification based on age showed that out of 973 crime data collected in 2013, criminals within the age range of 31-40 committed the highest number of crimes with 263 cases. In each analysis, there are always sub-analysis which re-analyzes the crime cases using all other parameters except the selected one.

Fig. 3. List of Crime Cases Yet to be investigated.

Fig. 4. List of Investigated Crimes.

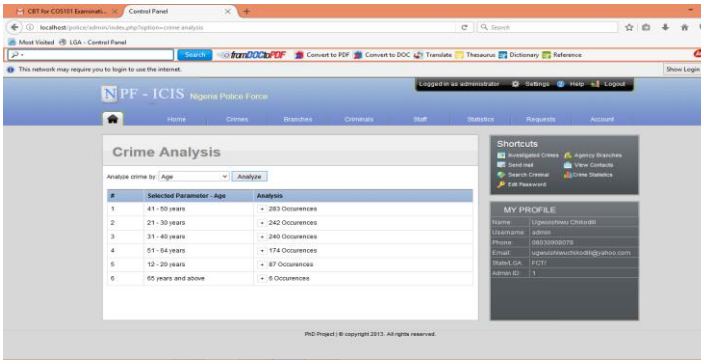


Fig. 5. Crime Analysis by Age Range.

Table I shows a clearer picture of the result in Fig. 5. The sub-analysis of the result on Fig. 5 is shown on Fig. 6. Result from Fig. 6 shows that Beach Junction Nsukka is a location with the highest number of crimes.

Table II shows a comprehensive crime statistics of all crime types committed in 2013 based on the collected. The result is presented state by state and the total occurrence of each crime type within the year is given. This crime distribution is based on the state of origin of the criminals i.e. citizens of the country that committed crime in Enugu state. Data mining results on classification showed that out of 973 crime data collected in 2013, Anambra state has the highest number of criminal cases with 192 crimes. It may also interest a reader to note that because this research was carried out in the Eastern part of the country, the concentration of these criminals' states of origin was spread mostly in the South eastern part of the country.

TABLE I. THE CRIME ANALYSIS RESULT BY AGE RANGE AS SHOWN IN FIG. 5

#	Age Range	Analysis
1	41 - 50 years	228 Occurences
2	21 - 30 years	229 Occurences
3	31 - 40 years	263 Occurences
4	51 - 64 years	168 Occurences
5	12 - 20 years	83 Occurences
6	65 years and above	2 Occurences

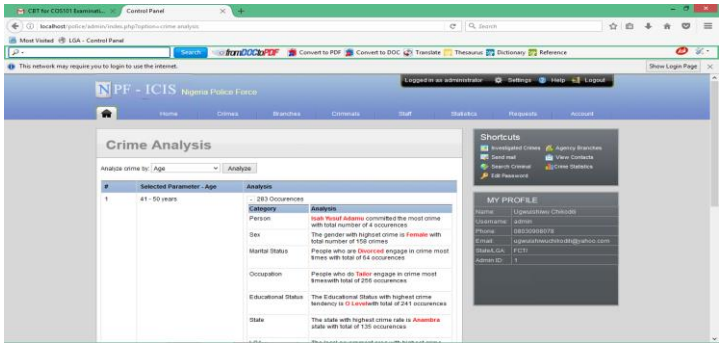


Fig. 6. The Sub-Analysis of the Result shown in Fig. 6.

TABLE II. CRIME STATISTICS FOR 2013 BASED ON THE CRIMINAL'S STATE

#	Criminal's State	Abduction	Assault	Assault Occ. Harm	Car Snatching	Car Theft	Certificate Forgery	Conduct	Defilement	false witness	Fire Incident	Forgery of Police ID	G. Harm	House Breaking	Illegal Roadblock	Indecent Assault	Keke NAPEP Snatching	Kidnapping	M. Damage	M/Stealing	Malicious Damage	Missing Person	MT Stealing	Murder	OBT	Phone Snatching	Pocket Picking	Raping	Robbery	Sexually Harassment	Stealing	Store Breaking and S	Threat to life	Threatening Violence	Vandalization	Wounding
1	Abia	2	3	3	5	1	6	4	5	1	1	1	5	1	3	2	3	3	3	2	3	4	1	3	2	0	1	4	4	3	5	4	1	2	5	3
2	Adamawa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Akwa Ibom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Anambra	5	8	7	7	9	9	5	4	7	2	5	5	4	4	7	5	3	6	5	2	5	8	5	7	7	6	4	4	5	8	3	8	7	8	6

5	Bauchi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Bayelsa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Benue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Borno	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Cross River	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Delta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Ebonyi	1	1	1	1	2	0	1	0	2	0	1	1	3	3	1	1	0	2	1	0	0	3	0	4	1	1	1	1	0	1	0	1	0	0	2
12	Edo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Ekiti	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Enugu	1	0	1	0	2	1	2	1	4	2	3	1	0	2	2	3	1	1	2	3	0	1	1	2	1	0	1	3	4	2	4	1	0	2	1
15	Gombe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Imo	1	1	1	1	0	2	3	2	4	2	1	2	2	1	2	2	1	2	0	0	2	1	0	2	0	3	2	0	0	2	2	2	2	2	1
17	Jigawa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Kaduna	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Kano	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Katsina	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Kebbi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Kogi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Kwara	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Lagos	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Nassarawa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	Niger	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Ogun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Ondo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Osun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	Oyo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	Plateau	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	Rivers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Sokoto	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	Taraba	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Yobe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	Zamfara	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	FCT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Total	**10**	**13**	**13**	**14**	**14**	**18**	**15**	**12**	**18**	**7**	**11**	**14**	**10**	**13**	**14**	**14**	**8**	**14**	**10**	**8**	**11**	**14**	**9**	**17**	**9**	**11**	**12**	**12**	**12**	**18**	**13**	**13**	**11**	**17**	**13**

V. CONCLUSION

A Real- Time Integrated Crime Information System was developed and data analysis performed. It is intended that when this application is fully deployed, data analysis should be done on crime data available in the RICIS centralized database. This work used classification and Association rule algorithms to identify useful crime trend and patterns. These machine learning algorithms (Psuedocode) were implemented using PHP and other programing language tools. The results of this data mining will support LEAs in effective decision making, management of their limited resources and will considerably reduce crime rate and increase the security standard of the nation.

A. Limitations of the Study

There was a big challenge during data collection because most of these Nigerian LEAs do not have electronic crime data. The data I got was manually picked into a table format one after the other and this made things difficult. Nigeria Custom Services, Nigeria Police and NAFDAC offices were visit by the researcher but data were collected from only two police stations – Eleme Police station and Nsukka police station, the major reason being that there was no electronic database to access.

B. Recommendations

The system is used by virtually every citizen of the country to fight against crime. Specifically, this application is meant for Nigerian LEAs to manage societal crime issues. LEAs of any country, other governmental agencies, private and public organization can equally use this system with a little modifications. Government can use this application to determine the efficiency of LEAs on crime management by taking the statistics of crime handled within a specified time.

WEKA and other mining software tools may be considered in the data analysis instead of developing new mining application, increase in the data size and mining parameters, and as well data mining techniques such as clustering and linear regression may be may be adopted for a more improve analysis.

Comments: The data collected are not sufficient to make for conclusive evidence. How does this provide for privacy of persons under investigation? The supposed ‘criminals’ are suspects, or have they been convicted by a court of competent jurisdiction?

REFERENCES

[1] Bharati M. Ramageri. “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305. Available at http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf [accessed Jan 26, 2017].

[2] M. Keyvanpoura, M. Javidehb, M. Ebrahimia (2011). Detecting and investigating crime by means of data mining: a general crime matching framework Procedia Computer Science 3 (2011) 872–880.

[3] Neha D. and B.M. Vidyavathi, “A Survey on Applications of Data Mining using Clustering Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 126 – No.2, September 2015. Available at http://www.ijcaonline.org/research/volume126/number2/neha-2015-ijca-905986.pdf[accessed Jan 27, 2017].

[4] Xingan Li, “Application of Data Mining Methods in the Study of Crime Based on International Data Sources”, 2014. Available at https://tampub.uta.fi/bitstream/handle/10024/95108/978-951-44-94192.pdf? sequence=1[accessed Feb 08, 2017].

[5] D. Tyagi, and S. Sharma. (2018). “An Approach to Crime Data Analysis: A Systematic Review.” International Journal of Engineering Technologies and Management Research, 5(2:SE), 67-74. DOI: 10.5281/zenodo.11975.

[6] C. Chauhan and S. Sehgal, "A review: Crime analysis using data mining techniques and algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 21-25, doi: 10.1109/CCAA.2017.8229823.

[7] S.Yamuna and N. S. Bhuvaneswari, “Data mining Techniques to Analyze and Predict Crimes” The International Journal of Engineering and Science (IJES) Vol 1 No. 2, PP 243-247, 2012. Available at http://www.theijes.com/papers/v1-i2/AJ01202430247.pdf [accessed Feb 08, 2017].

[8] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin and M. Chau, "Crime data mining: a general framework and some examples," in Computer, vol. 37, no. 4, pp. 50-56, April 2004, doi: 10.1109/MC.2004.1297301.

[9] R. Bhargava, P. Singh and R. S. Sangwa (2018). Analysis of Crime Data Using Data Mining Algorithm. International Journal of Engineering Sciences & Research Technology 7(2), 675-681.

[10] Akgöbek, Ömer (2013). A rule induction algorithm for knowledge discovery and classification. Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 21, DO - 10.3906/elk-1202-27.

[11] H. Chen, W. Chung, Yi Qin, M. Chau, J. Jie Xu, G. Wang, R. Zheng and H. Atabakhsh, “Crime Data Mining: An Overview and Case Studies”, 2003. Available at https://www.researchgate.net/publication/2870463_Crime_Data_Mining_An_Overview_and_Case_Studies [accessed Jan 26, 2017].

[12] S. P. Deshpande, and V. M. Thakare (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS) 1(1), DOI : 10.5121/ijdps.2010.1103 32.

[13] UKEssays (2018). Survey Of Data Mining Techniques On Crime Data Criminology Essay. https://www.ukessays.com/essays/criminology/survey-of-data-mining-techniques-on-crime-data-criminology-essay.php?vref=1.

[14] K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana, “Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model”, International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 2, No. 1, 2012. Available at http://ijcsits.org/papers/Vol2no12012/5vol2no1.pdf [Accessed on 26th Jan, 2017].

[15] Z. S. Zubi and A. A. Mahmmud, “Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention”, International Journal of Computers, Vol 8, 2014, pp.39-45. Available at http://www.naun.org/main/NAUN/computers/2014/a022007-096.pdf. [accessed 26th Jan, 2017].

[16] Z. S. Zubi and A. A. Mahmmud, “Using Data Mining Techniques to Analyze Crime patterns in the Libyan” Available at http://www.wseas.us/e-library/conferences/2013/Budapest/IPASRE/IPASRE-09.pdf [Accesses on 27th Jan,2017].

[17] L. McClendon and N. Meghanathan,” Using Machine Learning Algorithms to Analyze Crime Data”, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015. Available at http://airccse.org/journal/mlaij/papers/2115mlaij01.pdf [Accessed 26th Jan, 2017].

[18] H. Hassani, Xu Huang, E. S. Silva, and M. Ghodsi. “A Review of Data Mining Applications in Crime” 2016. Available from: https://www.researchgate.net/publication/301579904_A_Review_of_Data_Mining_Applications_in_Crime [accessed Jan 27, 2017].

[19] J. Agarwal, R. Nagpal and N. R. Sehgal, “Crime Analysis using K-Means Clustering”, International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, December 2013.Available at https://pdfs.semanticscholar.org/dcb6/bff7931d085e7bf6ff004b0d28fccfca22df.pdf [accessed Feb 08, 2017].

[20] M. Nayak, V. Yadav, Y. Patil (2019), Crimerate Prediction using Datamining. International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), 8(4), 50-52.

[21] K. B. Al-Janabi and H. K. Abdullah (2010). Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention Procedures. Iraqi conference for Information technology (ICIT).

[22] S. Varan Nath, "Crime Pattern Detection Using Data Mining", 2006. Available at http://cs.brown.edu/courses/csci2950-t/crime.pdf [accessed 26th Jan, 2017].

[23] U Mande, Y. Srinivas, J.V.R.Murthy / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 4, July-August 2012, pp.149-153.

[24] S. Rajkumar, P. M. Sakkarai, J. J. Soundarya, P. Varnikasree (2019). Crime Analysis and Prediction Using Data Mining Techniques. Proceedings of the First International Conference on New Scientific Creations in Engineering and Technology (ICNSCET-19), 602-607.

[25] S. Kim, P. Joshi, P. Kalsi, and P. Taheri (2018). Crime Analysis through Machine Learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

[26] P. Clark & T. Niblett, "The CN2 Induction Algorithm". Machine Learning Journal vol.3 No 4, Pp2, 1988. http://www.cs.utexas.edu/users/pclark/papers/cn2.pdf.

[27] J. Stefanowski (2010). Lecture note on Induction of Rules, Institute of Computing Sciences, Poznan University of Technology, Poznan, Poland.

[28] Wikipedia,"CN2 algorithm". Available at https://en.wikipedia.org/wiki/CN2_algorithm accessed on 17/04/16.

Human Face Recognition from Part of a Facial Image based on Image Stitching

Osama R. Shahin, Rami Ayedi, Alanazi Rayan, Rasha M. Abd El-Aziz, Ahmed I. Taloba
Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia

***Abstract*—Most of the current techniques for face recognition require the presence of a full face of the person to be recognized, and this situation is difficult to achieve in practice, the required person may appear with a part of his face, which requires prediction of the part that did not appear. Most of the current forecasting processes are done by what is known as image interpolation, which does not give reliable results, especially if the missing part is large. In this work, we adopted the process of stitching the face by completing the missing part with the flipping of the part shown in the picture, depending on the fact that the human face is characterized by symmetry in most cases. To create a complete model, two facial recognition methods were used to prove the efficiency of the algorithm. The selected face recognition algorithms that are applied here are Eigenfaces and geometrical methods. Image stitching is the process during which distinctive photographic images are combined to make a complete scene or a high-resolution image. Several images are integrated to form a wide-angle panoramic image. The quality of the image stitching is determined by calculating the similarity among the stitched image and original images and by the presence of the seam lines through the stitched images. The Eigenfaces approach utilizes PCA calculation to reduce the feature vector dimensions. It provides an effective approach for discovering the lower-dimensional space. In addition, to enable the proposed algorithm to recognize the face, it also ensures a fast and effective way of classifying faces. The phase of feature extraction is followed by the classifier phase. Displacement classifiers using square Euclidean and City-Block distances are used. The test results demonstrate that the proposed algorithm gave a recognition rate of around 95%, to validate the proposed algorithm; it compared to the existing CNN and Multibatch estimator method.**

Keywords*—Face recognition; image stitching; principal component analysis; Eigenfaces distance classifiers; geometrical approach

I. Introduction

Image stitching is the method used for consolidating various photographic images with an overlapping manner of view to get a sectioned display or high-resolution image. Most regular methodologies of image stitching require correct covers amongst images and indistinguishable exposures to create consistent results. Moreover, by using image stitching in computer vision and PC design applications, some digital cameras can stitch their photographs together internally [1]. Arrangement of the images may comprise at least two digital images taken of a solitary scene in various circumstances, from various sensors, or various perspectives. Image stitching strategies are classified into two general methodologies: feature-based techniques and direct techniques [2]. Feature-based techniques expect to identify a connection between the images through unmistakable features separated from the prepared images whereas direct techniques were dealing with all pixels of the parts of the image stitched. The feature-based strategy has the advantage of being more robust against direct techniques, and it can naturally find interlinkages between disorganized arrangements of images. Image stitching comprises three stages: Image Acquisition, Image Registration, and Image Blending.

Face recognition has wide applications in security, validation, surveillance, and distinct forensic evidence. Regular identification strategies such as ID cards and passwords are not considered as reliable as previously thought due to the various methods of hacking secret keys and others. As an option, biometrics, which is characterized by being a physical identification mark or belonging to a specific person, is not the same as others. The distinction between different individuals in the Known person database is the focal point of face recognition. In the future, face recognition is expected to oversee unlimited frameworks, such as access control for aircraft station security, smart home applications, structures, and faces, as well as for checking and monitoring buildings and vehicles and intelligent human-PC collaboration. In recent years, confrontational recognition has created limitless application fields from acknowledged assertion and proof to collaboration and correspondence between human devices and computers through video applications based on face recognition. Face recognition algorithms are regularly classified into three categories, which are comprehensive methods, feature-based methods, and hybrid methods.

The holistic methods category symbolizes the entire facial region as a high-dimensional vector that contributes to a classifier. Principle Component Analysis (PCA) is a successful agent technique for all-encompassing face recognition strategies, including Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA). Local methods, on the other hand, extract neighborhood characteristics from facial territories; for example, the eyes, mouth, nose, and cheeks. These features are utilized to characterize faces. Finally, in the hybrid method, both the holistic and local methods are utilized to perceive and distinguish a face [3]. The essential challenge confronting any calculation for facial recognition is the deficit in the appearance of the face that is expected to perceive. The motivation behind this work is to construct complete face images by utilizing the stitching image algorithm.

II. Related Work

Over the last two decades, numerous analysts have proposed and executed different display image stitching frameworks. In the authors have introduced new systems on image stitching based on the histogram-matching algorithm [4]. Histogram coordinating is utilized for image adjustment, so the images stitched have a similar level of brightness. At this point, the paper embraces the SIFT algorithm to separate the key features of the images and plays out the harsh coordinating procedure. This work followed by the RANSAC algorithm for fine matches finally ascertained the most suitable scientific mapping model between the two images and as indicated by the mapping relationship, a straightforward weighted normal algorithm was utilized for image blending.

Authors in [5] provided a specialized examination for the fast image stitching calculation based on SIFT. Firstly, the images are separated into squares. The component sorts these neighborhood image squares and is resolved. The element purposes of the nearby image squares are removed utilizing diverse streamlined techniques adaptively. Secondly, we utilize coordination to achieve the changed framework and the RANSAC algorithm connected to expel incorrect coordinating point sets. Finally, the stitched image can be achieved by image blending.

Another proposal [6] has provided a new approach for image stitching techniques using (DTW). This work proposes a novel technique that uses the Dynamic Time Warping (DTW) algorithm to coordinate sets of images for image stitching. They additionally perform a measurement-reducing plan that shrinks the computational multifaceted nature of the standard DTW algorithm without influencing its execution. The viability of their proposed technique is shown in the stitching of 50 sets of restorative X-beam images and its execution contrasted with those of standardized cross-relationship (NCC), minimum average correlation energy (MACE) channels, total of-square-contrasts (SSD), and the entirety of absolute-contrasts (SAD). Their technique likewise beats two generally utilized stitching programs accessible on the web called Hugin and Auto-stitch.

The work in [7] gives an in-depth review of the current image mosaicing algorithms by ordering them into a few classes. For each class, the principal ideas will be clarified, and afterward, the adjustments made to the fundamental ideas by various analysts clarified. Moreover, this paper additionally discusses the focal points and burdens of all the mosaicing classes. Several previous investigations have led the field of face detection and recognition in settled images through a different scheme of frameworks. Shahin and EL-SAYED proposed a face recognition system based on the Geometrical Approach. The geometrical approach consists of two phases, namely, the detection phase and the recognition phase. The edge of the tested face is detected in the first phase after some pre-processing operations have occurred on the tested image. The second phase attempts to identify the angles of the face outline. After calculating the angle vector of the tested face, it will compare with the training set angles vectors via a neural network to identify the face [8].

Recognizing the weaknesses and strengths of machine learning techniques is critical for real life applications as well as being a prerequisite for determining extensive research and development requirements. Papers on intense analysis model can be found in literature as part of either job that specifically focuses on the attributes of deep models, or ii) work which explores the attributes of deep models as component of that other participation. Papers in the first collective, like ours, typically investigate different models and current legal findings that address a variety of deep models as their main contribution, whereas papers with in second category introduce a new classification model and then analyze its attributes.

This work proposed a discriminative feature-learning approach to recognize the face using convolution neural networks (CNNs) [9]. The proposed center loss algorithm is needed for the task of face recognition. In this algorithm, they train robust CNNs to obtain the deep features with two key learning objectives, intra-class compactness, and inter-class dispensation, as much as possible, which are necessary to the process of face recognition. Kasar et al [10] proposed a strategy for face recognition using artificial neural networks (ANN). They examined the face recognition methods proposed by numerous specialists utilizing ANN, which are used as a part of the field of pattern recognition and image processing.

Furthermore in [11] proposed a strategy for face detection and recognition based on (PCA) – (LDA) and square Euclidean distance with the Viola-Jones algorithm. Their proposed strategy is based on the appearance-construct features that concentrate on the whole face image as opposed to neighborhood facial features. The initial phase in the face recognition framework is face location. The Viola-Jones' face location technique equipped for handling images to a great degree while accomplishing high identification rates is utilized. Feature extraction and measurement-reducing strategy connected after face recognition. The principal component analysis (PCA) technique is generally utilized as a part of pattern recognition. The Linear Discriminant Analysis (LDA) technique, used to overcome the disadvantages of PCA, has effectively connected to face recognition. It is accomplished by projecting the image onto the Eigenface space by PCA and then applying unadulterated LDA over it. Subsequently, Square Euclidean Distance (SED) was utilized. This distance classifier is required to identify the similarity between the tested face images with those located in the training set.

Finally, conducted a comprehensive survey on pose-invariant face recognition [12] (PIFR). They discussed the intrinsic challenges in PIFR and exhibited a complete audit of built-up systems. They characterized the current PIFR strategies into four classes, namely, pose-robust feature extraction methodologies. They described and assessed the inspirations, systems, geniuses/cons, and the execution of agent approaches.

The key contributions of the proposed work are summarized as,

- Initially, the face images dataset is trained in a system.
- At first, an input testing image is taken from the dataset.

- Consequently, image stitching process operation is performed to determine the missing part of the face.
- Moreover, feature extraction is done through the Eigen face approach through PCA and geometrical method.
- Finally, the displacement classifier that utilizes Square Euclidean and City- block distances is used for categorizing the faces.

III. Methodology

The proposed system developed in this paper is divided into two phases. The first phase is stitching the face image for the face that needs to recognize, while the second phase is face recognition. Fig. 1 illustrates the scheme of the proposed system.

Initially, the input face image was taken from the dataset. The image undergoes a face-stitching operation to find the missing part of the face. Through this, the complete face image is obtained, as the human faces are symmetrical on either side. After the image stitching operation, the features extraction is accomplished to classify the face image. Using the selected features, the face image is categorized as either known or unknown face by a distance classifier, which utilizes the Eigen-face approach with PCA and geometrical method.

A. Image Stitching

Image stitching is the method utilized for obtaining a more extensive field of perspective of a scene from a succession of halfway perspectives. It is an alluring exploration zone given its extensive variety of applications, including movement identification, determination upgrade, checking worldwide land use, and face insertion. These procedures can be classified into two groups: direct techniques and feature-based systems as shown in Fig. 2.

In a direct technique, every pixel located in the image is compared with each other, which is an exceptionally complex method. The direct way to perform an alignment between two images is to shift one image that corresponds to another by comparing pixels of the two images under testing. This comparison will depend on the rows or columns of each image and the mean square error (MSE) for each row or column will be calculated and will take as a reference to compare the given images. However, MSE is an example of the error metric. The [13] principal obstacle of direct methods is that they have a constrained scope of union.

The direct method utilizes data from all pixels. It iteratively refreshes a gauge of homography with the goal that a specific cost of work is limited. To speed up the error metric search process, hierarchical motion estimation is used. In hierarchical motion, an image pyramid is first [14] created and a search process over a fewer number of discrete pixels will perform at coarser levels.

Thirdly, Fourier-based [15] stitching depended on performing a convolution in the spatial domain resembles the summation of one signal with its conjugating of the other. In a parametric motion, a single constant translation vector with a correspondence map will be used. Finally, due to the feature points that may not be accurately located, an incremental motion refinement algorithm can calculate a more accurate matching score. However, incremental motion refinement needs more calculations than other algorithms [16], so they consider time-consuming, which reflects a decrease in its performance.

In contrast, feature-based systems are progressively more prominent and broader in mosaicing. This is particularly the result of the quality of new calculations and types of invariant features that have evolved over recent years. In Schmid represent a survey on key points detection and implements an experimental comparison to determine the repetitive features of detectors [17] and the information content available at each detected key point. The feature matching process will occur after detecting the features and key points. In addition [18] the feature matching process will determine which feature comes from locations in different images. The fundamental qualities of strong locators incorporate invariance to image scale invariance, interpretation invariance, and turn changes [19].

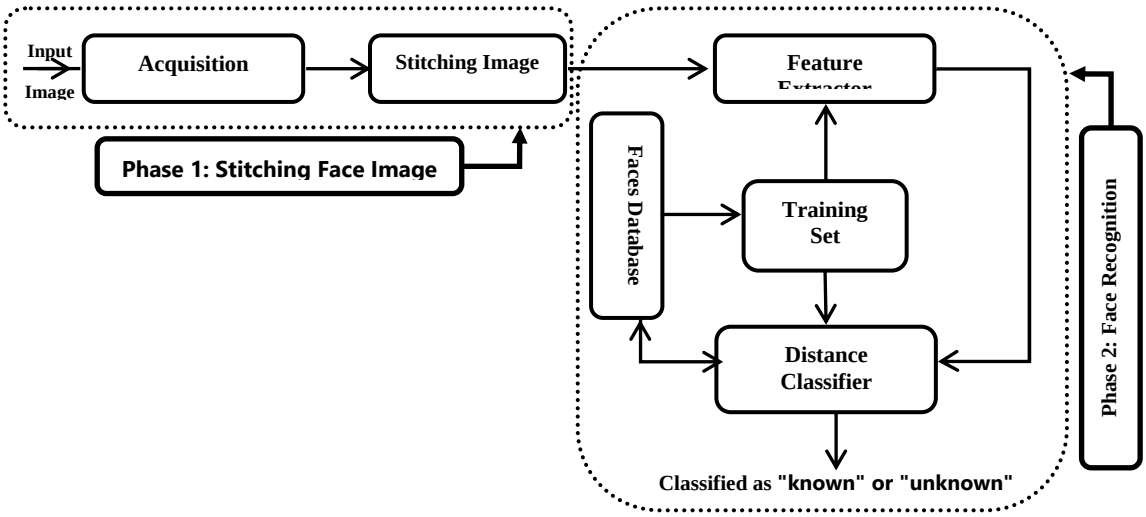


Fig. 1. Outline of the Typical Face Recognition System.

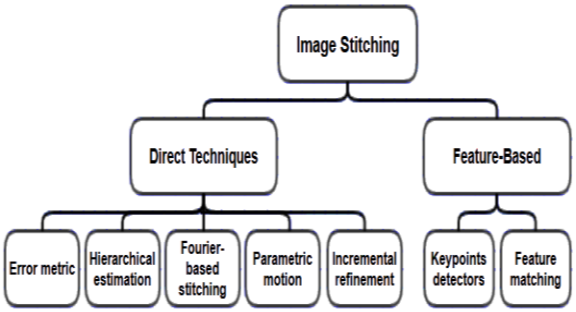


Fig. 2. Images Stitching Methodology.

Phase 1: Stitching the Face Image.

In this phase, the process of stitching the face image will be accomplished and completed according to the steps summarized in Table I.

TABLE I. PROPOSED FACE STITCHING APPROACH

1: Tested Image Acquisition: the image of the face to identify is processed. The facial image in this work assumed an incomplete face, which represents the main challenge in any face recognition system.

2: Nose Area Detection: the human face is symmetric around a certain line. To find such a line, the process of the nose detector is required. Here, the Viola-Jones Nose Detection Algorithm [20] was used. Then the centroid of the nose region is equivalent to the centroid of the rectangle that surrounds it. The centroid lies at the intersection point of the rectangle diagonals. The diagonals intersect at height $(\frac{h}{2})$ from reference y-axis and at width $(\frac{b}{2})$ from reference x-axis.

3: Vertical Line Drawn: A vertical line is drawn perpendicular to rectangle width (b) passing through the centroid. The whole face is symmetric around this vertical line.

4: Cropped Image: The image of the uncompleted tested face image cropped to obtain a cropped image I_1. The dimension of the cropped image I_1 from a pixel that has coordinate $x = 0$ to width equals $(x_0 + \frac{b}{2})$ from reference x-axis and from a pixel that has coordinate $y = 0$ to height equals the height of the vertical line that drawn in the previous step, which is equal to the height of the original tested image from reference y-axis. Whereas (x_0, y_0) represent the origin, i.e., [21] coordinate of the first pixel in the cropped nose detector image.

5: Flipped Image: A horizontal reflection of I_1 is performed to generate the missing part of the face, this part will be denoted as I_2.

6: Stitched Image: The first step is to calculate the relative positions of the obtained images and to produce a vacant set of images in the computer memory where these images will be assigned. The following stage is identifying the purpose of best correlation, which is performed by sliding contiguous image edges in the two headings until the point where the best match of edge features is found. The normalized cross-correlation coefficient for the case above is defined as in equation (1):

$$\text{Cross Correlation} = \frac{\sum_{x=0}^{L-1}\sum_{y=0}^{K-1}(w(x,y)-\bar{w})\left(f(x+i,y+j)-\bar{f}(i,j)\right)}{\sqrt{\sum_{x=0}^{L-1}\sum_{y=0}^{K-1}(w(x,y)-\bar{w})^2}\sqrt{\sum_{x=0}^{L-1}\sum_{y=0}^{K-1}\left((f(x+i,y+j)-\bar{f}(i,j)\right)^2}} \quad (1)$$

Where $w(x, y)$ represents a pixel value of the image to place; $\bar{w}$ is the mean value of all pixels included in the selected – cropped -box area $f(x + i, y + j)$ represents a pixel value of the composite image inside the box area. However, $\bar{f}(i, j)$ are the mean value of all pixels of the composite image within the box area and parameters K, L represents the box dimensions in the number of pixels included [21].

7: Image Blending: After all the input images had aligned with each other, we will use a multi-band image blending approach to produce seamless panoramic views by choosing a suitable compositing surface.

Fig. 3 depicted the outline of the image stitching process. Fig. 3(a) depicts the input test image. To find the symmetric line of the face, the nose detection technique is used and the nose-detected image is shown in Fig. 3(b). In addition, the centroid point of the nose is shown in Fig. 3(c) and the vertical line through that point is given in Fig. 3(d). The image is cropped through that vertical line to obtain the cropped image and it is shown in Fig. 3(e). Now, the image flipped to obtain the reflected part of the cropped image. It is shown in Fig. 3(f). The search area of the stitch is shown in Fig. 3(g). After face stitching is done, the complete face image would be obtained and it is shown in Fig. 3(h) and Fig. 3(i) is the final face image after the image blending operation is performed.

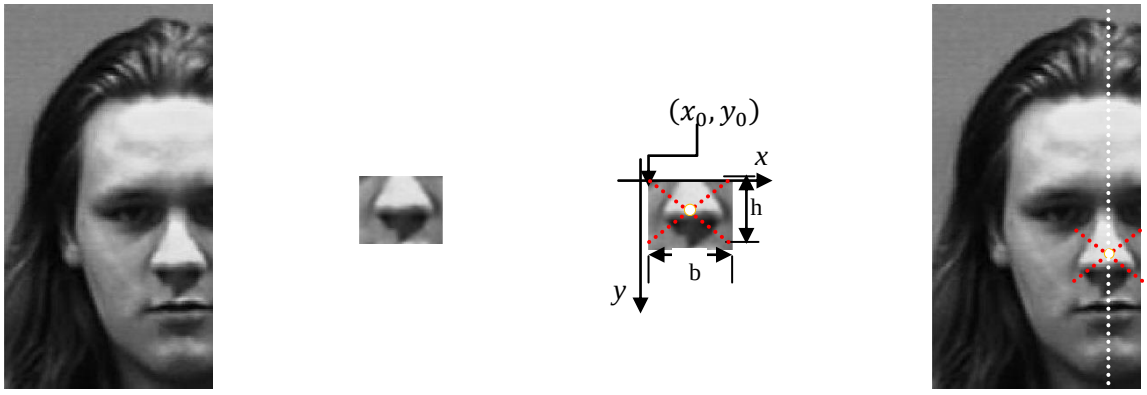
B. Face Recognition Techniques

Face Recognition has been an interesting topic for many computers science engineers who deal with artificial intelligence. The computer first detects the face and then for recognition, a step will be performed. Face recognition is considered a pattern recognition task performed precisely on human faces [8]. The outcome of this process is to classify either a face as "known" or "unknown" which compares the given unknown faces with stored known faces. The face image must be with a uniform background to avoid problems concerning the background complexities. However, it may be affected by the change in facial expressions. For example, laughter and crying change the mouth and eyes opening size, and aging also plays an important role as the face detail changes. Much research in computer recognition of faces has focused on identifying individual face features such as the nose, eyes, head outline, and mouth. The closest match between stored data and face image achieves recognition.

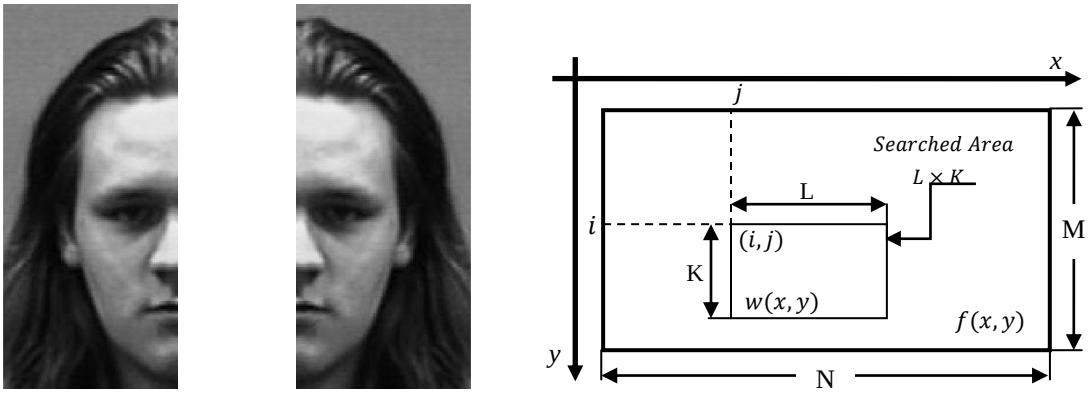
PCA is a standard technique used to distinguish patterns and signal processing. It is a statistical method employed to reduce data dimensions and extract features, which is an essential step in facial recognition. The analysis of basic compounds involves a mathematical procedure that converts several interrelated variables into several non-interrelated variables called basic components. These components are linked to the original variables by orthogonal transformation and are defined in such a way that the first primary component has the highest variance and the second fundamental component has the second-highest contrast and so on. Go back to Osama Shahin [8] present a face recognition approaches that depend on the geometry of the head which is done by calculating the angles for the head circumference and then storing these readings in a vector that describe a given face and will be used for comparison with other vectors that represents other cases.

Phase 2: Face Recognition Algorithm

The idea of the proposed system is to identify human faces if they are recorded in the database of the system as well as categorize individuals whose images are not recorded in the database as unqualified or as strangers through the process of automatic identification [22] and identification of persons. In this phase, Eigenfaces for recognition algorithm and Geometrical Approach for Face Detection and Recognition [8] will be used.



(a) Uncompleted Tested Face (b) Nose Detector (c) Centroid of the Nose Region (d) Vertical Line Drawn



(e) Cropped Image (f) Flipped Image (g) Searching area of Stitching

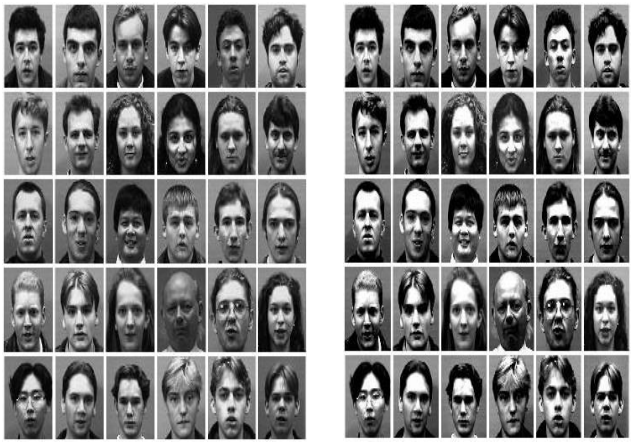
(h) Stitched Image (i) Image Blending

Fig. 3. Outline of the Image Stitching Process.

IV. Experimental Results

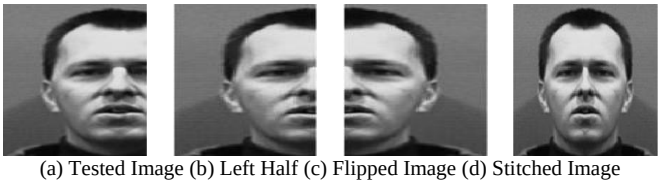
The system was tested on database FACES94. The training dataset contains a total number of 3080 images of 123 individuals grouped and classified into three categories (male, female, and male staff) taken with a little variation in head position. The image had a resolution of 180×200 datasets [23]. The training sample of face images is shown in Fig. 4(a). These training sample images normalized to minimize blunders caused by lighting conditions. The normalized face images as shown in Fig. 4(b).

Fig. 5 shows the test image that is fed into the proposed system for classification. The quality of the image stitching process is located by observing the seam lines between the stitched images.



(a) (b)

Fig. 4. (a): Initial Training, (b): Normalized Training Image Set.



(a) Tested Image (b) Left Half (c) Flipped Image (d) Stitched Image

Fig. 5. Tested and Stitched Image.

As previously mentioned, the experiments were performed on the FACES94 database with various numbers of training images. The percentage of discrimination calculated using the method of analyzing only the basic compounds in the extraction of properties and neural networks in discrimination and depending on the mainframe. The percentage of the recognition rate [24] is dependent on the number of Eigenfaces with the highest value of Eigenvalues. When 100 Eigenfaces were taken, the recognition rate was 94.2% with the Euclidean distance classifier and was equal to 92.3% when the City-Block classifier is used. When taking 150 Eigenfaces, the recognition rate rose to 94.6% with the Euclidean distance classifier and 93.1% with the City- Block classifier. This percentage then dropped to 92.2% and 91.8% respectively and then to 91.7% and 90.4% with the Euclidean and City-Block classifiers respectively when 200 and 250 Eigenfaces were taken. On taking 300 Eigenfaces, the percentage increased again to 95.1% and 93.7% respectively when Euclidean and City-Block classifiers were used. Fig. 6 and Table II show the result of the recognition process using Eigenfaces.

A comparative recognition error rate for glasses-persons is wearing glasses – and no glasses – persons are not wearing glasses – recognition using the eigenfaces and geometrical approaches methods are depicted in Table III.

Recognized Image Input Image

Fig. 6. Correct Face Recognition Result for FACES94 Database.

TABLE II. RECOGNITION RATES FOR THE PROPOSED ALGORITHM

No. of Eigenfaces	Recognition Rate % Euclidean Distance	Recognition Rate % City-Block Distance
100	94.2	92.3
150	94.6	93.1
200	92.2	91.8
250	91.7	90.4
300	95.1	93.7

TABLE III. A COMPARATIVE RECOGNITION ERROR RATES

Approach	Error rate % for glasses	Error rate % for no glasses
Eigenfaces approach	45.2	20.2
Geometrical approach	25.5	7.7

After the process of the recognition is performed, the calculation of Image Quality Metrics (IQM) between the stitched image and the original one is needed. In this work, MSE and CR [25] were chosen as quality factors. Table IV shows the advantages and disadvantages of several existing face detection techniques.

TABLE IV. ADVANTAGES AND DISADVANTAGES OF VARIOUS EXISTING FACE DETECTION TECHNIQUES

Methods	Advantages	Disadvantages
Feature face detection	More precise and low execution time	Maximum learning time
Geometric face detection	Effective approach and easy to implementation	Low precise and more false alarm
Haar like feature face detection	Feature extraction has been improved, and there are very few false alarms now.	Maximum execution time and implementation difficulty

A. Mean Square Error (MSE)

MSE is needed to measure image quality. Mean Square Error must be zero in a perfect case because it is the difference between the original images and the stitched ones. However, in a practical case the smaller the value of MSE, the better the quality of the stitched image.

$$\mathrm{MSE} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\mathrm{OI}(i,j) - \mathrm{SI}(i,j)}{m * n} \quad (2)$$

Where "OI" is the original image, "SI" is the stitched image, "M" & "N" is the numbers of rows and columns in both images respectively.

B. Correlation Coefficient (CR)

CR was used in measuring closeness between the original image and the stitched one. The correlation Coefficient should equal 1 in a perfect case. However, in a practical case value of CR near one is a significant figure.

$$C_r = \frac{\sum_m \sum_n (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{(\sum_m \sum_n (Xi - \bar{X})^2)(\sum_m \sum_n (Yi - \bar{Y})^2)}} \quad (3)$$

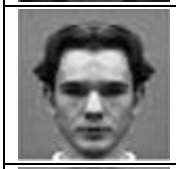
Where "$\bar{X}$" & "$\bar{Y}$" is the original image & stitched image average values respectively, "M" & "N" are the numbers of

rows and columns in both images respectively. Table V depicted MSE and Cr, the calculations for a sample of stitched and original images.

The accuracy and error rate comparison of the existing and proposed mechanism is shown in Fig. 7(a) and (b).

From the above figure, it is clear that the proposed method achieves higher accuracy of 99.78% and a reduced error rate of 0.22% compared to the existing CNN [9] and Multibatch estimator method approaches [26].

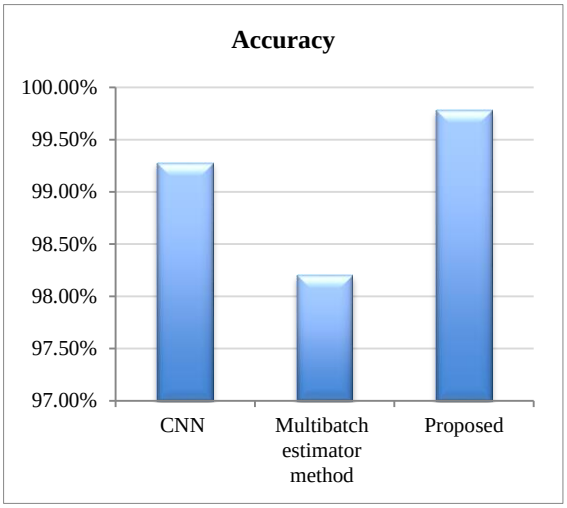
TABLE V. MSE AND CR ARE CALCULATIONS FOR A SAMPLE OF STITCHED AND ORIGINAL IMAGES

Cases		MSE	Cr
Stitched Image	Original Image		
		7.2011	0.9902
		13.2465	0.9734
		12.3301	0.9855
		9.3312	0.9977
		8.2135	0.9924

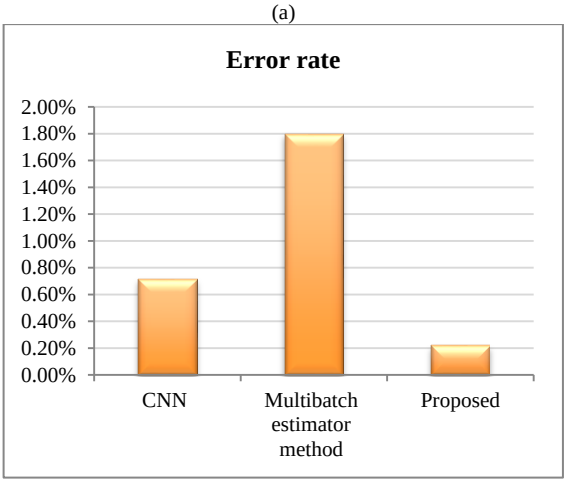
The proposed image stitching approach with Eigen-face and geometrical approach is compared with the existing CNN [9], Voting, and Random Subspace with Random Forest approaches [26]. The comparison of proposed and existing face recognition mechanisms is framed in Table VI.

TABLE VI. COMPARISON OF PROPOSED AND EXISTING FACE RECOGNITION METHODS

Parameter	Accuracy	Error rate
CNN	99.28 %	0.72 %
Multibatch estimator method	98.20 %	1.80 %
Proposed	99.78 %	0.22%



(a)



(b)

Fig. 7. (a) Accuracy Comparison of Existing and Proposed Methods, (b) Error rate Comparison of Existing and Proposed Methods.

V. CONCLUSION

This work proposes an improved algorithm for human face recognition by using two approaches: Eigenfaces and geometrical methods from part of a facial image, which is based on image stitching. The geometrical approach was accurate especially when the person wears glasses than Eigenface, but the Eigenface approach is quick and simple for the face recognition problem. Face recognition from the perspective of PCA connected on two distance classifier systems with a certain training dataset. The training dataset contained a total number of 3080 images of 123 individuals are grouped and classified into three categories (male, female, and male staff) and was taken with a little variation in head position. The test results identified that PCA gave the best results with the squared Euclidean distance methodology, with a recognition rate of 95.1%, which was more noteworthy than the city-Block distance strategy. The recognition rate also varied with the number of Eigenfaces used in the experiment. In future development, much improved methods can be ready for improved performance. With a larger data set, a greater number of pixels could be learned. Components that are less expensive can be suggested.

REFERENCES

[1] Ghosh and N. Kaabouch, "A survey on image mosaicing techniques," Journal of Visual Communication and Image Representation, vol. 34, pp. 1-11, 2016.

[2] S. Mistryand A. Patel, "Image stitching using harris feature detection," International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 4, 2016.

[3] J. K. Essel, "Head tilt classification using FFT-PCA/SVM algorithm," PhD diss., University of Ghana, 2018.

[4] J. Zhang, G. Chen and Z. Jia, "An image stitching algorithm based on histogram matching and SIFT algorithm," International Journal of Pattern Recognition and Artificial Intelligence, vol. 31, no. 04, 2017.

[5] Chen Yue, Yan Zhao, and S. G. Wang, "Fast image stitching method based on SIFT with adaptive local image feature," Chinese Optics, vol. 9, no. 4, pp. 415-422, 2017.

[6] S. Adwana, I. Alsaleh and R. Majed, "A new approach for image stitching technique using dynamic time warping (DTW) algorithm towards scoliosis X-ray diagnosis," Measurement, vol. 84, pp. 32-46, 2016.

[7] D. Vaghela and P. Naina, "A review of image mosaicing techniques," International Journal of Advance Research in Computer Science and Management Studies, vol. 2, no. 3, 2014.

[8] O. R. Shahin and E. Ayman, "Geometrical approach for face detection and recognition," Minufiya Journal of Electronic Engineering Research (MJEER), vol. 18,no.1, 2008.

[9] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A discriminative feature learning approach for deep face recognition",In European Conference on Computer Vision Springer, pp. 499-515, 2016.

[10] M. M. Kasar, D. Bhattacharyya and T. H. Kim, "Face recognition using neural network: a review," International Journal of Security and Its Applications, vol. 10, no. 3, pp. 81-100, 2016.

[11] N. H. Barnouti, S. S. Al-Dabbagh, W. E. Matti and M. A. Naser, "Face detection and recognition using viola-jones with PCA-LDA and square euclidean distance," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 5, pp. 371-377, 2016.

[12] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 7, no. 3, 2016.

[13] S. Pravenaa and R. Menaka, "A methodical review on image stitching and video stitching techniques," International Journal of Applied Engineering Research, vol. 11, no. 5, pp. 3442-3448, 2016.

[14] C. Guo, L. Wang and F. Deng, "The auxiliary model based hierarchical estimation algorithms for bilinear stochastic systems with colored noises," International Journal of Control, Automation and Systems,vol.18, no. 3, 2020.

[15] R. Perrot, P. Bourdon and David Helbert, "Confidence-based dynamic optimization model for biomedical image mosaicking," JOSA, vol. 36, no. 11, 2019.

[16] R. Szeliski, "Feature detection and matching in computer vision," Texts in Computer Science. Springer, London, 2011.

[17] Ma, J., X.Jiang, A.Fan, J.Jiang, and J.Yan, "Image matching from handcrafted to deep features," International Journal of Computer Vision, 2020.

[18] M. Karpushin, G. Valenzise and F. Dufaux, "Good features to track for RGBD images", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1832-1836, 2017.

[19] M. Z. Bonny and M. S. Uddin, "Feature-based image stitching algorithms", International Workshop on Computational Intelligence (IWCI), pp. 198-203, 2016.

[20] K. Vikram and S. Padmavathi, "Facial parts detection using viola jones algorithm," in Proceedings of the Advanced Computing and Communication Systems (ICACCS), pp. 1-4,2017.

[21] V. Rankov, R. J. Locke, R. J. Edens, P. R. Barberand B. Vojnovic, "An algorithm for image stitching and blending," In Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XII, vol. 5701, pp. 190-199, 2005.

[22] A. L. Machidon, O. M. Machidon and P. L. Ogrutan, "Face recognition using eigenfaces, geometrical PCA approximation and neural networks," in Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP), pp. 80-83, 2019.

[23] The Database of FACES94, http://cmp.felk.cvut.ci/space lib/faces/ faces94.html.

[24] O. R. Shahin and M. Alruily,"Vehicle identification using eigenvehicles," in Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-6, 2019.

[25] U. Sara, M. Akter andM. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study," Journal of Computer and Communications, vol. 7, no. 3, pp. 8-18, 2019.

[26] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz and A. Shashua, "Learning a metric embedding for face recognition using the multibatch method", in Advances in Neural Information Processing Systems, pp. 1388–1389, 2016.

Comparative Heart Rate Variability Analysis of ECG, Holter and PPG Signals

Galya N. Georgieva-Tsaneva, Evgeniya Gospodinova
Institute of Robotics, Bulgarian Academy of Sciences, Sofia, Bulgaria

***Abstract*—The article presents a demonstrative software system with included procedures for input, preprocessing and mathematically based analysis of cardiac data. The created program has the ability to work with the following signals: ECG, holter recordings, PPG signals. The presented system uses real cardiological data from patients and obtained with modern medical devices - electrocardiography, continuous holter monitoring, photoplethysmography device and others. The presented system allows mathematically based study of cardiac records through the use of linear, nonlinear, fractal and wavelet based methods. A comparative analysis was made of the results obtained in the evaluation of the HRV parameters in the three types of signals used. The difference between HRV time series (cardiac intervals and HRV analysis) obtained by examination of individuals diagnosed with heart failure and healthy individuals is graphically presented. The findings indicate that studies of heart rate variability on ECG, Holter and PPG records can be used to support the cardiac practice of physicians.**

***Keywords*—Heart rate variability; cardiovascular diseases; mathematical analysis; holter records; software system**

I. Introduction

The research conducted in recent years has unequivocally shown that heart rate variability (HRV) reflects an individual's health status. HRV [1] presents the variation in time between successive heartbeats (time intervals between heart beats), which variation depends on internal and external conditions. Internal conditions include the work of the physiological systems of the human body, their effective interaction and the general health of man as a result. External conditions include the impact of external circumstances (temperature, weather, emotions, stress, etc.) on the body. HRV is the ability of the human body and in particular of the heart to adapt to constantly changing external circumstances through compensatory reactions.

Usually high values of HRV parameters are an indicator of good health and excellent regulation of the autonomic nervous system. Low values of HRV parameters are an indicator of deteriorating health and disorders in the regulation of the body. The study of many quantitative characteristics of HRV makes it possible to assess how the variability in the parameters of the cardiac series [2] reflects the response of the human body to internal and external factors affecting its physiological and mental health.

HRV can be used to assess the regulation of the autonomic balance of the human body, the work of the heart system and the condition of blood vessels, blood pressure, the work of the digestive system, the nervous system and others.

Heart rate variability can be measured by various mathematically based methods: linear and nonlinear methods, methods based on fractal and wavelet theory, Detrended fluctuation analysis (DFA), Poincare method, methods for estimating the Hurst parameter and many others.

The article uses 3 types of cardiac signals: holter recors, write down from monitoring device for continuous observation; photoplethysmographic (PPG) and electrocardiographic (ECG) recorded with PPG device, capable of simultaneous recording of ECG and PPG signals. When recording the input data, the holter device works in parallel with the PPG device.

In ECG and Holter recordings, heart rate variability parameters are evaluated on RR intervals (determined by the input values of ECG and Holter recordings, R is the maximum point in QRS), and in PPG signal evaluations are performed on PP intervals (P peaks - maximum amplitude deviations of input signal).

The purpose of this article is to present the results of the project "Investigation of the application of new mathematical methods for the analysis of cardiac data", funded by the Research Fund, obtained with a demonstration software system for research and mathematical analysis of cardiac signals and data (obtained by Holter device, and developed demonstration device for write down of ECG and PPG signals). The paper presents the results obtained in time analysis, frequency analysis, DFA and calculation of the Hurst parameter of heart rate variability on three investigated types of real cardiac signals: ECG, Holter records and PPG.

The aim of the article is to present research on the cardiovascular system, made by means of a PPG device developed by the authors and a software system for processing and analysis of cardio data. The software system is designed to work with 3 types of cardio data: PPG, ECG and Holter records, performing preprocessing in accordance with the specific type of cardio data.

The article raises and seeks a solution to the following questions:

- Are the use of the three types of examined cardio (ECG, PPG and Holter records) equally effective in the assessment of HRV and in the analysis of the parameters of HRV data of healthy and sick individuals.
- The problem with the difference of HF parameters in patients with heart failure from those of the studied control healthy group was studied.

- Can PPG, ECG and Holter records be used to differentiate between healthy and sick individuals (and in particular patients diagnosed with heart failure)?

The rest of this document is organized as follows:

Section II provides an overview of related research in the scientific literature on heart rate variability. Section III focuses on the Methodology, HRV data analysis procedures used in this paper (performing mathematical analysis in time domain, frequency domain; fractal analysis and wavelet analysis). Section IV presents the results of the performed experimental analyzes in numerical and graphical form. Finally, Section V presents the conclusions drawn from the results obtained and Section VI presents the direction of future work and perspectives.

II. Review of HRV Studies

Healthy individuals are characterized by good heart rate variability, which results from the body's internal forces to adapt to environmental challenges (physical and psychological). The functioning of the cardiovascular system of a healthy human body can be described by complex mathematical models based on variability in the action of the heart.

HRV is a method for determining the work of the heart, which has been the subject of extensive scientific research over the last two decades on both healthy and diseased individuals. Despite numerous publications on this subject, the method for mathematically based analysis of heart rate variability has not yet been well studied. New methods for the study of HRV are emerging, which need extensive research before being adopted in the daily practice of cardiologists and to be standardized. Particularly valuable in this regard are the findings made on the studied real cardiac records of patients with various heart diseases.

The authors of [3] emphasize the importance of the age and sex of the subjects and the duration of the records made for the results of the HRV analysis.

The research, conducted by Murthy et al. [4] by spectral analysis on PPG records of 5 patients (with atrial fibrillation and myocardial infarction) and 5 healthy individuals showed that PPG signals can be used in the analysis of heart disease. PPG signals were recorded from the earlobes from the earlobes of the subjects.

In publication [5] the authors record and analyze ECG and PPG signals, the recording of which (within 5 minutes) is done simultaneously in time. HRV estimates were performed in the time and frequency domains, and nonlinear mathematical analyzes were performed. However, the recording of the signals is made in ideal conditions (without movement of the studied individuals).

The study done in [6] proposes to improve the efficiency of localization of the maximum deviations in the photoplethysmographic signal through a probabilistic approach based on Bayesian training.

The authors of [7] propose an algorithm based on continuous wave transform (CWT) to detect the maximum deviations of the PPG signal. The algorithm also uses a combination of functions obtained from the applied wavelet transform and indicators evaluating the self-similarity of PPG signals to detect damaged areas of the studied signal.

Studies in [8] compare PPG and ECG signals in terms of determining the HRV time series and accuracy when working with both types of signals. The authors propose a modified algorithm for detecting PP peaks in PPG signals. The obtained results show low statistical errors in determining the HRV time series for both types of studied signals.

Fractal methods for HRV analysis are used by the authors of [9], studying patients with congestive heart failure (CHF) and healthy subjects. The data used are taken from a public database (Physiobank).

The effect of artifacts in cardiac signals (obtained with PPG sensors) caused by movement on HRV and its evaluation using the statistical parameters SDNN and RMSSD was studied in [10]. The authors examined cardiac data obtained from devices worn on the wrist by 22 young and healthy individuals.

A study on 50 healthy young volunteers was presented in [11]. The HRV study was performed on simultaneously recorded photoplethysmographic and electrocardiographic signals using a mobile device. The presented HRV results were obtained using popular public Kubios software.

III. HRV Data Analysing Procedures

The procedures used in the paper for pre-processing and mathematical analysis of the input cardiac data are presented in Fig. 1. The preprocessing includes:

- ✓ Reduction of signal interference, noise and so on;
- ✓ Determination of maximum deviations (R peaks (QRS complexes) in ECG and Holter signals, and P peaks in PPG signals);
- ✓ HRV time series obtained;
- ✓ Determination of normal to normal (NN) intervals from RR intervals time series or PP intervals time series.

The detection of the main points in the cardiac signals - the points with maximum amplitude deviations (R peaks in ECG signals and P peaks in PPG signals) [12] is the starting point in the HRV analysis.

Normal NN intervals are obtained by excluding abnormal strokes such as ectopic strokes from RR/PP intervals (outside the right atrium’s sinoatrial node).

Each of the three types of signals studied in the paper (Holter records, ECG, and PPG signals) has its advantages and disadvantages compared to other ways of recording the activity of the heart. Therefore, each case of treatment and prevention of a patient should be considered separately and to determine which method of HRV testing is most appropriate.

ECG signals are obtained in a non-invasive, popular, and widespread way. Electrocardiograms (in cases when they are recorded in an inpatient setting in a polyclinic or hospital) require the placement of several electrodes on the human body in specific places and each of them must be firmly fixed. This

is uncomfortable with frequent use of this method (for example, for daily measurements or if several measurements per day are required).

Long-term monitoring of the heart (if you need 24 hours of records and longer for a continuous monitoring of heart activity in risk groups of individuals) is appropriate to perform with a Holter device [13].

Photoplethysmographic signals are an alternative to ECG signals, they are easier to record, PPG devices [14, 15], through which these signals are recorded, are light and comfortable for longer carrying. PPG sensors are small and can be easily integrated into various lightweight and easily portable devices, smart phones and smart watches. In the last few years, with the improvement of the technology for production and miniaturization of PPG sensors, PPG technology enters the daily life of more and more people and becomes an integral part of it. Photoplethysmography determines the time between heartbeats by continuous monitoring of changes in blood volume in part of the peripheral microvasculature [16]. This non-invasive method for measuring pulse waves can also be the basis for HRV analysis.

Determining the health status of an individual is an issue that more and more people are interested in today. Advances in technology have led to an increasing miniaturization of digital sensors, which has led to an increase in the possibility of more accurate and easy continuous monitoring of individuals (eg blood pressure monitoring [17], a parameter that is crucial for the development of a number of diseases) if necessary with the help of mobile devices.

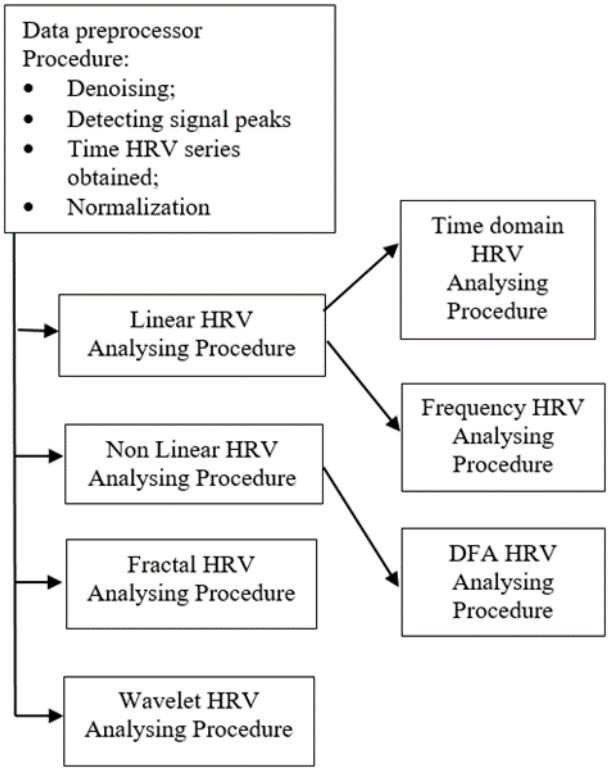


Fig. 1. HRV Analysing Prosedures.

One of the disadvantages of PPG signals is their influence by artifacts [18], which are obtained during the movement of the studied individual. This may adversely affect the HRV score if inaccuracies are found in the localization of the peaks in the PPG signal. In addition, the physiological processes carried out with the help of the heart and the influence of human skin (which may have different characteristics in individual individuals) make it difficult to study HRV in PPG signals.

In this paper, studies of the three types of signals (ECG, Holter, PPG) for recording heart activity were made in terms of assessing heart rate variability.

The article examines HF on short-term cardiac records (2 to 24 hours). Short-term HRV is affected by the cardiovascular system, central nervous system, respiratory system, baroreceptors (blood pressure sensors) and others. Nonlinear estimates of HRV quantify the unpredictability of cardiac time series.

A. HRV Time Domain Analysing Procedure

The parameters in the time domain give numerical expression of the quantitative characteristics of HF for time periods from 2 to 24 hours [19]. The following statistical parameters in time domain were examined in the present study [20]:

- Mean RR (Mean PP) - the mean value of RR and PP intervals;
- SDNN - standard deviation (sd) of normal RR (PP) intervals;
- SDANN - standard deviation of the average normal RR (PP) intervals for each 5 min segment of a 24 h record;
- RMSSD - square root of the mean squared differences between successive RR (PP) intervals;
- SDindex - mean of the standard deviation of all normal RR (PP) for each 5 min segment of the whole record.

The calculations of the presented parameters are performed using the HRV time domain analyzing procedure of the demonstration software system.

B. HRV Frequency Domain Analysing Procedure

The parameters in the frequency domain provide a quantitative assessment of the complexity of the model and the low ability to predict the values of the cardiac series.

The signal energy (Power) in the respective frequency band is determined by the HRV frequency domain procedure. Measurements in the frequency domain show the distribution of absolute power (in ms^2) and relative power (in normal units (nu)) in the studied three frequency bands (Very Low Frequency, Low Frequency and High Frequency).

Frequency domain parameters are based on spectral analysis for the following three components (for short-term records) presented in Table I.

The sum of the three powers (VLF, LF and HF) for short-term recordings gives the total signal power [21].

TABLE I. HRV PARAMETERS IN FREQUENCY DOMAIN

Power [ms^2]	Frequency range [Hz]	Interaction with the systems of the human body
Very Low Frequency (VLF)	0.003-0.04	Sympathetic nervous system
Low Frequency (LF)	0.04-0.15	Sympathetic and parasympathetic nervous system
High Frequency (HF)	0.15-0.4	Parasympathetic nervous system and respiratory sinus arrhythmia.
LF/HF	-	Gives an assessment of the sympathetic balance

The ratio (LF/HF) between LF and HF band powers, known as the sympathetic balance index, assesses the ratio between the activity of the two parts of the nervous system (sympathetic and parasympathetic).

C. HRV Fractal Analysing Procedure

The fractal properties of the three types of studied cardiological data in this paper are determined by analyzing the fluctuations in the time series through the parameters Alpha1, Alpha2 and Hurst.

Detrended fluctuation analysis examines the correlations between RR (PP) interval series in different time scales. DFA defines two parameters: Alpha1 (α_1) – describes short-term fluctuations, Alpha2 (α_2) - provides information on the long-term fluctuations of the studied signal.

The traditional R/S statistical method, most often used in the scientific literature, is used to determine the Hurst parameter in the studied cardiological data [22, 23].

D. HRV Wavelet Analysing Procedure

In the present study, wavelet-based method are used to determine the Power Spectral Density (PSD) of heart rate variability. A graphical method was chosen for the study of global PSD, which allows for visual comparison of the spectral properties of the analyzed signals.

To account for the differences between the HRV parameters determined on the time series of the studied three types of signals, the root mean square error (MSE) is calculated:

$$MSE = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \quad (1)$$

Where: x – time series of one signal;

y - time series of the second signal;

N- the number of intervals.

IV. RESULTS

Twenty-four individuals diagnosed with Heart Failure were studied (10 men and 14 women aged 52.8 ± 1.6). Records were also made of 12 healthy volunteers without cardiac disease.

The following recordings were made: ECG, Holter and PPG signals of the subjects. Holter recordings are made with a holter monitoring device purchased for the purposes of the research project. ECG and PPG signals are recorded with a multisensor device created according to the scientific project. The created demonstration software determines the RR intervals in the ECG signals and the PP intervals in PPG signals.

In Fig. 2 shows the RR intervals obtained from an ECG signal recorded using a multifunctional PPG device of an individual diagnosed with Heart Failure.

In Fig. 3 presents the RR intervals obtained from the Holter record recorded using the Holter Heart Failure Patient Monitoring Device.

Fig. 4 shows the PP intervals obtained from a PPG signal recorded using the portable PPG device of an individual diagnosed with Heart Failure.

Fig. 5 shows the PP intervals obtained from a PPG signal (via a PPG device) of a healthy individual. The comparison of Fig. 4 and Fig. 5 shows greater variability in the determined PP intervals, which is also proved by the calculated values of the HRV parameters using the methods used in the study (presented in Table II).

All made records of the three types of cardiac signals shown in the figures and participating in the present study were two hours.

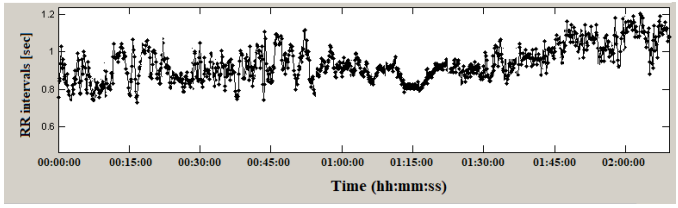


Fig. 2. RR Intervals (ECG Signal, Heart Failure).

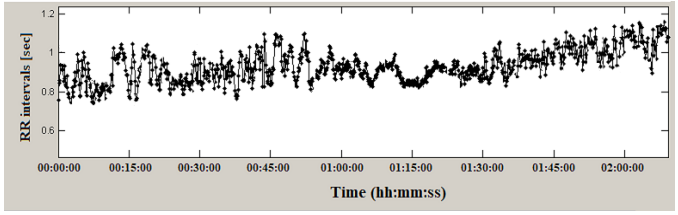


Fig. 3. RR Intervals (Holter Signal, Heart Failure).

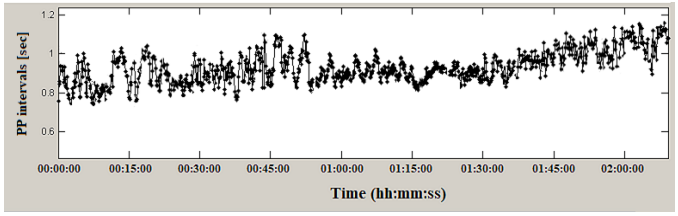


Fig. 4. PP Intervals (PPG Signal, Heart Failure).

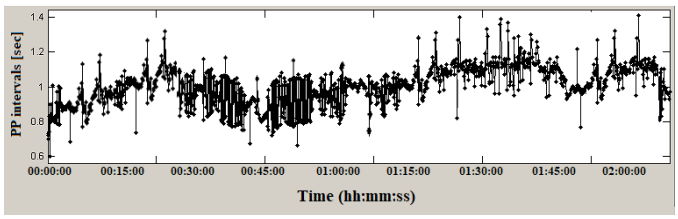


Fig. 5. PP Intervals (PPG Signal, Healthy Individual).

TABLE II. HRV PARAMETERS FOR ECG, HOLTER AND PPG

Parameters		Group 1 ECG (mean± sd)	Group 2 Holter (mean± sd)	Group 3 PPG (mean± sd)
Time Domain Analysis	MeanRR(PP) [ms]	684.22 ±214.68	661.33 ±189.13	692.11 ±223.83
	SDNN [ms]	124.08 ±16.88	118.77 ±24.32	132.66 ±32.09
	SDANN [ms]	98.56 ±34.21	112.01 ±35.43	104.67 ±31.08
	RMSSD [ms]	28.18±8.65	26.35 ±14.15	31.06 ±18.98
	SDindex [ms]	61.33 ± 26.11	64.07 ±22.18	63.88 ±26.44
Frequency Domain Analysis	Power VLF [ms^2]	3098.51 ±654.22	3127.06 ±487.34	2995.78 ±586.39
	Power LF [ms^2]	688.22 ±183.06	691.89 ±243.99	704.05 ±433.01
	Power HF [ms^2]	586.23 ±204.55	582.99 ±244.13	602.33 ±212.03
	Power LF [nu]	0.54±0.19	0.54±0.16	0.53±0.87
	Power HF [nu]	0.46±0.23	0.46±0.43	0.47±0.68
	LF/HF ratio	1.17±0.78	1.19±0.81	1.17±0.93
DFA	Alpha1	1.22±0.24	1.18±0.19	1.23±0.48
	Alpha2	1.28±0.27	1.21±0.22	1.17±0.31
R/S method	Hurst	0.72±0.26	0.74±0.71	0.68±0.14

Table II presents the obtained results of the HRV analysis in Time Domain, Frequency Domain, DFA and hurst parameter (R/S method) obtained from a study of Heart Failure Individuals (n=20).

The results presented in Table II show lower mean values for all three types of cardio signals examined on the frequencies in the Low Frequency band in Power in the indicators of patients diagnosed with Heart Failure compared to the values of healthy patients (691.89 ± 243.99 ms^2 for Holter records versus 1170 ± 416 ms^2 normal values (for healthy people) given in the HRV Standard [20]). The Power HF values for the three types of cardio signals tested were also lower than those recommended in the HRV Standard for Healthy Individuals (582.99 ± 244.13 ms^2 for Holter records versus 975 ± 203 ms^2 normal values given in [20].

The calculated LF/HF ratio, giving information about the balance between Low Frequency band and High Frequency band, has a value of 1.19 ± 0.81 (against 1.5-2.0 recommended value for healthy people [20]).

In the time domain, the following studied parameters show lower values than those recommended for healthy individuals according to [20]: SDNN (max 132.66 ±32.09 ms for PPG against 141±39 ms in [20]); SDANN (max 112.01 ±35.43 ms for Holter records against 127±35 ms in[20]).

Fig. 6 shows the DFA results for Heart Failure Individual. The two studied parameters are alpha1 (shown in cyan color) and alpha2 (shown in red color). The obtained parameters for short correlation values are alpha1=1.26 and for long correlation alpha2=1.28. Fig. 6 shows that alpha1 and alpha2 have similar values in diseased individuals. This shows that in heart disease there is a tendency for short-term and long-term correlations in the HRV series to equalize.

Fig. 7 shows the DFA results for a Healthy Individual. The obtained values for the short correlations are as follows: alpha1 =0.87 and long correlations alpha2=1.24. Fig. 7 shows higher values for the alpha2 parameters relative to alpha1.

Fig. 8 presents a global PSD, drawn using a wavelet-based graphical method of one of the studied Heart Failure Individual (on a record of cardiac data made with a Holter device). The graph shows relatively low values of signal power in all three studied frequency bands (VLF, LF and HF).

Fig. 9 presents a global PSD, drawn by means of a wavelet-based graphic method of one of the examined volunteers without disease (the cardiological record was made with a Holter device). The graph shows high values of signal power in each of the three frequency bands (VLF, LF and HF).

The comparative analysis of the presented global PSDs, shown graphically in Fig. 8 and Fig. 9 shows a decrease in the power of the studied cardiac signal in the studied three frequency bands. This is an indicator of reduced heart rate variability in individuals with Heart Failure.

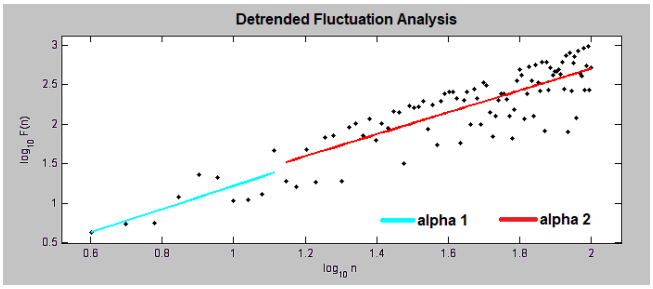


Fig. 6. DFA for Heart Failure Individual.

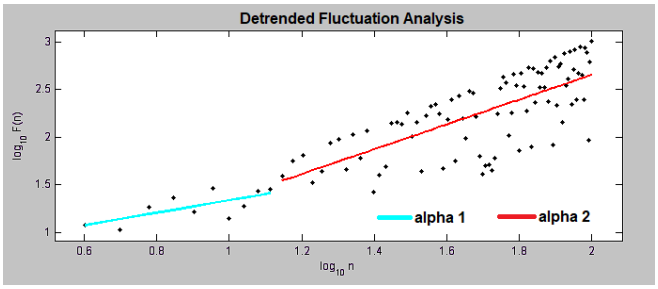


Fig. 7. DFA for an Individual without Disease.

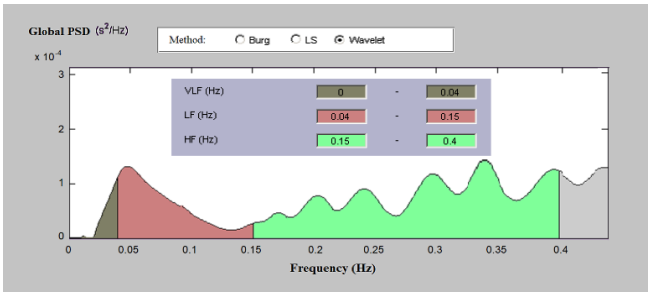


Fig. 8. PSD for Heart Failure Individual.

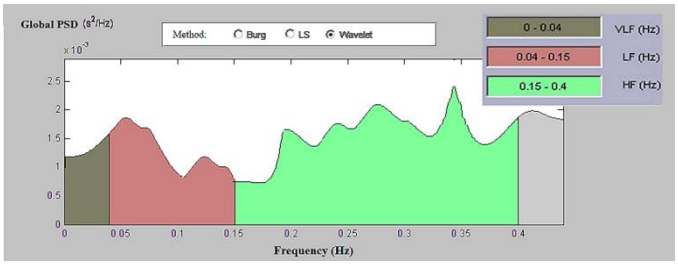


Fig. 9. PSD for an Individual without Disease.

Table III shows the mean squared error calculated by formula (1) for the signals ECG-Holter (MSE_{G1-G2}), ECG-PPG (MSE_{G1-G3}) and Holter-PPG (MSE_{G1-G3}). The smaller the MSE parameter show the closer the two studied time series have values, from which it follows that the two studied methods for determining the cardiac interval give similar results.

TABLE III. RELATIVE ERROR BETWEEN ECG, HOLTER, PPG RECORDS

Parameters		MSE_{G1-G2} **[%]**	MSE_{G1-G3} **[%]**	MSE_{G2-G3} **[%]**
Time Domain Analysis	MeanRR(PP) [ms]	1.34	3.31	0.6
	SDNN [ms]	0.64	1.47	0.88
	SDANN [ms]	1.49	0.69	0.83
	RMSSD [ms]	2.27	3.13	5.27
	SDindex [ms]	4.01	3.18	3.58
Frequency Domain Analysis	Power VLF [ms^2]	2.96	4.92	5.93
	Power LF [ms^2]	3.04	4.07	1.69
	Power HF [ms^2]	4.33	6.71	2.78
	Power LF [nu]	0.04	1.65	1.97
	Power HF [nu]	0.1	1.37	2.06
	LF/HF ratio	0.49	0.08	1.02
DFA	Alpha1	2.44	1.36	3.9
	Alpha2	0.88	1.11	0.82
R/S method	Hurst	0.67	1.03	0.94

From the comparative analysis of the studied data pairs it follows that the relative errors for all examined parameters are less than 4.33% for the ECG-Holter pair, less than 6.71% for the ECG-PPG pair and less than 5.93% for the Holter-PPG pair. The calculated relative errors are small and we can assume that the results obtained in the study of heart rate variability through the three types of signals studied are similar and reliable.

The presented numerical and graphical results are obtained through a demonstrative software system created in the MATLAB software environment.

V. DISCUSSION AND CONCLUSION

The article aims to consider the possibility of interchangeable use of three types of cardio signals in the study of heart rate variability. To solve this problem, MSE errors were determined between the evaluation parameters of each of the three studied types of cardio signals. The results (Table III) show low values of the calculated MSE errors (maximum value 6.71 for MSE_{G1-G3} in determining of Power HF), which is an indicator of the ability to use for correct research each of the three types of cardio signals (ECG, PPG, Holter records). This conclusion is of practical importance, as it proves the possibility of using PPG signals, which have recently become more and more common in human everyday life, for correct analyzes for health purposes.

The other purpose of the article is to examine the possibility of distinguishing healthy individuals from sick cardiovascular patients (more specifically, patients diagnosed with heart failure). The results presented in Table II show that the indicators Power LF, Power HF and LF/HF differ significantly from the corresponding indicators given in the HRV Standard, which is an indicator of the ability to differentiate between healthy controls and patients with heart failure.

The graphical representation of the alpha1 and alpha2 indicators obtained from the application of DFA visually show that these two parameters have similar values in diseased individuals (Fig. 6). The graphical representation of these parameters in Individual without disease (Fig. 7) shows higher values for the alpha2 parameters relative to alpha1. This finding was confirmed for all 24 individuals in the study diagnosed with Heart Failure.

The graphical representation of Global PSD shows low values of global PSD for Heart Failure Individuals (Fig. 8) compared to the values of this parameter in Individuals without disease (Fig. 9). Studies have therefore shown the possibility of graphically distinguishing patients from healthy individuals.

The article presents software procedures for determining heart rate variability, based on mathematical methods for studying three types of signals: ECG, Holter records, PPG. The presented software procedures perform a study of the parameters of heart rate variability in the time domain, frequency domain, apply DFA and determine the hurst parameter of the studied time series.

The parallel analysis of the studied three types of signals shows similar results in the study of heart rate variability and therefore the three methods for HRV analysis can be used equally. The choice of the specific method can be made according to each specific individual case.

The obtained numerical and graphical results show reduced variability of heart rate in the studied individuals with Heart Failure (for example for Holter records Power LF is 691.89 ±243.99 ms^2, Power HF is 582.99 ±244.13 ms^2) compared to healthy individuals. The coefficient showing the state of balance between LF and HF (1.19 ± 0.81) also has lower values in patients with arrhythmia compared to this coefficient in healthy people (1.5-2.0).

The studies in this paper were made on real cardiac records (three types of data were taken: ECG, Holter and PPG) of patients diagnosed with heart failure by a cardiologist as well as on several healthy volunteers.

The use of PPG signals to assess HRV and continuous monitoring of cardiac activity in patients with heart disease in need of long-term monitoring has the following advantages: easy, patient-friendly measurement; efficient signal processing and HRV, use of measuring devices that have increasingly popular and affordable hardware and software solutions; low price and convenience in their use.

VI. Future Work

Procedures for mathematical analysis of the three types of studied HRV signals are to be created with other nonlinear methods, as well as with wavelet-based methods for determining and detailed study of numerical parameters. The results of this study indicate the possibility of a study of HRV parameters in patients with other heart diseases (myocardial infarction, ischemic heart disease, syncope) and creation of an information database for patients.

Acknowledgment

This research work was carried out as part of the scientific project “Investigation of the application of new mathematical methods for the analysis of cardiac data” No KP-06-N22/5, date 07.12.2018, funded by the National Science Fund of Bulgaria (BNSF).

References

[1] U.R., Acharya, Suri, J.S., Spaan, J.A.E., Krishnan, S.M. “Advances in Cardiac Signal Processing”. Springer: Berlin. 2007.

[2] G. Ernst. “Heart Rate Variability”. Springer-Verlag London, 2014.

[3] F. Shaffer and J. P. Ginsberg. „An Overview of Heart Rate Variability Metrics and Norms“. Frontiers in Public Health. 2017. Vol.5:258 (pp.1-17). https://doi.org/10.3389/fpubh.2017.00258. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/.

[4] V.S. Murthy, S. Ramamoorthy, N. Srinivasan, S. Rajagopal, M.M. Rao. “Analysis of photoplethysmographic signals of cardiovascular patients”. Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2204-2207, 2001. https://ieeexplore.ieee.org/document/1017209, DOI: 10.1109/IEMBS.2001.1017209.

[5] G. Lu, F. Yang, J. A. Taylor, J. F. Stein. “A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects”. Journal of Medical Engineering & Technology. Vol. 33, No. 8, 2009, pp. 634-641, https://www.tandfonline.com/doi/abs/10.3109/03091900903150998?jou, rnalCode=ijmt20 https://doi.org/10.3109/03091900903150998.

[6] A. Alqaraawi, A. Alwosheel, A. Alasaade. “Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach”. Healthcare Tech. Letters. Vol. 3, No. 2, pp.136-42, 2016. DOI: 10.1049/htl.2016.0006. https://pubmed.ncbi.nlm.nih.gov/27382483/.

[7] A. Neshitov, K. Tyapochkin, E. Smorodnikova, P. Pravdin. “Wavelet Analysis and Self-Similarity of Photoplethysmography Signals for HRV Estimation and Quality Assessment”. Sensors (Basel). 2021 Vol. 13, No.21(20):6798. DOI: 10.3390/s21206798.

[8] D. Janković, R. Stojanović. “Flexible system for HRV analysis using PPG signal”. In: Badnjevic A. (eds) CMBEBIH, 2017, pp. 705-712. IFMBE Proceedings, vol 62. Springer, Singapore. https://doi.org/10.1007/978-981-10-4166-2_106.

[9] S. Kuntamalla, R. G. R. Lekkala “Reduced Data Dualscale Entropy Analysis of HRV Signals for Improved Congestive Heart Failure Detection”. Measurement Science Review, Vol. 14, No. 5, pp. 294-301, 2014. https://www.sciendo.com/article/10.2478/msr-2014-0040.

[10] A. Rossi, D. Pedreschi, D. A. Clifton and D. Morelli. “Error Estimation of Ultra-Short Heart Rate Variability Parameters: Effect of Missing Data Caused by Motion Artifacts”. Sensors 2020, 20(24), 7122; https://doi.org/10.3390/s20247122.

[11] Ch. K. K., M. Manaswinib, K.N.Maruthyc, A.V.S. Kumard, K. M. Kumar. “Association of Heart rate variability measured by RR interval from ECG and pulse to pulse interval from Photoplethysmography”. Clinical Epidemiology and Global Health. Vol. 10, 2021, 100698. https://www.sciencedirect.com/science/article/pii/S2213398421000026.

[12] B.S. Chandra, C. S. Sastry, S. Jana. “Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion”. IEEE Trans. Biomed. Eng. Vol. 66, No. 3, pp.710-7, 2019.

[13] T. Todorov, G. Bogdanova, N. Noev, N. Sabev. „Data management in a Holter Monitoring System“ TEM Journal, Vol. 8, No.3, pp. 801-805. 2019.

[14] D. J. Plews, B. Scott, M. Altini, M. Wood, A. E. Kilding, P. B. Laursen. “Comparison of heart-rate-variability recording with smartphone photoplethysmography, Polar H7 chest strap, and electrocardiography”. International Journal of Sports Physiology and Performance, Vol. 12, No. 10, pp.1324-28, 2017. DOI: 10.1123/ijspp.2016-0668. https://pubmed.ncbi.nlm.nih.gov/28290720/.

[15] S. Botman, D. Borchevkin, V. Petrov, E. Bogdanov, M. Patrushev, N. Shusharina. „Photoplethysmography-Based Device Designing for Cardiovascular System Diagnostics. International Journal of Biomedical and Biological Engineering“. vol. 9(9), pp. 689-693. 2015.

[16] Clint R. Bellenger, Dean Miller, Shona L. Halson, Greg Roach and Charli Sargent.”Wrist-Based Photoplethysmography Assessment of Heart Rate and Heart Rate Variability: Validation of WHOOP”. Sensors. 2021, 21, 3571. https://doi.org/10.3390/s21103571.

[17] M. Elgendi, R. Fletcher, Y. Liang, N. Howard, N. Lovell, D. Abbott, K., Lim, and R. Ward. „The use of photoplethysmography for assessing hypertension“. Npj Digital Medicine 2:60; 2019. https://doi.org/10.1038/s41746-019-0136-7.

[18] Th.Wittenberg, R. Koch, N. Pfeiffer, N. Lang, M. Struck, O. Amft, and Eskofier, B. "Evaluation of HRV estimation algorithms from PPG data using neural networks" Current Directions in Biomedical Engineering, Vol. 6, No. 3, 2020, pp. 505-509. https://doi.org/10.1515/cdbme-2020-3130.

[19] S. Siecinski, P. S. Kostka and E. J. Tkacz. “Heart Rate Variability Analysis on Electrocardiograms, Seismocardiograms and Gyrocardiograms on Healthy Volunteers”. Sensors. 20, 4522. 2020, DOI:10.3390/s20164522.

[20] M. Malik. “Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, Heart rate variability - standards of measurement, physiological interpretation, and clinical use”. Circulation. 1996. Vol.93, pp.1043–1065. Available: https://www.escardio.org/static_file/Escardio/Guidelines/Scientific-Statements/guidelines-Heart-Rate-Variability-FT-1996.pdf.

[21] F. Shaffer, R. McCraty, C.L. Zerr. “A healthy heart is not a metronome: an integrative review of the heart’s anatomy and heart rate variability”. Front Psychol. 2014, Vol. 5:1040 (pp. 1-19), https://doi.org/10.3389/fpsyg.2014.01040.

[22] U.R. Acharya, K.P. Joseph, N. Kannathal, C.M. Lim, J.S. Suri, “Heart rate variability: a review”, Med. Bio. Eng. Comput, Vol. 44, pp.1031-1051, 2006.

[23] B. Malia, S. Zuljb, R. Magjarevic, D. Miklavcic, T. Jarma, “Matlab-based tool for ECG and HRV analysis”, Biomedical Signal Processing and Control, 2014, Vol. 10, pp. 108-116. http://dx.doi.org/10.1016/j.bspc.2014.01.011.

Multistage Relay Network Topology using IEEE802.11ax for Construction of Multi-robot Environment

Ryo Odake, Kei Sawai, Noboru Takagi, Hiroyuki Masuta, Tatsuo Motoyoshi
Graduate School of Engineering, Toyama Prefectural University
Toyama, Japan

***Abstract*—This paper describes an information gathering system comprising multiple mobile robots and a wireless sensor network. In general, a single robot searches an environment using a teleoperation system in a multistage relay network while maintaining communication quality. However, the search range of a single robot is limited, and it is difficult to gather comprehensive information in large-scale facilities. This paper proposes a multistage relay network topology using IEEE802.11ax for information gathering by multi-robot. In this multi-robot operation, a mobile robot carries wireless relay nodes and deploys them into the environment. After a network is constructed, each robot connects to this network and gathers information. An operator then controls each robot remotely while monitoring its end-to-end communication quality with each mobile robot in the network. This paper proposes a method assuming the end-to-end throughput with multiple mobile robots. The validity of the proposed method is then inspected via an evaluation experiment on multi-robot teleoperation. The experimental results show that the network constructed with the proposed topology is capable of maintaining the communication connectivity of more than three mobile robots.**

Keywords*—Multi-robot system; IEEE802.11ax; information gathering; multistage relay network; network topology

I. Introduction

After a disaster, disaster reduction activities are performed in the affected area to prevent the damage from spreading. In the implementation of disaster reduction activities, information should be gathered to determine the damage status [1-2]. Helicopter aerial photographs, existing infrastructure (such as surveillance cameras), and rescue teams can be used to gather such information [3-6]. However, in enclosed spaces, such as tunnels and underground malls, it is difficult to gather information using helicopters. In some cases, existing infrastructure cannot be used due to malfunction or lack of power supply. Moreover, there is a risk of injuring humans or inducing secondary disasters during information gathering by rescue teams. Therefore, the use of mobile robots is widely considered to gather information in enclosed spaces after disasters [7-11].

Two communication methods are adopted for mobile robots: wired and wireless. Wired communication maintains a stable communication quality and power supply from cables [12]. However, cables may be disconnected and communication with the mobile robots can be interrupted when cables become tangled with obstacles or the wheels of the mobile robot. Wireless communication offers high runability, as there is no physical restriction by cables [13-14]. However, in wireless communication, mobile robots may become isolated when radio waves are hindered by obstacles. Therefore, in an enclosed space after a disaster, it is necessary to use the communication method that best matches the purpose and the situation of the disaster area [15-16]. This paper discusses a method for gathering information using wireless communication in environments where it is difficult to operate a mobile robot using wired communication.

Robot wireless sensor networks (RWSNs) involve the teleoperation of mobile robots using wireless communication [17-19]. In an RWSN system, a mobile robot expands its search range by deploying a relay node called a sensor node (SN) in its path (Fig. 1). In a network using a multistage relaying such as RWSNs, the communication quality decreases as the number of relays and the distance between nodes increase. Therefore, it is difficult to maintain the communication connectivity of mobile robots in a multistage relay network, and single robots are mainly operated. Moreover, the search range of a single robot in a large-scale facility is limited. This paper proposes a multistage relay network topology for constructing a multi-robot environment. The experiment in this paper constructed a multistage relay network with the proposed topology and measured the communication quality to confirm the validity of the proposed topology for constructing a multi-robot environment. This experiment used a bandwidth compression throughput to measure the communication quality at the packet level.

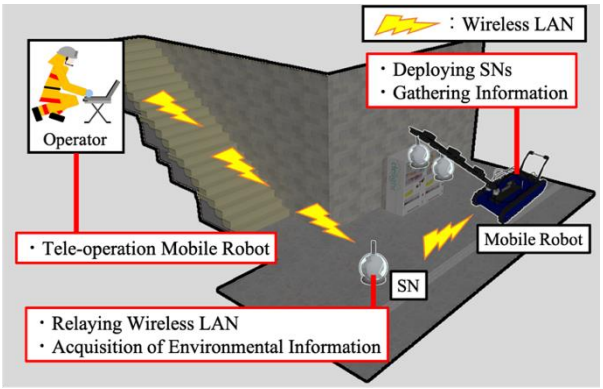


Fig. 1. RWSN.

II. Teleoperation Method for Mobile Robots using Multistage Relay Network

In the teleoperation of mobile robots by wireless communication, the search range is limited due to radio-wave damping. Therefore, wireless communication is not effective for gathering information in large-scale facilities. An RWSN relays SNs to expand the search range of a mobile robot. It also extends the network construction range by teleoperating the mobile robot with each node relayed in advance (Fig. 2). This allows the operator to teleoperate the mobile robot without the risk of a decrease in communication quality due to the increase in the number of relays. An RWSN deploys SNs in consideration of the communication quality between nodes, thereby constructing the network flexibly according to the environment. Therefore, RWSNs have scalability and flexibility, which make them effective for searching in environments with many obstacles.

However, in a multistage relay network, such as an RWSN, packets are processed for transfer within the SN. Hence, the transmission rate decreases and the packet reception interval become misaligned as the number of relays and the distance between nodes increase. Consequently, the mobile robot becomes less operable; it can become isolated due to moving out of the communication range by mistake or failing to maintain communication connectivity. Therefore, in an RWSN, it is common to operate a single robot to maintain the communication quality required for its teleoperation with the mobile robot. However, a single robot encounters limitations in searching and gathering comprehensive information in large-scale facilities. Given the importance of multi-robot systems for solving the problem of using RWSNs in large-scale facilities, and the paper aims to construct a multi-robot environment using a multistage relay network. The next section shows the flow of the multi-robot system using a multistage relay network and the requirements for the network to teleoperate multiple robots.

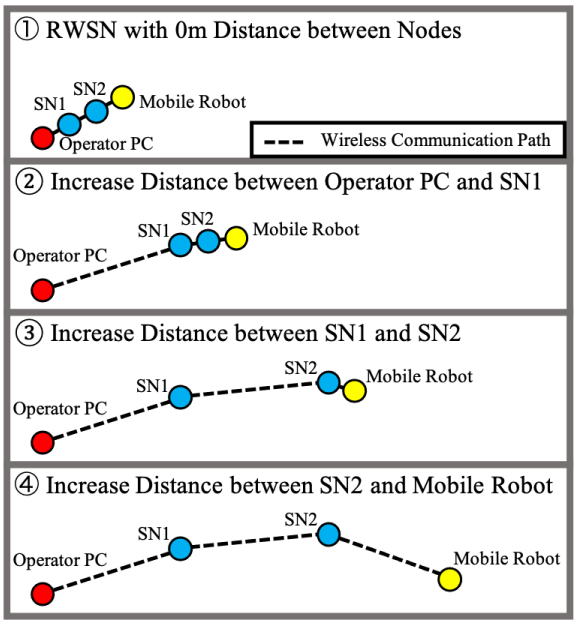


Fig. 2. Increasing Distance between Nodes while Maintaining Communication Connectivity in RWSNs.

III. Requirements of Multistage Relay Network

Let mobile robot <1> be the mobile robot that deploys the SN. The flow from the construction of the network to the operation of the mobile robots is shown in Fig. 3.

- The construction range of a multistage relay backbone network is expanded with the mobile robot <1> as shown in Fig. 2.
- Multiple mobile robots are connected to the constructed backbone network.
- Mobile robot <1> and the other mobile robots search within the network construction range while switching the nodes to be connected.

The requirements of the multistage relay backbone network for multi-robot operation are as follows:

1) Remote control of a total of three or more mobile robots
2) A throughput of 20.0 Mbps or higher in the communication path between each operator PC and each mobile robot.

About three mobile robots with five cameras each are needed to search 100 m of an enclosed space. Therefore, in requirement (1) set the number of teleoperated mobile robots to three. Moreover, the mobile robot used for searching in the nuclear power plant damaged by the Great East Japan Earthquake gathered information based on five camera images and required a throughput of 20.0 Mbps. Therefore, in requirement (2), the throughput required for the teleoperation of a single mobile robot is set to be 20.0 Mbps or higher.

Many wireless teleoperation systems for mobile robots are based on TCP/IP protocols. TCP/IP is highly suitable for the communication of mobile robots because most of the control system of a mobile robot involves a PC. Therefore, socket communication is often adopted for mobile robot communication, and information communication by packet transmission and reception is typical. Therefore, RWSNs adopt wireless LAN as their communication method. However, the theoretical value of IEEE802.11b/g used in RWSNs is 54 Mbps, which is an insufficient throughput for multi-robot operation. This study constructs a network using IEEE802.11ax, which has a theoretical value of about 1,200 Mbps. The next section describes the proposed multistage relay network topology, which takes into account the abovementioned requirements.

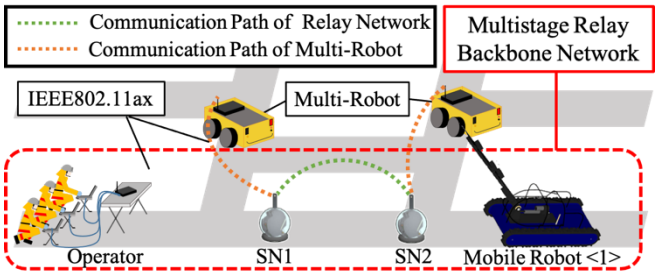


Fig. 3. Gathering Information by Multi-robot using Multistage Relay Network.

IV. Proposed Network Topology

A network topology is a form of connection that shows how devices such as operator PCs are connected to each other. From the flow of the multi-robot operation in Section 3, the multistage relay part of the proposed topology should be a static network and an RWSN to prevent any change in communication quality due to the dynamic communication path change. The proposed network topology for a multi-robot system is shown in Fig. 4. The devices of each node are in Fig. 5, and Fig. 6 depicts how the nodes are connected. In this network topology, each node in the multistage relay network is equipped with an access point (AP) and an adapter device; this creates a communication path between the operator PC and the mobile robot <1>. Then, the multi-robots are equipped with adapter devices, which are connected to the APs installed in each node of the multistage relay network to construct a communication path between each operator PC and each mobile robot. In this topology, a network is constructed for each node, and each robot can comprehensively search within the network construction range while switching the AP to which it is connected.

In case of SN failure or battery power insufficiency, it is necessary to replace the SN and reconstruct the multistage relay network. In this topology, the communication path between the operator PC and mobile robot <1> can be reconstructed with the same topology as that before the communication interruption by connecting the newly deployed SN to the AP of the node on the upload side that is still functioning.

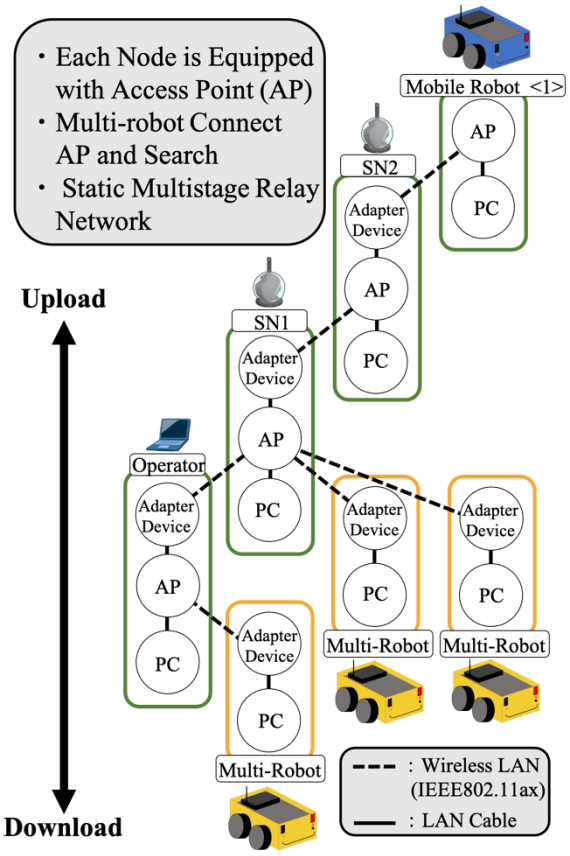


Fig. 4. Proposed Topology for Constructing a Multi-robot Environment.

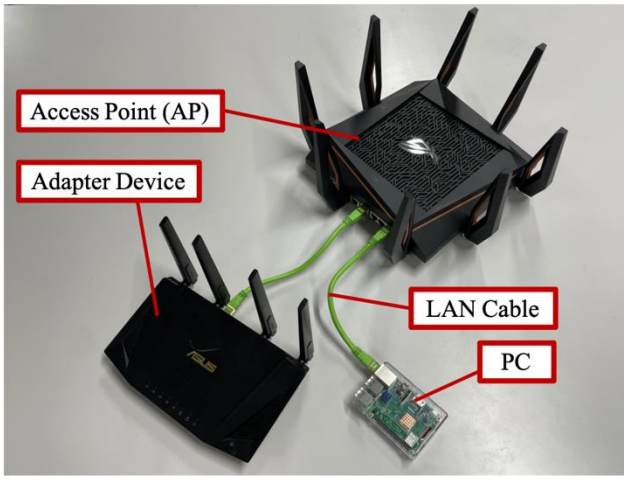


Fig. 5. Device of a Node Constituting the Multistage Relay Network.

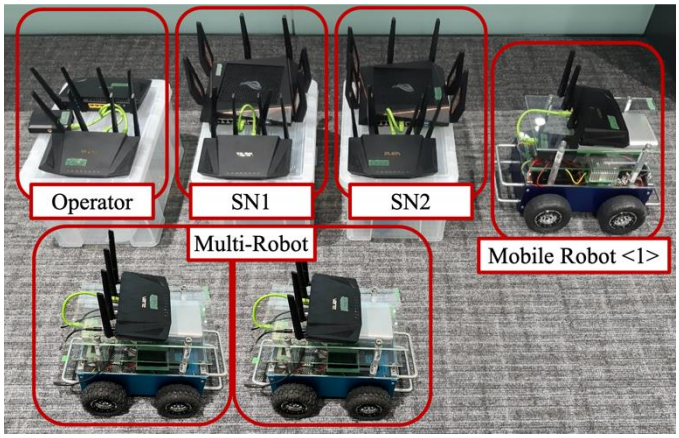


Fig. 6. Multistage Relay Network and Robots.

V. Evaluation of Communication Quality using Bandwidth Compression Throughput Measurement

This section define the communication quality characteristics that should be evaluated when teleoperating a mobile robot and explain how to measure them.

The operator receives packets containing camera and sensor information acquired by the mobile robots, and teleoperates the mobile robots (Fig. 7). Therefore, the transmission speed between the operator PC and the mobile robots should be monitored to maintain stability of the teleoperation of the mobile robots. This paper evaluates the transmission speed in a TCP/IP-compliant communication path as the throughput at the packet level. Throughput (bps) indicates the transmission speed received by the receiving PC per second. This system is expected to send a large amount of sensed information from the mobile robots. Therefore, this paper uses a band compression throughput measurement that can determine the upper limit of the transmission speed.

Bandwidth compression throughput represents the maximum number of data receivable per second by transmitting a large number of measurement packets from the transmitting side. Additionally, this measurement method send as many packets as the computer can process to measure the upper limit of the throughput accurately. The packet size of the

transmission packet is set to 1,400 bytes, which is the upper limit. In this system, the bandwidth compression throughput is calculated assuming that the packet size is *B* (byte), the total number of received packets is *n*, and the time required to complete the measurement is *t* (s). Then, the band compression throughput, *Th* (bps), can be expressed by the following equation:

$$Th = \frac{8Bn}{t} \quad (1)$$

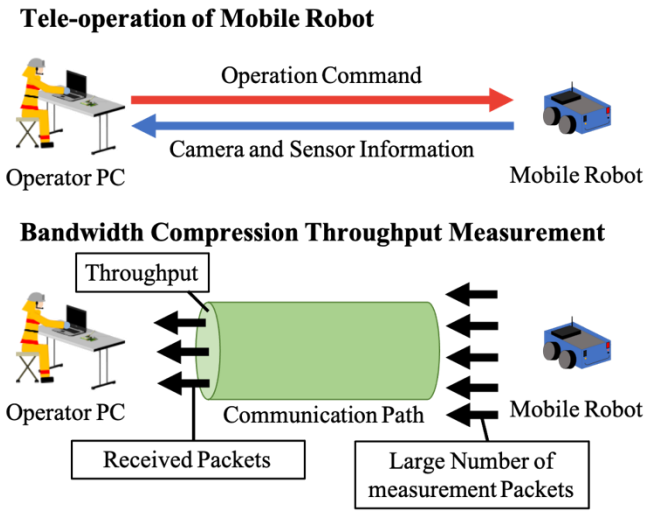


Fig. 7. Evaluation of Communication Quality of Teleoperation of Mobile Robots.

VI. Possibility of Constructing a Multi-robot Environment with Proposed Multistage Relay Network

A. Experiment on Increasing Distance between Nodes

This experiment examined the change in throughput with increasing distance between nodes to confirm whether such distance can be increased during network construction. This experiment measured the bandwidth compression throughput to determine the change in throughput with increasing distance between nodes. The distance between SN1 and SN2 was increased from 0 m to 5 m at 5 m intervals until communication was lost, and the bandwidth compression throughput was measured 10 times in both directions at each point. Then, this experiment loaded the network by sending 50,000 packets of 1,400 bytes. The upload direction was from SN1 to SN2, and the download direction was from SN2 to SN1, as shown in Fig. 4. A straight, paved 250 m road was used (Fig. 8). Furthermore, this experiment used a Raspberry Pi 4 Model B as the PC, ASUS RT-AX3000 as the router, and a CAT8 LAN cable.

Fig. 9 and 10 show the experimental results in the upload and download directions, respectively. The figures state the minimum, average, and maximum throughput values at each point. After the bandwidth compression throughput measurement in the download direction at the 230 m point, SN1 lost connection to the AP of SN2, so the upstream direction could be measured to the 225 m point and the download direction to the 230 m point. Thus, the distance between nodes can be extended, so this experiment constructed a multi-robot environment with a multistage relay network, as explained in the next section.

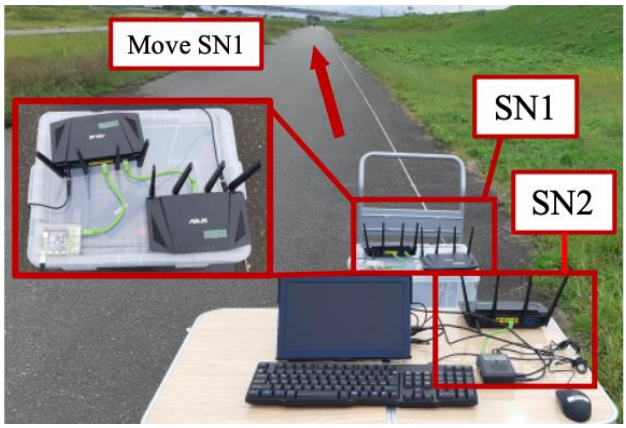


Fig. 8. Experimental Environment; Increase in Distance between Nodes.

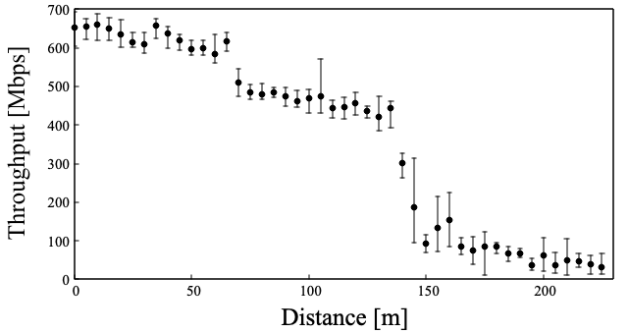


Fig. 9. Throughput of Upload; Increase in Distance between Nodes.

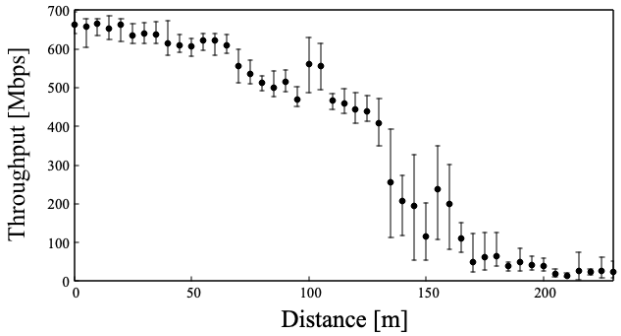


Fig. 10. Throughput of Download; Increase in Distance between Nodes.

B. Construction of Multi-robot Environment using Proposed Multistage Relay Network.

The purpose of this experiment is to confirm whether it is possible to construct a multi-robot environment that satisfies the requirements of the proposed network topology. This experiment measured the bandwidth compression throughput to confirm whether the constructed network can maintain the throughput required for multi-robot teleoperation. First, the distance between each node between the operator PC and mobile robot <1> in the proposed topology was increased from 0 m to 90 m at 10 m intervals. Then the bandwidth compression throughput was measured five times in both directions between the operator PC and the mobile robot <1> at

each point. A mobile robot was connected to the constructed network (the distance between each node was 90 m) and teleoperated, and bandwidth compression throughput measurement was performed five times in both directions between the operator PC and the mobile robot every 10 m. This experiment used a Raspberry Pi 4 Model B as the PC, ASUS RT-AX3000 and GT-AX11000 as the router, and a CAT8 LAN cable.

Fig. 11 shows the experimental environment of this section. This experiment was performed indoors with a straight 90 m line, so the operator PC and SN2 were deployed at the 0 m point, and SN1 and mobile robot <1> were moved to increase the distance between each terminal in the fold. A mobile robot connected to the constructed network switched the nodes to which it was connected is the following manner:

- The mobile robot was teleoperated by connecting to the operator's AP from the 0 m point to the 90 m point (point where SN1 was deployed).
- The mobile robot was teleoperated by connecting to the AP of SN1 from the 90 m point to the 180 m point (point where SN2 was deployed).
- The mobile robot was teleoperated by connecting to the AP of SN2 from the 180 m point to the 270 m point (point where mobile robot <1> was deployed).

Fig. 12 and 13 show the throughput values between the operator PC and mobile robot <1> during the expansion of the network construction range. Fig. 14 and 15 show the throughput values between the operator PC and the mobile robot during the teleoperation of the mobile robot. Fig. 12–14 shows the minimum, average, and maximum throughput values at each point. As stated in Fig. 11 and 12, the range of the multistage relay network was expanded to a distance of 90 m between each node (270 m in total), and the throughput was maintained at more than 60 Mbps in both directions. Fig. 13 and 14 also show that the throughput between the operator PC and the mobile robot remained over 60 Mbps.

These results indicate that a multistage relay network constituted by the proposed topology can maintain the required throughput for multi-robot teleoperation. Therefore, the robots connected to the multistage relay network constructed in this experiment could operate stably.

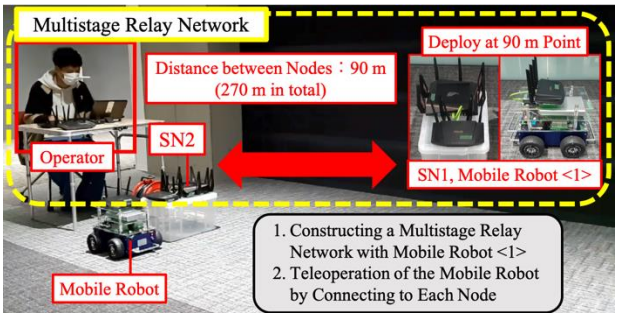


Fig. 11. Experimental Environment, Construction of Multistage Relay Network and Teleoperation of Mobile Robot.

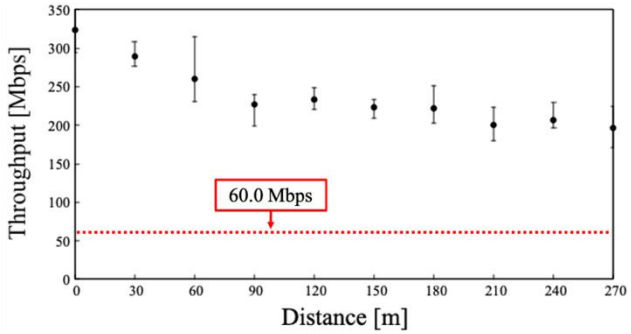


Fig. 12. Throughput of Upload between Operator PC and Mobile Robot <1> in Triple-hop Multistage Relay Network.

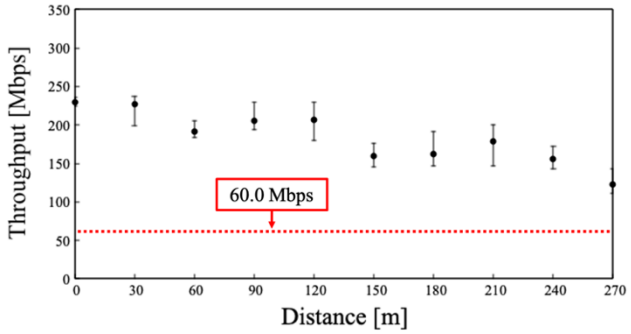


Fig. 13. Throughput of Download between Operator PC and Mobile Robot <1> in Triple-hop Multistage Relay Network.

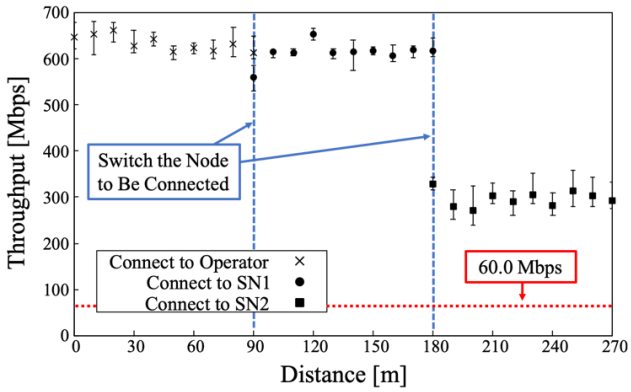


Fig. 14. Throughput of Upload between Operator PC and Mobile Robot during Mobile Robot Teleoperation.

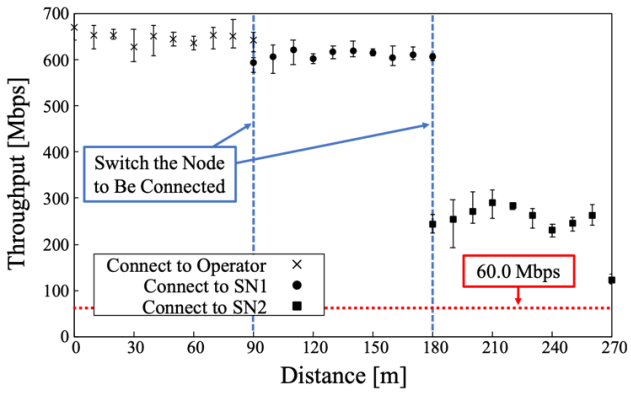


Fig. 15. Throughput of Download between Operator PC and Mobile Robot during Mobile Robot Teleoperation.

VII. Discussion

As the distance between nodes increased, the throughput between the operator PC and mobile robot <1> was measured up to the 225 m point. Fig. 9 and 10 show the throughput decreased by more than 100 Mbps when the distance between nodes was about 140 m. This decrease in throughput was deemed to have been caused by fallback, which decreases throughput in relation to a decrease in electric field strength. In the transmission and receiving of data between nodes, a delay or disconnection occurs if the amount of communicable data is exceeded. Furthermore, in wireless LANs, communication connectivity between nodes is maintained by fallback, which sets an upper limit on the transmission speed according to the electric field strength [20]. Therefore, the experiment of Section 6 assumed that the decrease in throughput at 140 m was due to the decrease in electric field strength caused by the increase in distance between nodes. This result suggests that it is necessary to monitor not only the throughput but also the electric field strength in a multi-robot system.

The multistage relay network was expanded to 270 m, maintaining a throughput of more than 60 Mbps. Therefore, more than three mobile robots could be teleoperated at 270 m in the network constructed in this experiment. A comparison between Fig. 9 and 10, and Fig. 12 and 13 indicates that the multistage relay network provided a larger search area. Moreover, in the teleoperation of the mobile robot, a throughput of more than 60 Mbps was maintained between the operator PC and the mobile robot after the construction of the multistage relay network. Thus, more than three mobile robots could be teleoperated when connected to any node of the network constructed in this experiment.

An RWSN cannot maintain the required throughput for multi-robot teleoperation, but the proposed method can ensure stable communication connectivity for multi-robot operation. The proposed topology is also effective in intricate environments because it can be used to construct a multistage relay backbone network with a high throughput.

VIII. Conclusion

This paper proposes a multistage relay network topology for constructing an environment with mobile robots for information gathering in large-scale facilities. Teleoperation of a mobile robot using a multistage relay network is effective for exploring in intricate environments and obstacle spaces due to its flexibility and scalability. However, in a multistage relay network, the communication quality decreases due to the increase in the distance between nodes and the number of relays, so a single robot is typically operated to maintain communication connectivity. As the search range of a single robot is limited, this paper considers the importance of operating a multi-robot system in a large-scale facility. A multistage relay network consisting of the proposed topology can connect multiple mobile robots to each node. These robots change the search range by switching the nodes to which they are connected. The experiment of Section 7 connected a mobile robot to a multistage relay network consisting of the proposed topology and confirmed its effectiveness by measuring the throughput. Findings indicated that this topology is effective for constructing a multi-robot environment in a large, enclosed space.

The proposed topology is effective for the teleoperation of multi-robot systems; however, there is a communication time loss when a mobile robot switches between connected nodes. Additionally, the mobile robot may become isolated if it fails to maintain the communication quality required for teleoperation in the communication path after changing nodes. Therefore, the future aim to develop a method of changing the communication route while maintaining the communication connectivity of a mobile robot.

References

[1] Yoshiaki Kawata, “The great Hanshin-Awaji earthquake disaster, damage, social response, and recovery,” Journal of Natural Disaster Science, Vol. 17, No. 2, pp.1-12, 1995.

[2] L. Ernesto Dominguez-rios, Tomoko Izumi, Yoshio Nakatani, “A disaster management platform based on social network system oriented to the communities self-relief,” IAENG International Journal of Computer Science, Vol. 42, No.1, pp.8-16, February 2015.

[3] Sabarish Chakkath, “Mobile robot in coal mine disaster surveillance,” IOSR Jourbal of Engineering, Vol. 2, No. 10, pp. 77-82, 2012.

[4] Keiji Sakuradani, Keigo Koizumi, Kazuhiro Oda, Satoshi Tayama,” Development of a sloap disaster monitoring system for expressway operation and maintenance control,” Jounal of GeoEngineering, Vol. 13, No.4, pp.189-195, December 2018.

[5] F. Kurz, D. Rosenbaum, J. Leitloff, O. Meynberg, P. Reinartz, “A real time camera system for disaster nad traffic monitoring,”. https://core.ac.uk/download/pdf/11146229.pdf

[6] Jingxuan Sun, Boyang Li, Yifan Jiang, Chih-yung Wen, “A camera-based target detection and positioning UAV system for search and rescue (SAR) Purposes,” Sensors 2016, Vol. 16, No. 11, 1778. https://doi.org/10.3390/s16111778.

[7] Masataka Fuchida, Shota Chikushi, Alessandro Moro, Atsushi Yamashita, Hajime Asama, “Arbitrary viewpoint visualization for teleoperation of disaster response robots,” Journal of Advanced Simulation in Science and Engineering, Vol. 6, No. 1, pp.249-259, 2019.

[8] Hemanth Reddy A, Balla Kalyan, Ch. S. N. Murthy, “Mine Rescue Robot System – A Review,” Procedia Earth and Planetary Science, Vol.11, pp. 457-462, 2015.

[9] Trupti B. Bhondve, Prof.R.Satyanarayan, Prof. Moresh Mukhedkar, “Mobile rescue robot for human body detection in rescue operation of disaster,” International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, No.6, pp.9876-9882, June 2014.

[10] Zia Uddin, Mojaharul Islam, “Search and rescue system for alive human detection by semi-autonomous mobile rescue robot,” International Conference on Innovations in Science, Engineering and Technology, Octorber 2016.

[11] Xuewen Rong, Rui Song, Xianming Song, Yibin Li, “Mechanism and explosion-proof design for a coal mine detection robot,” Procedia Engineering, Vol. 15, pp.100-104, 2011.

[12] Tomoaki Yoshida, Keiji Nagatani, Satoshi Tadokoro, Takeshi Nishimura, Eiji Koyanagi, “ Improvements to the rescue robot Quince toward future indoor surveillance missions in the Fukushima Daiichi Nuclear Power Plant,” Field and Service Robotics, pp. 19-32, December 2013.

[13] Albert Ko, Henry Y. K. La, “Robot assisted emergency search and rescue system with a wireless sensor network,” International Journal of Advanced Science and Technology, Vol. 3, pp.69-78, Febuary 2009.

[14] Andrew Wichmann, Burcu Demirelli Okkalioglu, Turgay Korkmaz, “The integration of mobile (tele) robotics and wireless sensor networks: A survey,” Computer Communications, Vol. 51, No.15, pp. 21-35, September 2014.

[15] Yasushi Hada, Osamu Takizawa, “Development of communication technology for search and rescue robots,” Journal of the National

Institute of Information and Communications Technology, Vol. 58, pp. 131-151, 2011.

[16] Carlos Marques, Joao Cristovao and Paulo Alvito, "A search and rescue robot with tele-operated tether docking system," Industrial Robot: An International Journal, Vol. 34, No. 4, pp. 332-338, 2007.

[17] Yuta Koike, Kei Sawai, Tsuyoshi Suzuki, "A study of routing path decision method using mobile robot based on distance between sensor nodes," International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 3, pp. 25-31, 2014.

[18] Kei Sawai, Ju Peng, Tsuyoshi Suzuki, "Throughput Measurement Method Using Command Packets for Mobile Robot Teleoperation Via a Wireless Sensor Network," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, pp 348- 354, 2016.

[19] Tsuyoshi Suzuki, Ryuji Sugizaki, Kuniaki Kawabata, Yasushi Hada, Yoshito Tobe, "Autonomous deployment and restoration of sensor network using mobile Robots," International Journal of Advanced Robotic Systems, Vol. 7, No. 2, pp. 105-114, 2010.

[20] Jeba Sonia J, Julia Punitha Malar Dhas, "A rate adaptation algorithm for IEEE802.11 wireless networks for commercial applications," Journal of Chemical and Pharmaceutical Sciences, Vol. 9, No.4, pp. 1904-1908, 2016.

Use of Value Chain Mapping to Determine R&D Domain Knowledge Retention Framework Extended Criteria

Mohamad Safuan Bin Sulaiman, Ariza Nordin, Nor Laila Md Noor, Wan Adilah Wan Adnan
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM)
Shah Alam, Malaysia

***Abstract*—Implementing a knowledge retention (KR) strategy is crucial to overcome the loss of expert knowledge due to employee turnover and retirement. The knowledge loss phenomenon caused organizations to face enormous risks which affect performance. KR frameworks and models are made available beyond research and development (R&D) organizations, to address knowledge retention strategies for administrative, operational, and manufacturing organizations. For research-intensive portfolios within R&D organizations, using the available KR frameworks requires fitting. The difficulty to address knowledge loss due to the uniqueness of the R&D organization's knowledge artifacts requires an extended KR framework. Before designing the extended KR framework, it is crucial to determine the framework's additional criteria. The paper reports the use of value chain mapping to determine the extended criteria of the KR framework fit for R&D organizations. The value chain mapping method identifies the knowledge activities in the R&D using Porter Value Chain (PVC) as the reference model. The output is a Knowledge Chain Model (KCM) that defines the critical points of knowledge loss in the R&D value chain. These critical points are project-based expert critical knowledge focus, project-based tacit knowledge transfer, and project-based knowledge repository which are nominated extended criteria of the KR Framework fit for R&D organizations.**

Keywords*—Knowledge retention framework; research and development; porter value chain; knowledge management; knowledge loss; research intensive portfolio; knowledge chain model

I. Introduction

Knowledge loss is a phenomenon that undeniably brings risk to organizational sustainability. The knowledge loss is spurred by either aging experts leaving the organization for retirement or expert or knowledgeable workers leaving for better job offers in other organizations. Knowledge loss is critical drawing much attention from the organizational knowledge management teams and risk managers to overcome the phenomenon using solutions pertaining to retention of critical knowledge loss. Despite having excellent KM practices in place, organizations still have a slimmer chance of facing critical knowledge loss [1].

Most of the studies of retention of critical knowledge loss were conducted on small and medium enterprises (SMEs), oil and gas, education, and manufacturing sectors [2,3,4]. These studies were more focused on the operational knowledge of the organizations. To address the critical knowledge loss in the organization, this research specificity focused on the critical knowledge loss in the research and development (R&D) organizations. Like other organizations, R&D organizations also face knowledge loss when their expert or knowledgeable researchers leave the organization. The loss indefinitely affects the performance of the R&D organization since the absence of expert researchers' knowledge hinders the completion of remaining or future R&D projects. The need to understand knowledge loss in an R&D organization is important since R&D is a strategic investment to produce better and newly featured products of technology either in a business or government. Overlooking knowledge loss in R&D organizations may be a strategic loss for the economic growth of a country. The current literature reported limited studies conducted for R&D organizations to address the critical knowledge loss of R&D activities.

The paper reports on a study to determine extended criteria required to fit the KR framework for an R&D organization. The use of value chaining mapping is elaborated to define criteria that are synthesized from the R&D and knowledge chain analysis. This study used the Porter value chain (PVC) model as the basis to develop the R&D value chain and knowledge chain model (KCM) to identify the knowledge activities in the R&D value chain.

II. Literature Review

Knowledge Retention (KR) is one of the important KM and organizational strategies to minimize critical knowledge loss [5]. Many KR Frameworks and models were proposed to overcome the knowledge loss but there is a lack of focus being given to KR for R&D organizations. The literature review of this study consists of four parts. The first part discusses frameworks and models for retention of critical knowledge loss to help understand the issues and context of the existing setup of managing critical knowledge loss.

The second part focuses on the R&D organizations to show the uniqueness of the organizations as compared to the operational-based organizations. The third part reviews the Porter Value Chain (PVC) model that is used in this study as the basis for synthesizing the R&D value chain. The fourth part reviews the knowledge chain model that is used to map the

knowledge activities with the R&D value chain. These reviews are towards the objective of identifying criteria that will be used as the foundation for the development of the KR framework for R&D organizations.

A. Frameworks and Models for Retention of Critical Knowledge in Organization

Several frameworks related to KR were reviewed for this purpose. This study has reviewed the frameworks and models proposed by Arif, Egbu, Alom & Khaflan [6], Boyles, Kirschnick, Kosilov, Yanev & Mazour [7], Doan & Rosenthal-sabroux [8], Levy [1] and Wamundila and Ngulube [9]. The summary of the frameworks is shown in Table I.

Early work on the KR model [6] for the construction organizations is based on a case study performed at construction companies in the United Arab Emirates (UAE). This model proposed a method to assess the KR capabilities of an organization and suggests opportunities for improvement. The model emphasizes a four-stage KR process that covers socialization, codification, knowledge construction, and knowledge retrieval.

Similarly in the same year, Boyles, Kirschnick, Kosilov, Yanev & Mazour [7] proposed a comprehensive cycle of retention processes in the case of retirement implemented in nuclear industries that emphasized several important stages that included (1) Conduct of risk assessment, (2) Determination and Implementation of the plan, (3) Monitoring and evaluation. Each process has sub-processes that further elaborate the retention of critical knowledge in detail. In addition, Boyles, Kirschnick, Kosilov, Yanev & Mazour [7] suggested a dedicated and separate self-assessment process in the case of employees who are leaving and transferring to other organizations or departments.

Continuing work on KR is seen in the work of Levy [1] who proposed a retention framework that is based on case studies performed in Israel between 2007 and 2010. The case studies were performed in seven organizations in banking, ministerial- level of a government department, national services, and defense industries. The model consists of three main stages for retention, which focused on the implementation stage for vertical knowledge transfer and eliminates assessment stage. The stages include (1) Scope, (2) Transfer, and (3) Integration. In the context of the organization, Levy [1] has underlined three types of organizational response to the phenomena of knowledge loss which include (1) Avoidance, (2) Engagement, and (3) Reaction. In this context, Knowledge retention is in the need for Engagement and Reaction-type of organizations because of inappropriate KM practice at the organizational level.

Wamundila and Ngulube ([9] proposed a retention framework for higher education institutions which the case study was performed at the University of Zambia (UNZA). The proposed framework has focused on (1) Identifying KR challenges at the organizational level, (2) Acknowledge the need and purpose for KR at the organizational level, (3) Preparedness of tacit and explicit knowledge integration, (4) Understanding the dimensions of KR which primarily encompasses knowledge assessment, acquisition, and transfer.

TABLE I. SUMMARY OF COMPONENTS IN EXISTING KR FRAMEWORKS

Authors	Components of the Framework
Arif et al. (2009) [6]	1. Personalization/Socialization (Individual knowledge) 2. Codification/ Externalisation (Conversion - Tacit to Explicit) 3. Combination (Organizational Memory: knowledge saved in IT/ Support Systems) 4. Internalisation (Retrieval - Explicit to Tacit: Retrieving Knowledge for Reuse)
Boyles et al. (2009) [7]	1. Conduct Risk Assessment 2. Determine and Implement Plan 3. Monitor and Evaluate
Doan et al. (2011) [8]	1. Top Management Support 2. ICT Tools 3. Knowledge Retention Process 4. Critical knowledge (Initiation, 5. Implementation and evaluation) 6. Business Process Focus 7. Human Resource Practices 8. Knowledge Retention Strategy 9. Learning Culture
Levy (2011) [1]	1. Initiating the process 2. Scope 3. Transfer 4. Integration 5. Structured Process 6. Structured Result
Wamundila and Ngulube (2011) [9]	1. Identify knowledge retention challenges 2. Acknowledge need and purpose for knowledge retention 3. Integrate tacit and explicit knowledge 4. Dimensions of knowledge retention (knowledge assessment, acquisition, and transfer)

Doan & Rosenthal-Sabroux [8] proposed a reference model of knowledge retention for Small and Medium-Sized Enterprises (SMEs). The model consists of several elements that are believed to be critical for an effective KR implementation. Doan & Rosenthal-Sabroux [8] suggested the model can be used as a starting step of the KR initiative and as a template to assess the KR maturity level.

Based on the summary in Table I, all frameworks and models have variations of components which some have similar, and some have their unique components. In addition, data that have been acquired to propose the frameworks are mainly from operational-based organizations. Whilst an analysis by Sulaiman [10] has underlined several issues on the applicability and completeness of KR frameworks for knowledge-intensive organizations which found such limitations in the existing frameworks that lack of technology used to help the assessment process during the implementation of KR and lack of study had been done in R&D organization and remain as recommendations for future exploration. These limitations have motivated this study to be conducted.

B. Research and Development (R&D) Organizations

Research and development (R&D) organization is an example of a knowledge-intensive organization where KR is deemed to be important. Before further discussion it is good to clarify the definitions of R&D. Research is defined in a few categories which includes.

1) Basic research that its objective to gain more complete knowledge of the studied subject without a specific application in mind with the advancement of scientific knowledge without working for long-term economic or social benefits and with no positive efforts to apply it (pure basic research) and produce a broad base of knowledge to form the background to the solution of problems (oriented basic research) without a specific commercial goal (oriented basic research) [11,12].

2) Applied Research is the acquisition of knowledge to determine the means to achieve a specific and recognized need by discovering new scientific knowledge that has specific commercial objectives concerning products, processes, or services [11,12].

Development is defined as the systematic use of the knowledge or understanding gained from research, directed toward the production of useful materials, devices, systems, or methods, including the design and development of prototypes and processes. [11]. It is also called experimental development which means a systematic work, drawing on existing knowledge gained from research and practical experience that is directed to producing new materials, products, and devices; to installing new processes, systems, and services; or to improving substantially those already produced or installed [12].

R&D is found to be the most important component in any part of modern businesses that creates new, robust, and better products, processes, and the way people do things. Elements of R&D in the organizations and firms have a strong influence on the success rate in their business and help them gain a competitive advantage over other firms [13].

R&D also plays important role in the economic sector in a nation, making a profit in business enterprise, effective in technology-based governmental agencies (e.g., the US Department of Defence), and the higher investment in R&D activities by a nation ($355 billion in the United States in 2007), [14]. Korea is one of the many examples, where R&D spending is on more advanced industries that foster productivity growth and proven that the productivity impact of R&D is stronger in more high-tech industries and during economic downturns [15].

R&D organizations are different from other organizations based on four elements 1) People, 2) Ideas, 3) Funds and 4) Culture [14]. Managing an R&D organization is not simple and good management of research is not only the critical difference among the organizations, but the research itself is the most difficult to manage as compared to other functional activities [14].

The importance of R&D and the uniqueness of R&D organizations due to their people, ideas, funds, and culture raise the need to further explore KR in R&D organizations. This agrees with the findings from Sulaiman [10] who also suggested the need to further explore KR in the R&D organization. As a first step, it is necessary to identify where critical knowledge loss could occur in R&D organizations before addressing the KR of R&D organizations. For this purpose, the Porter value chain (PVC) and knowledge chain model (KCM) analyses can be used to identify the critical knowledge loss in R&D organization value chain activities. PVC and KCM are further elaborated in the following sub-sections.

C. Porter Value Chain (PVC)

Porter [16] introduced the value chain concept to describe a set of activities that an organization carries out to add value to its customer. The concept is now formally known as the Porter value chain (PVC) and is an established mechanism to understand the value chain in operational, business, and manufacturing organizations. Various authors have used the PVC in their research settings ever since. Relating to the knowledge value chain, the PVC has been used by Holsapple and Singh [17] for mapping a proposed knowledge chain model and Jordan et al. [18] used PVC as the basis for the product value chain in his framework for evaluating R&D impact and supply chain.

The PVC model consists of nine value-adding activities with five primary and four secondary activities toward competitive advantage as illustrated in Fig. 1.

The primary value-added activities consist of inbound logistics, operations, outbound logistics, marketing and sales, and service and are defined as in Table II.

The secondary activities (Table II) involve corporate infrastructure, human resource management, technology development, and procurement. In the primary activities, that PVC shows the important value-added at each stage starting from input materials until the finished product that can be marketed. While secondary activities are in support of the whole range of primary activities.

The mapping of PVC to specific organization activities highlights the value-added at each phase of the organization's primary activities and identifies the organization's competitive advantage. The mapping of PVC activities to organizational value chain was used by Tomasevic and Stojanovic [19] and Sobotka [20] for educational institutions and Rapcevi [21] for public sectors.

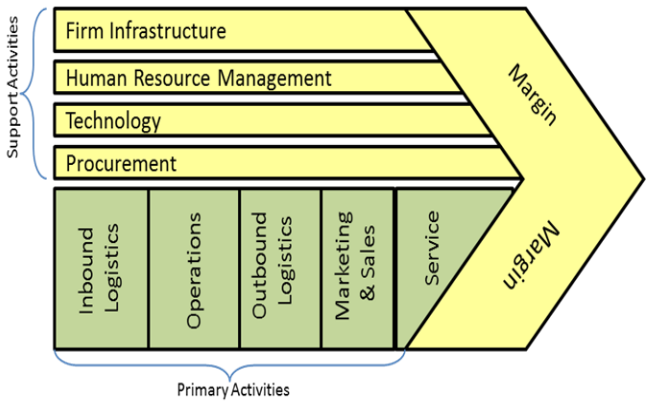


Fig. 1. Porter Value Chain (PVC) Model.

TABLE II. DEFINITION OF THE PVC ACTIVITIES

Activity	Definition
Primary	
1. Inbound Logistics	Receiving, storing, and distributing materials to manufacturing premises.
2. Operations	Transforming inputs into finished products
3. Outbound Logistics	Storing and distributing products
4. Marketing and Sales	Promotion and sales efforts
5. Service	Maintain or enhance product value through post-sale services
Secondary	
1. Corporate Infrastructure	Support for the entire value chain including general management, planning, finance, accounting, legal services, government affairs, and quality management
2. Human Resource Management	Recruiting, hiring, training, and development of employees
3. Technology Development	Improving product and manufacturing process
4. Procurement	Purchasing input

D. Knowledge Chain Model (KCM)

The KCM was proposed by Holsapple and Singh [17] and was based on a descriptive KM framework developed via a Delphi study involving an international panel of prominent KM practitioners and academicians [22]. The model has five primary and four secondary KM activities in the KCM as in Fig. 2 [23].

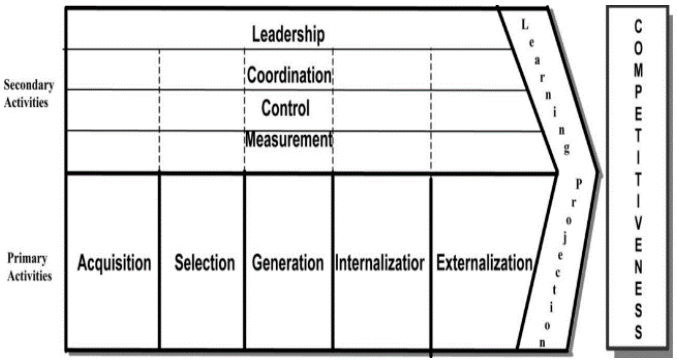


Fig. 2. Knowledge Chain Model (KCM).

Five primary KM activities include Knowledge acquisition, selection, generation, internalization, and externalization, and four secondary KM activities include knowledge leadership, coordination, control, and measurement. The definition of the primary and secondary KM activities is in Table III.

The reviewed KR frameworks from previous works are different from one another. The differences have shown the gaps between the frameworks. A small but important gap has shown that a lack of data was acquired from R&D organizations as the basis of producing the existing KR Framework. Therefore, the importance and unique characteristics of R&D organizations are reviewed and have shown some differences between operational-based and R&D organizations. The reviewed PVC model has shown the useful technique to derive the R&D organization value chain activities as been used in the operational-based organization as well as another mapping for the organizational value chain. While KVC is reviewed, to address important knowledge activities at the organizational level that might lead to critical knowledge loss.

TABLE III. DEFINITION OF THE PRIMARY AND SECONDARY KM ACTIVITIES

Primary Activities	Secondary Activities
Knowledge Acquisition: Acquiring knowledge from external sources and making it suitable for subsequent use. Knowledge Selection: Selecting needed knowledge from internal sources and making it suitable for subsequent use. Knowledge Generation: Producing knowledge by either discovery or derivation from existing knowledge. Knowledge Internalization: Altering the state of an organization's knowledge resources by distributing and storing acquired, selected, or generated knowledge. Knowledge Externalization: Embedding knowledge into organizational output for release into the environment.	Knowledge Leadership: Establishing conditions that enable and facilitate fruitful conduct of KM Knowledge Coordination: Managing dependencies among KM activities to ensure that proper processes and resources are brought to bear adequately at appropriate times Knowledge Control: Ensuring that needed knowledge processors and resources are available in sufficient quality, subject to security requirement Knowledge Measurement: Assessing values of knowledge resources, knowledge processors, and their deployment

III. METHODOLOGY

The research method employed is a two-step procedure that focused on (1) the PVC mapping on R&D processes to produce R&D organization value chain activities and (2) the knowledge chain model mapping on R&D organization knowledge chain activities to identify possible loss of critical R&D knowledge.

The PVC mapping on R&D processes was conducted based on the R&D processes as described by several kinds of literature that include the work of [23-29]. An interpretive analysis based on PVC is then used to map the operational-based organization with the R&D organization value chain. This analysis is purposely to identify the differences between operational-based and R&D organizations and produce the R&D organization value chain.

In the next step, the knowledge chain model by Holsapple and Singh [17] is used to map each R&D organization's value chain activities with knowledge activities. The result of the mapping is used to identify knowledge activities at each R&D value chain activity that might lead to the possible loss of critical R&D knowledge.

By knowing the critical point of R&D knowledge loss from the findings of those mappings, several criteria for retention of critical knowledge loss would be suggested for KR Framework in R&D organizations.

IV. Data and Analysis

The mapping of PVC onto the R&D organizational process is mainly used in R&D standard processes [26]. The R&D standard processes are well-structured and similar to the PVC structure [26].

Standard R&D processes into two levels: the organizational level and the project level [26]. The organizational level contains organizational processes such as R&D Planning, Portfolio Management, Idea Management, Intellectual Property Management, Infrastructure Management, Human Resource Management, Organizational Performance Management, and Quality Management The structure of the standard R&D processes at the organizational level is shown in Fig. 3.

The project level contains support processes and fulfillment processes. Support processes consist of Project Planning, Project Monitoring, Gate Assessment, Collaboration Mgt. and Risk Mgt. while, fulfillment processes consist of Concept Modelling, Business Feasibility, Specification Definition, Design, Development, Prototype, Market Test, and Market Launch. The standard R&D processes at the project level are shown in Fig. 4.

Besides standard process mapping [26], some other R&D processes are also considered in the mapping includes [27,28,28] to support the works from Yoon, Lee, Lee & Yoon [26] and some have additional perspectives [28] Table IV shows the summary of R&D processes, used in this mapping.

Fig. 3. Standard R&D Processes at the Organizational Level.

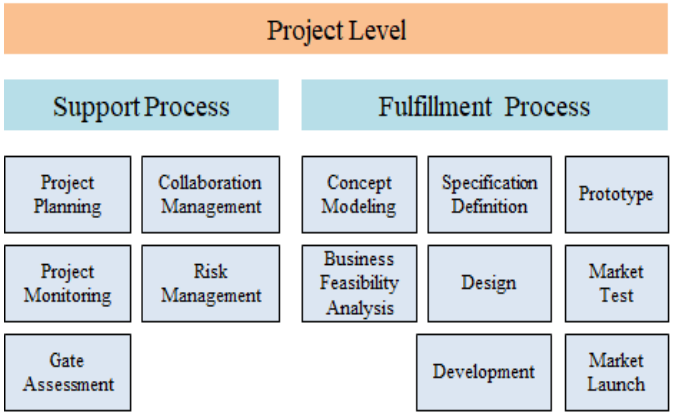


Fig. 4. Standard R&D Processes at the Project Level.

TABLE IV. Summary of R&D Components and Processes

Authors	R&D Components/ Processes	/Processes
Yoon, Lee, Lee & Yoon. (2015) [26]	• Organizational Process • Support Process • Fulfillment Process	• R&D Planning • Portfolio Mgt. • Idea Mgt. • Intellectual Property Mgt. • Infrastructure Mgt. • Human Resource Mgt. • Organizational Performance Mgt. • Quality Mgt. • Project Planning • Project Monitoring • Gate Assessment • Collaboration Mgt. • Risk Mgt. • Concept Modeling • Business Feasibility • Spesification Definition • Design • Development • Prototype • Market Test • Market Launch
Martin (2014) [28]	• Foster Ideas • Focus Ideas • Develop Ideas • Prototype And Trials • Regulatory, Marketing, And Product Development Activities • Launch	
Kalypso (2018) [27]	• R&D\Strategy	• Portfolio Management • Innovation • Project Management • Ip Management • Sourcing • Talent Management • Regulatory Compliance • R&D Operation
Rousselon Saadn & Erickson (1991) [29]	• Priority Setting • Portfolio • Management • Project Management • Strategic Planning	• Program Analysis • Portfolio Analysis • Portfolio Adjustment • Portfolio Assignment • Project Planning • Project Budgeting • Manpower/Resource Planning • Scheduling • Active Monitoring • Project Analysis • Technology Forecasting • Strategic Planning • Market Forecasting

Based on some similar structures that exist in secondary activities on PVC and R&D standard processes [26], this study focused the analysis on the primary activities (PVC) and R&D standard processes at the project level. Each R&D process in Table III is grouped based on the PVC activities. However, the mapping of the R&D processes into PVC activities is rather difficult because the definition of inbound logistics, operations, and outbound logistics is more focused on the product. Therefore, some adjustments on the terminology of the value chain activities and definitions are suggested because R&D standard processes are more focused on the research project rather than the product itself. For Inbound Logistics, Operations, Outbound Logistics, and Marketing and Sales, new terminology of the value chain activities for R&D are R&D Inputs, R&D Work Processes, R&D Outputs, and Realization respectively. The result of the mapping is shown in Table IV and the adjustment on the definition is suggested as follows:

- R&D Inputs: Inputs to the research project that is needed to conduct Research Activities such as project team member, planning, and funding.
- R&D Work Processes: Activities needed to achieve the objective of the research project such formulation of the research project, development of ideas, data collection, development of concept, model, theories by utilizing the research input.
- R&D Outputs: The output of the research project such as patents, Innovations, products, and publications acquired from research activities.
- Realization: Established and tested R&D output such as products and processes are packaged to be marketed and applied to targeted industries and commercialize the application of the Research Output to many industries.
- Service: Activities of maintaining the R&D output such as marketed product and processes or usually called after-sales service.

Table V indicates the mapping of the R&D processes onto the PVC components and is based on the above definitions. Mapping Inbound Logistics onto R&D Inputs include business feasibility, specification definition, portfolio analysis, portfolio adjustment, portfolio assignment, and sourcing.

R&D Work Processes include project monitoring and analysis, concept modeling, design, development, scheduling, active monitoring, R&D operation, fostering and developing ideas, prototype and trials, product development activities, and gate assessment.

R&D Outputs include Innovation and Intellectual Property and mapping Marketing & Sales onto Realization include market launch, test, and forecasting, launch, and marketing.

Based on the mapping analysis, it can be understood, Inbound Logistics can be mapped with Research Inputs, Operations with R&D Work Processes, and Outbound Logistics with Research Outputs and Marketing and Sales, with Research Realization [30].

TABLE V. MAPPING R&D PROCESSES ONTO PORTER VALUE CHAIN (PVC)

Primary Activities (*Project Level*)				
Inbound logistics *(R&D Input)*	**Operations (*R&D Work Processes)***	**Outbound logistics *(R&D Output)***	**Marketing & Sales *(Realization)***	**Service**
Business Feasibility	Project Monitoring	Innovation	Market Launch	
Specification Definition	Concept Modelling	Intellectual Property	Market Test	
Portfolio Analysis	Design		Market forecasting	
Portfolio adjustment	Development		Launch	
Portfolio assignment	Prototype		Marketing	
Sourcing	Scheduling			
	Active Monitoring			
	Project Analysis			
	R&D Operation			
	Foster Ideas			
	Focus Ideas			
	Develop Ideas			
	Prototype and trials			
	Product dev. activities			
	Gate Assessment			

None of the R&D processes could be mapped onto the service activities because it is not part of R&D processes and usually under the technical department after the technology has been transferred within a period. The result of the mapping is shown in Fig. 6.

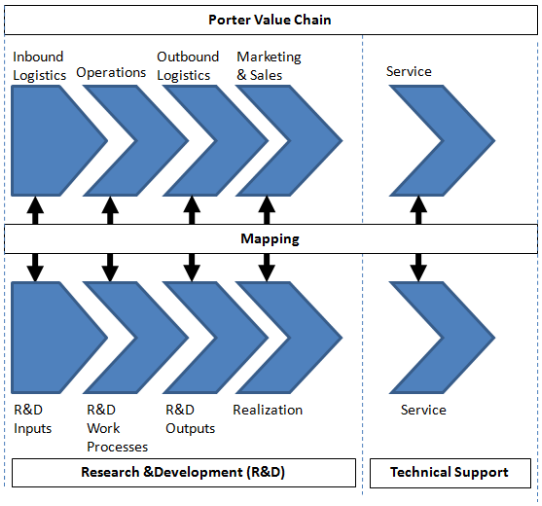


Fig. 5. Result of the PVC Mapping onto R&D Processes.

As referred to Fig. 5, four primary value chain activities for the R&D value chain were identified which include R&D inputs, R&D work processes, R&D Outputs, and Realization. Whilst Service is more towards technical support and was not considered as part of R&D processes. As a result, the proposed R&D value chain is as in Fig. 6.

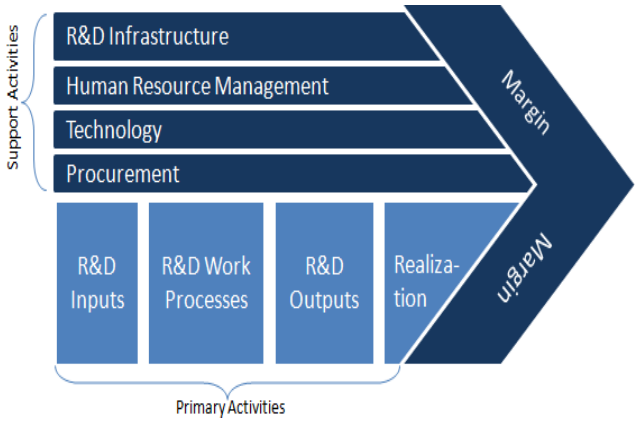


Fig. 6. The R&D Organization Value Chain.

The primary activities in the R&D organization value chain as in Fig. 6 are then mapped with the knowledge chain model to identify the point of critical R&D knowledge loss. The knowledge chain model (KCM) of Holsapple and Singh [17] is used in this study to identify knowledge chain activities in each of the R&D value chain activities. The definition of each primary knowledge chain activity is suggested by Holsapple and Singh [17] as in Table III.

From the definition of primary knowledge chain activities, it is understood, each R&D value chain has at least one knowledge chain activity. Knowledge acquisition and selection occur at R&D Inputs, knowledge generation and internalization occur at R&D Work Processes Activities, knowledge generation and externalization occur at R&D Outputs and knowledge externalization occurs at Realization. The result of the mapping of knowledge value chain activities on R&D value chain activities is shown in Fig. 7.

At this stage, each R&D value chain activity has knowledge activities and at some R&D activities, it produces and generates new knowledge.

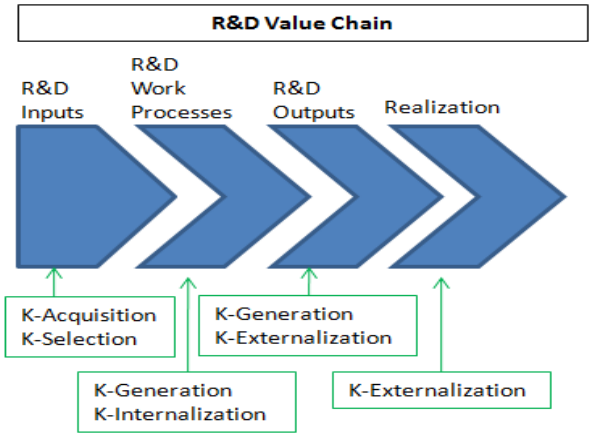


Fig. 7. R&D Value Chain and Knowledge Chain Activities.

V. Discussion

As this study is focusing on the criteria for knowledge retention framework, further discussion on the context of knowledge aspects for retention of critical knowledge in R&D organization which is based on the R&D value chain and knowledge activities is relevant.

From the analysis of mapping PVC onto R&D Processes, it is discovered that the R&D organization's value chain has different primary activities with similar support activities. These can be seen in the adjustment of the definitions at the primary activities as well as the absence of the R&D process at Service activity. The differences proved that R&D and operational-based organization is different obviously at the primary activities which perhaps the existing KR framework is not fit for R&D organizations.

From the KCM mapping onto primary R&D value chain activities, it is found that knowledge activities occur in each R&D value chain activity. It is also believed that each knowledge activity such as knowledge acquisition, selection, generation, internalization, and externalization has intensively and extensively occurred from the beginning until the end of the R&D projects.

R&D Project team which is considered as R&D knowledge workers play an important role in each R&D and its knowledge activities. In addition, an expert in the project team plays a more significant role in the R&D project. According to Joe, Yoong & Patel [31], "experts are a powerful source of value creation within organizations and are people who have deep specialized knowledge of a subject, who are tested and trained, especially by experience. Expert demonstrates higher levels of efficiency, performs tasks with greater accuracy and cost-effectiveness and holds subject-specific knowledge, such as on methods and procedures, including knowledge of how to deal with problems and new situations". Expert knowledge is also a valuable organizational resource [32]. To a certain extent, the experts often do not realize that they possess unique valuable knowledge, and for cognitive reasons, they are not able to express this knowledge [33]. In the situation, if an expert suddenly leaves the organization during the development of a prototype, where the knowledge generation is just started to be implemented and no one has the capability similar or nearly like the expert, it is critical and affects the project progress as well the performance of the R&D organization. This implies that an R&D expert is one of the critical points to be seriously considered for the criteria of KR in R&D organizations.

By looking into the R&D value chain and its knowledge activities, the tacit and explicit knowledge of each R&D project team will possibly accumulate over time and at each R&D value chain activity. The knowledge of the team gradually and proportionally increased over time and chain activities. It means that knowledge accumulated at R&D Outputs activities is more than knowledge accumulated at R&D Inputs. The knowledge accumulated over time and value chain activities is an asset to the organization. Losing experts or project team members at the final stage of the R&D value chain activities are more critical than at the earlier R&D stages because more efforts and resources must be invested to make the project progress as planned. Therefore, the accumulated

knowledge over time and R&D chain activities is found to be important to be seriously considered where some strategic action could be possibly done to capture the knowledge as it progressed.

From the above arguments, expert knowledge and accumulated R&D project knowledge are two important factors to be seriously considered for the criteria setting of KR Framework for R&D organization based on the following findings:

- R&D organizations are expertise-oriented and not solely product-oriented organizations. Therefore, in the context of retention of critical knowledge loss, more focus should be given to expert knowledge instead of knowledge of the product and innovation of the R&D projects at every R&D value chain activity.
- Each R&D project must have project team members that manage the accumulated R&D project knowledge at each R&D value chain activity. The accumulated knowledge in each R&D value chain is important starting from the R&D Inputs until Realization activities and it should be easily stored and transferred.
- Each project team consists of several persons that might have specific and diversified knowledge of the working R&D projects and come from different generations. Transfer and sharing activities of tacit knowledge between members should be_done as early as at the R&D Inputs to prevent knowledge loss.
- Each R&D value chain activity has its knowledge activities which might hold the critical point of losing R&D knowledge due to retirement, aging, and turnover factors. Therefore, each R&D project shall have documentation at every R&D value chain activity and stored in organizational memory to prevent the loss of explicit knowledge.

VI. Proposed Criteria for KR Framework

Based on the previous findings derived from the PVC and KCM analyses in Section V, this study underlines several criteria on the perspective of critical R&D knowledge loss to be used as the basis of developing the KR Framework for R&D organizations. Details of the criteria are as follows:

- Focused on the KR should be given to the retention of the R&D expert knowledge and potential project team that might leave the organization. Therefore, the KR Framework for R&D organization should have a thorough assessment component and shall be focusing on the critical R&D expert knowledge so that the potential leaving critical experts or members of the project team and a successor could be easily identified and risk of losing the critical knowledge of the project team could be minimized and retained as early as possible. It is also recommended to use some technology [34], to assist the preparation of the assessment process and prevent the waste of organizational resources [10].
- For expert knowledge transfer, the minimum knowledge gap between an expert and a successor and within the R&D project team would be a possible chance of easy transfer of critical knowledge. Therefore, the KR framework for R&D organizations should consider mapping and binding experts and successors for tacit knowledge transfer. In the case of the expert and successor have multiple R&D projects. The transfer shall be based on the project's point of view.
- For accumulated R&D project knowledge, each knowledge activity in R&D value chain activities should retain the R&D project as it's progressed to minimize critical knowledge loss due to retirement and employee turnover. Therefore, the KR framework should consider a mechanism for easy transfer, store and retrieve R&D project knowledge as the project progressed. It is also important to consider the use of technology so that knowledge stored procedure would be easily reinforced, and the organizational memory would be easy and well-structured, and organized.

VII. Conclusion and Future Work

Most studies on KR frameworks are based on operational-based organizations and a lack of focus had been given to R&D-based organizations. As a first step towards KR frameworks for R&D organizations, the first part of this study used mapping of PVC onto R&D processes to identify the differences between both organizations. From the mapping analysis, the differences were found at primary activities of both organizations and proposed an R&D organization value chain. The second part of this study used the mapping of KCM onto the proposed R&D organization value chain to identify knowledge chain activities in the R&D value chain. From the KCM mapping analysis, expert and accumulated R&D knowledge factors were identified as a critical point of potential R&D knowledge loss. Several criteria used as a basis for the development of the KR framework for R&D organizations were proposed which were based on several findings from the result of the mapping analyses.

Some empirical study is suggested for future works to further extend and examine the findings and streamline the recommended criteria of the KR framework for R&D organizations.

Acknowledgment

Special thanks to the Government of Malaysia, Malaysian Nuclear Agency, and Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) for support during the conduct of this study.

References

[1] Levy M. Knowledge retention: minimizing organizational business loss. Journal of Knowledge Management. 2011 Jul 19.

[2] Ramona T, Alexandra B. Knowledge retention within small and medium-sized enterprises. Studies in Business and Economics. 2019 Dec 1;14(3):231-8.

[3] Haughton R. Exploring knowledge retention strategies to prevent knowledge loss in project-based organizations (PBOs) (Doctoral dissertation, Walden University).

[4] Sumbal MS, Tsui E, Durst S, Shujahat M, Irfan I, Ali SM. A framework to retain the knowledge of departing knowledge workers in the manufacturing industry. VINE Journal of Information and Knowledge Management Systems. 2020 Jan 10.

[5] Sanz R, Hovell J. Knowledge retention framework and maturity model: improving an organization or team's capability to retain critical knowledge. Knowledge Management for Development Journal. 2021 Aug 28;16(1):8-27.

[6] Arif M, Egbu C, Alom O, Khalfan MM. Measuring knowledge retention: a case study of a construction consultancy in the UAE. Engineering, Construction, and Architectural Management. 2009 Jan 9.

[7] Boyles JE, Kirschnick F, Kosilov A, Yanev Y, Mazour T. Risk management of knowledge loss in nuclear industry organizations. International Journal of Nuclear Knowledge Management. 2009 Jan 1;3(2):125-36.

[8] Doan QM, Grundstein M, Rosenthal-Sabroux C. A reference model for knowledge retention within small and medium-sized enterprises.2011.

[9] Wamundila S, Ngulube P. Enhancing knowledge retention in higher education: A case of the University of Zambia. South African Journal of Information Management. 2011 Jan 1;13(1):1-9.

[10] Sulaiman MS, Abdul R, Nordin A, Noor NL. Improving Knowledge Preservation Strategy at Organizational Level through Knowledge Loss Risk Assessment (KLRA). 2016.

[11] National Science Board.Science and Engineering Indicators 2008, Two volumes (1, NSB 08-01; 2, NSB 08-01A). Arlington, VA: National Science Foundation.2008.

[12] Manual O. The measurement of scientific and technological activities. Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. 2005 Jul;30.

[13] Liu O, Wang J, Ma J, Sun Y. An intelligent decision support approach for reviewer assignment in R&D project selection. Computers in Industry. 2016 Feb 1;76:1-0.

[14] Jain R, Triandis HC, Weick CW. Managing research, development and innovation: Managing the unmanageable. John Wiley & Sons; 2010 Jun 18.

[15] Lee D. Role of R&D in the productivity growth of Korean industries: Technology gap and business cycle. Journal of Asian Economics. 2016 Aug 1;45:31-45.

[16] Porter, M. (1985) Competitive Advantage, New York: The Free Press.

[17] Holsapple CW, Singh M. The knowledge chain model: activities for competitiveness. Expert systems with applications. 2001 Jan 1;20(1):77-98.

[18] Jordan G, Mote J, Ruegg R, Choi T, Becker-Dippmann A. A Framework for Evaluating R&D Impacts and Supply Chain Dynamics Early in a Product Life Cycle. Looking inside the black box of innovation. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014 Jun 1.

[19] Tomašević I, Stojanović D, Simeunović B, Radović M, Andrić-Gušavac B. Creating Value in Higher Education Institutions. InToulon-Verona Conference" Excellence in Services" 2015 Aug 25.

[20] Sobotka B. Value chain in education sector illustrated with an example of Vocational Competence Certificate System. InForum Pedagogiczne 2016 (Vol. 2, pp. 305-316). Wydawnictwo Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie.

[21] Rapcevičienė D. Modeling a value chain in public sector. Social Transformations in Contemporary Society. Disponible en (último acceso noviembre de 2017): http://stics. mruni. eu/wpcontent/uploads/2014/08/STICS_2014_2_42-49. pdf. 2014.

[22] Davè KP. An investigation of knowledge management characteristics: Synthesis, Delphi study, analysis. University of Kentucky; 1998.

[23] Lee CC, Yang J. Knowledge value chain. Journal of management development. 2000 Nov 1.

[24] Saha A. Mapping of Porter's value chain activities into business functional units. Management Innovation Exchange. Retrieved on September. 2011;13:2017.

[25] Wang, C. L. and Ahmed, P. K. (2005). The knowledge value chain: a pragmatic knowledge implementation network. Handbook of Business Strategy 6 (1): 321-326.

[26] Yoon B, Lee K, Lee S, Yoon J. Development of an R&D process model for enhancing the quality of R&D: comparison with CMMI, ISO and EIRMA. Total Quality Management & Business Excellence. 2015 Aug 3;26(7-8):746-61.

[27] Kalypso, Manage R&D as a Business: R&D Management Framework accessed from http://viewpoints.io/uploads/files/RnDManagement.pdf, on 8 April 2021.

[28] Martin. Research and Development (R&D), Overview and Process, Cleverism, 2014 accessed from https://www.cleverism.com/rd-research-and-development-overview-process/ on 8 April 2021.

[29] Rousselon PA, Saadn KN, Erickson TJ. The evolution of third generation R&D. Planning Review. 1991 Feb 1.

[30] Yoshikawa H. Design methodology for research and development strategy. Japan: Center for Research and Development Strategy (CRDS). 2012 Feb.

[31] Joe C, Yoong P, Patel K. Knowledge loss when older experts leave knowledge-intensive organisations. Journal of Knowledge Management. 2013 Oct 18;17(6):913-27.

[32] Hammer M. The Getting and Keeping of Wisdom. Canada: Public Service Commission of Canada. 2002 Oct.

[33] Hinds PJ, Pfeffer J. Why organizations don't "know what they know": Cognitive and motivational factors affecting the transfer of expertise. Sharing expertise: Beyond knowledge management. 2003:3-26.

[34] Sulaiman MS, Nordin A, Noor NL. A review of knowledge retention frameworks for knowledge intensive organization. In2016 International Conference on Information Management and Technology (ICIMTech) 2016 Nov 16 (pp. 106-111). IEEE.

Adaptive Deep Learning based Cryptocurrency Price Fluctuation Classification

Ahmed Saied El-Berawi[1]
Computer Networks and Datacenter
Arab Academy for Science Technology and Maritime and Maritime Transport
Alexandria, Egypt

Mohamed Abdel Fattah Belal[2]
Faculty of Computers and Artificial Intelligence, Helwan University
Cairo, Egypt

Mahmoud Mahmoud Abd Ellatif[3]
College of Business, University of Jeddah, Saudi Arabia
Faculty of Computers and Artificial Intelligence, Helwan University
Cairo, Egypt

***Abstract*—This paper proposes a deep learning based predictive model for forecasting and classifying the price of cryptocurrency and the direction of its movement. These two tasks are challenging to address since cryptocurrencies prices fluctuate with extremely high volatile behavior. However, it has been proven that cryptocurrency trading market doesn't show a perfect market property, i.e., price is not totally a random walk phenomenon. Based upon this, this study proves that the price value forecast and price movement direction classification is both predictable. A recurrent neural networks based predictive model is built to regress and classify prices. With adaptive dynamic features selection and the use of external dependable factors with a potential degree of predictability, the proposed model achieves unprecedented performance in terms of movement classification. A naïve simulation of a trading scenario is developed and it shows a 69% profitability score a cross a six months trading period for bitcoin.**

Keywords*—Computer intelligence; cryptocurrency; deep learning; market movement; recurrent neural network; timeseries forecasting

I. Introduction

Cryptocurrency, or digital currency, is a virtual currency used for the exchange and transfer of assets. When compared to traditional currencies, which rely on central banking institutions, cryptocurrencies are built on the idea of decentralized control. As a result, a cryptocurrency is used to send money electronically without the involvement of a central or governmental authority. Within the last few years, and due to its uncontrollable and untraceable character, there has been a growing interest in cryptocurrency trading. The industry has grown tremendously for financial transactions and trading throughout the world. It also shows a continuing and growing trend for the near future. The market is expected to rise from 1.8 billion to 2.2 billion USD by 2026 according to https://coinmarketcap.com/charts/" (accessed Jul. 18, 2021). Moreover, when compared to traditional state-issued currencies, cryptocurrencies are extremely volatile. Their exchange rates cannot be assumed to be independently and identically distributed phenomena [1]. In a short period of time, the cryptocurrency sector has experienced exponential development and global popularity. Such increase in popularity received significant media attention, attracting more investors, researchers, regulators, and speculators to the field as promising booming business. Such increased popularity necessitates study into their dynamics and how they affect the financial sector and economies of countries in general.

The following sections discuss in short the intrinsic characteristics of cryptocurrency systems that relate to the problem statement addressed, and what are challenges, gaps and research questions of this work.

A. Cryptocurrency Characteristics

The value of each cryptocurrency is determined by its volume of transactions and price movements. Moreover, each cryptocurrency has its own ecosystem and operates differently from the other ones in terms of value variations, transaction speeds, usages, and volatility. Such independence makes forecasting cryptocurrencies prices a challenging task.

Another characteristic is that despite the massive bitcoin meltdown at the start of 2018, with a lot of volatility, people's interest in it has remained relatively steady. In response, in recent years, academics have presented a variety of approaches for predicting and modelling the price of cryptocurrencies as well as analyzing the volatility of the crypto market. Also, it is difficult to say exactly what drives the price of cryptocurrencies over time. However, P Katsiampa used the *asymmetric Diagonal BEKK* model to investigate the volatility dynamics of four main cryptocurrencies in this extensive analytical work that had been published in 2019 [2]. The conditional variances (the variance of a phenomenon given the value(s) of one or more factors in econometrics) of all four cryptocurrencies are strongly impacted by both prior squared errors and past conditional volatility, according to the study. Furthermore, big news has been proven to affect volatility dynamics. Finally, the crypto industry as a whole is known for its price fluctuations and trading volumes. With the growing interest in cryptocurrencies and their importance in the financial sector, extensive research and forecasting of cryptocurrencies' volatility dynamics is needed [3].

Also, because of the exponential growth in speculative activities, cryptocurrency markets are increasingly vulnerable to price fluctuations and degradation. It's still unclear if the biggest price fluctuations in cryptocurrencies are unpredictable or predictable over time. In other words, is it follows the Efficient Market Hypothesis (EMH) or not?

According to [4], markets are regarded as totally effective when they follow a random (weak-form) model in which future returns on the basis of prior data cannot be forecasted. Many recent studies have been conducted to investigate the efficiency and volatility characteristics of cryptocurrency prices. According to Palamalai et al. [5] the returns from the top cryptocurrencies show a persistency impact, indicating market inefficiency, i.e., not a random-walk process. Their findings have immediate ramifications for cryptocurrency market speculation. For example, the idea that bitcoin history data retains some degree of predictability about future values is correct. Another approach to figuring out predictability for the bitcoin problem is to see whether external elements may be integrated as additional features. Besides these, there are prediction strategies that are already in-use by traders. Most of them are based on heuristics and empirical conclusions, such as "Engulfing Pattern" and "Evening Star". Furthermore, some studies have also shown that social media sentiment analysis, particularly tweets about trading activities, have considerable prediction potentials, check Fig 1for factors that may contribute to the cryptocurrency pricing.

Based upon this discussion, prediction of price's value and movement direction (whether goes up or down) is a task that plays a vital role in the cryptocurrency economy. Even though the efficiency of the forecasting models has been improved in recent years, most improvements have been achieved by minimizing the error between predictions and real readings. However, due to high fluctuations rates, predicting the next exact value turned to be not that useful for trading purposes more than predicting whether it will go up or down. However, a review of relevant studies reveals that the majority of research activities have exclusively focused on the forecasting model's accuracy. Nonetheless, it is possible to further address the actual goal (which is increasing trading profitability) by setting the focus on the price movement directions rather than, or with addition to, the price values.

With this goal in hand, this work seeks to address the following research questions: 1) Are cryptocurrency prices predictable? And if so, is that due to inherent features of the pricing data? or, alternatively, because of external influences? 2) Which is more important in terms of profitability: forecasting the price value or predicting the price movement direction (increase or drop)? 3) What additions that deep learning algorithms potentially provide in this regard? Whether it's forecasting the price or deciding the price change's direction?

B. Organization of the Paper

The rest of the paper is organized as follows: Section II outlines various related and recent work in this new field with emphasis on deep learning models. Section III is devoted to this work; the dataset involved in the study, the features engineering, the proposed model architecture. Section IV shows the experiments setup, findings, and results. Finally, Section V concludes the study and discusses suggested future work.

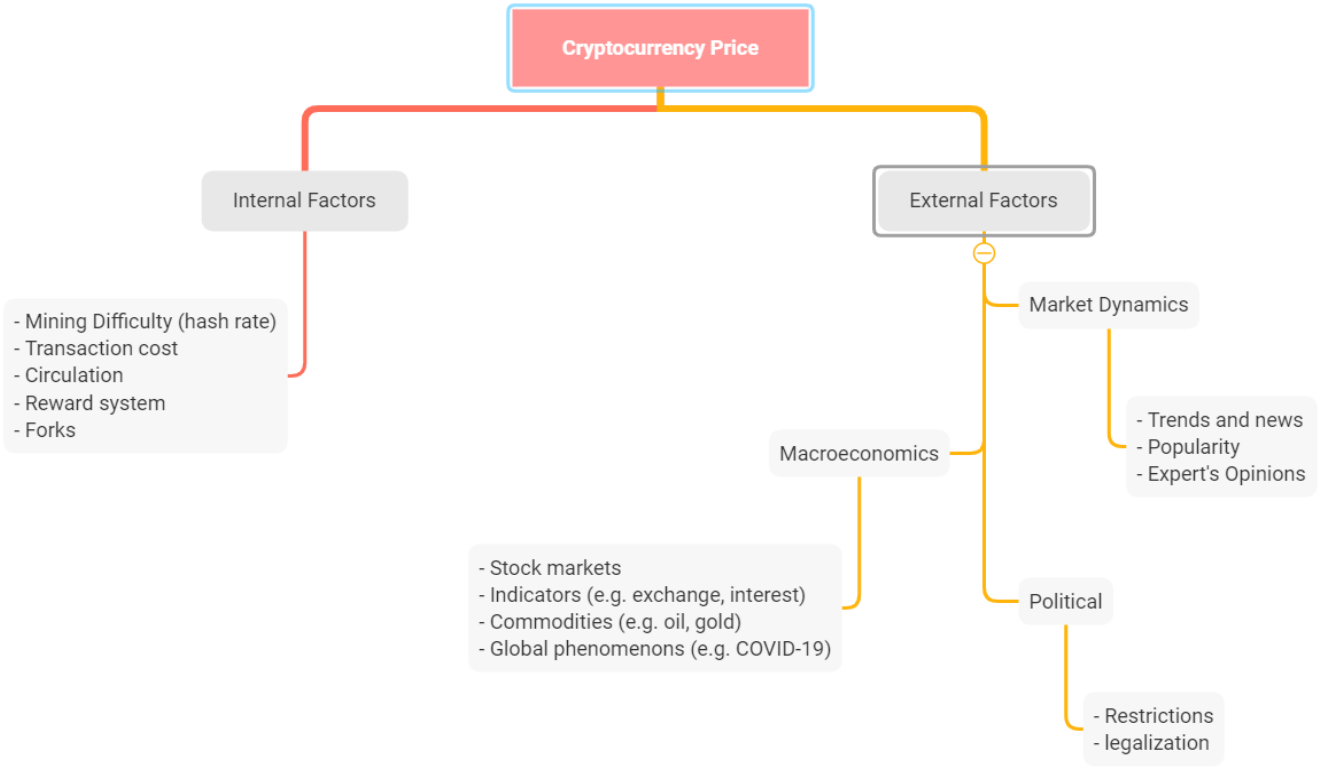


Fig. 1. Factors Affecting Cryptocurrency Prices.

II. Related Work

Over the last several years, there has been a significant surge in research interest in the subject of cryptocurrencies. Since 2015, the number of publications in this field has been increasing, and this trend is expected to continue in 20201, attributed to the growing popularity of cryptocurrencies and their widespread attention as an emerging financial sector. The field of cryptocurrency price prediction is still relatively new, with not much papers having been published; however, it is a very important topic that deserves to be explored because of the impact it is having on the financial system. Also, research trends are mostly focused on forecasting price.

A. Statistical and Traditional Techniques

A recent study by Palamalai et.al [5] evaluated the weak-form market efficiency of a top ten cryptocurrencies. They used random-walk testing with parametric and non-parametric methods which are resistant to "unknown structural breakdowns" and "asymmetric effects". The results confirmed the cryptocurrencies' weak-form inefficiency, and that it refuted the random-walk hypothesis. In essence, the results said that there is potential predictability within the cryptocurrency trading process. Their findings were consistent with which earlier research in 2017 (Hayes et al [3]) that demonstrated the impact of external factors on cryptocurrency predictability.

Holt-Winters exponential smoothing is a classic linear method for time-series forecasting that is commonly used. It splits the input data into several trend components. These components are used to forecast targets that have seasonal characteristics. Peng [6] argued that such traditional linear methods cannot be used to accurately predict cryptocurrency prices because they lack seasonality.

Chu et.al, [1] developed a GARCH method-based models. They investigated the log returns of exchange rates. The maximum likelihood method was used to fit the data. The models were evaluated based on several criteria, and the best-fitting models with the best forecasting performance were selected. Recently, Malladi et.al, [7] investigated the relationships between both Bitcoin and Ripple renderings and volatility, the authors used the Autoregressive Mobiles Average Exogenous Input (ARMAX), the Autoregressive Generalized Conditionally Heteroscedastic (GARCH) model, the Vector Autoregression (VAR) model, and Granger causality tests and showed that the Bitcoin crash of 2018 might have been described by these time series approaches. They also discovered that global stock and gold returns do not cause Bitcoin's returns, but Ripple's returns have a direct impact on Bitcoin pricing.

The impact of the three factors on cryptocurrency value was studied by Hayes et al [3] using cross-sectional data from 66 of the most commonly used cryptocurrencies. The factors are competition levels in the producers' network, production rates of the unit, and the algorithm difficulty used to mine crypto-monetary activity. The authors demonstrated that the three factors have shown a significant impact on price. Blau et.al, [8] used the regression to analyze the impact of the speculative trading, a 5-day bitcoin return, a 5-day volume sale, excess bitcoins, and a volatility estimate on the price of bitcoin of a number of other cryptocurrencies. The results showed that speculative trade has no impact on bitcoin prices. Speculative trade is the trade in future contracts, without obtaining the commodity that underlies it in reality. These traders buy or sell future contracts to re-sell it before their settlement date.

Another early proposition is described in [9] that applied autocorrelation and partial autocorrelation functions. The authors found that the first difference in the Bitcoin exchange rate is a weak-stationary time-series dataset. They created an ARIMA model to forecast future prices. They showed that ex-post forecasting's mean absolute percent error is 5.36 percent. In this work and in related references [14] it was investigated multiple variables that affecting the values of four cryptocurrencies (Bitcoin, Ethereum, Dash, litecoin, and Monaro). It has been shown that using weekly data and the enhanced Dickey–Fuller unit-root test and bound testing technique from 2010 to 2018. Market exchange, trading volume, and volatility all have an influence on the values of all four cryptocurrencies in the short and long run. This research used ARIMA as a forecast model.

Recently, Akcora et al. [10], [11], [12] have introduced a novel concept of chainlets, or blockchain motifs, to create a complex network of financial interactions on the blockchain that can be used to investigate link between bitcoin risk investing and different blockchain network aspects. Chainlets allow researchers to examine the influence of the blockchain's local topological structure on Bitcoin and Litecoin price development and evolution.

The authors in [13], [14], [15] described a strategy for forecasting changes in Bitcoin and Ethereum values using Twitter data and Google Trends [16]. Twitter is rapidly being utilized as a news source, alerting users about the currency and its rising popularity, affecting purchasing decisions. They discovered that tweet volume is a better predictor of price direction than tweet sentiment, which is always positive regardless of price direction. They were able to properly forecast the direction of price fluctuations using a linear model that takes as input tweets and Google Trends data. Aside from that, Shen et al. [17] argued that the number of tweets from Twitter, rather than Google trends, is a better indication of attention from more knowledgeable investors. They notice that the quantity of tweets has a substantial impact on the trade volume the next day.

Mohapatra et.al, [18] developed a novel KryptoOracle, a unique real-time and adaptable bitcoin price prediction engine based on Twitter emotions. The platform's integrative and modular design includes a Spark-based architecture for durable and fault-tolerant handling of massive amounts of incoming data, as well as real-time natural language processing including sentiment analysis, and an online learning-based prediction approach. Based on the accessible and ever-increasing volume and diversity of financial data, the experimental assessment indicated that the suggested platform may assist in the acceleration of decision-making, the identification of new possibilities, and the supply of more timely insights.

Bhambhwani et al. [19] used the dynamic ordinary least-squares approach was used to study the underlying determinants of cryptocurrency prices (Bitcoin, Ethereum, Monero, Litecoin, and Dash). The values of various currencies have been revealed to be influenced by their mining computer power and network connection. In [20], the authors looked into the correlation between Bitcoin market volatility and volatility in other traditional markets including gold, currencies, and stocks. On a daily, weekly, and monthly basis, they used data. Recent data shows a small but positive correlation between changes in Bitcoin volatility and changes in the trade weighted USD currency index volatility, according to correlations and regressions. Furthermore, a greater positive link between Bitcoin volatility and search demands on Bitcoin-related terms on Google has been observed. Furthermore, a VAR-analysis demonstrated that the volatility of the USD currency index, which is the lone driver of future Bitcoin volatility, is to some extent predictive of future Bitcoin volatility.

Vector autoregression method was investigated by Giudici et.al, [21]. The authors introduced partial correlations and correlation networks. The model assisted in determining the dynamics of cryptocurrency prices in various crypto exchange markets and allowed one to understand its correlation with other traditional market prices. The use of VAR correlation networks also allowed for the development of a model for predicting bitcoin price that makes use of the information contained in various correlation patterns among various exchange prices. Dos Santos et al. [22] investigated the dynamic behavior and predictability of bitcoin price dynamics (high and low). They used the fractionally cointegrated vector autoregressive (FCVAR) model to analyze bitcoin and dollar price patterns. The empirical analysis was carried out between January 2012 and February 2018. They compared fractionally cointegrated VAR to various other algorithms, and the results showed that fractionally cointegrated VAR performed better.

In [23] The authors utilized a modified Binary Auto Regressive Tree model to develop a short-term bitcoin price forecasting model (BART). BART is a hybrid method which mixes autoregressive models from the classification and regression trees (C&RTs). A short-term prognosis was made of the three most important cryptocurrencies, Bitcoin, Ethereum, and Ripple (from 5 to 30 days). The suggested technique has been identified in the prediction of crypto-currency time slower (fall) and transitional dynamics more accurately than ARIMA-ARFIMA models (change of trend). In RMSE according to the predicted horizon, the suggested model obtained 2,5 to 4,9 percent.

As reported by [15], sentiment analysis can be used as a computational tool to forecast the prices of bitcoin and other cryptocurrencies over various time intervals. The fact that currency prices fluctuate based on people's perceptions and opinions, rather than institutional money regulation, is a key feature of the cryptocurrency market. As a result, examining the relationship between web search and social media is crucial for projecting cryptocurrency prices. Because these social media platforms are used to influence purchasing decisions, this study forecasted the short-term prices of the major cryptocurrencies using Google Trends and Twitter. The study adopts and interpolates a novel multimodal approach to investigate the impact of social media on bitcoin pricing. The findings demonstrate that the psychological and behavioral views of people have a substantial influence on such highly speculative cryptocurrency values.

B. Machine Learning Techniques

Cryptocurrency is a highly volatile asset. Researchers were inspired to apply DL and ML paradigms to cryptocurrency issues as a result of this. The use of stock market price prediction techniques can help to improve precision. Because of its ability to identify the general trend and fluctuation, machine learning has become one of the most researched approaches in cryptocurrency price prediction in recent years.

One of the early work using machine learning techniques showed in [24], the authors used an SVM, an ANN, linear regression, and logistic regression to predict bitcoin price using blockchain data. The highest price accuracy was 55% for a NN classifier with two hidden layers, followed by logistic regression and SVM. In addition, the study mentions the use of several tree-based models and K-nearest neighbors in its analysis. In this study, only blockchain data was used for training and prediction, which resulted in limited predictability. According to the findings, using features directly extracted from bitcoin exchanges, such as financial flow features, would likely improve the accuracy of bitcoin price prediction. Hitam et.al, [25] forecasted the cryptocurrency future price using an optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO). The results of forecasting using basic SVM algorithms were found to be unreliable. Meanwhile, PSO's optimized version of SVM shows that it can accurately forecast future cryptocurrency prices, outperforming single SVM algorithms.

Mohanty et al. [26] used LSTM for bitcoin future price prediction. and Twitter data was used to predict public mood. This method selected some key features from the blockchain that had a significant impact on bitcoin demand and supply, and then used them to train a model that improved bitcoin price prediction in the future. The model demonstrated high precision and accuracy. Mittal *et al.* [27], identified the correlation between bitcoin price and Twitter and Google search patterns using machine learning techniques such as linear regression, polynomial regression, recurrent NN (RNN), and long short-term memory (LSTM)-based analysis. Tweet sentiment analysis performs the worst out of Google Trends, tweet volumes, and tweet sentiments. When LSTM, RNN, and polynomial regression were used to analyze Google Trends and tweet volume, the accuracy of the results improved.

Atsalakis *et al*, [28] proposed PATSOS, a neuro-fuzzy controller for predicting bitcoin's daily price change trend. The scheme outperformed two other computational intelligence models, the first of which was created using a simpler neuro-fuzzy approach and the second of which was created using artificial neural networks (ANNs). They also stated that the PATSOS system's performance was stable enough to be used for other cryptocurrencies.

The results obtained by applying hidden markov models in [29] to the historical cryptocurrency movements, and using LSTM to predict future movements suggested that the proposed approach had the lowest MSE, RMSE, and MAE when compared to traditional time-series prediction models, ARIMA, and conventional LSTM, demonstrating its effectiveness. This hybrid approach's parameters were further optimized using a genetic algorithm. This model did not account for the internal details of bitcoin transactions. As a result, they planned to consider additional features for future work in order to provide more information about the blockchain.

Radityo et.al, [30] showed for the next-day prediction, four variants of ANN were used to use the bitcoin exchange rate (closing price) on the American dollar. They Studied variety of ANN methods to predict the market value of one of the most used cryptocurrencies, Bitcoin. The ANN methods will be used to develop model to predict the close value of Bitcoin in the next day (next day prediction). This study compares four ANN methods, namely backpropagation neural network (BPNN), genetic algorithm neural network (GANN), genetic algorithm backpropagation neural network (GABPNN), and neuro-evolution of augmenting topologies (NEAT). The methods are evaluated based on accuracy and complexity. The result of the experiment showed that BPNN is the best method with MAPE 1.998 ± 0.038 % and training time 347 ± 63 seconds. Maiti et.al, [31] proposed a forecasting approach based on the chaotic co-movement of seven major cryptocurrencies, non-linear forecasting models have been proposed and implemented. For lags 0 and 0-3 the LSTM outperforms the ANN, while for large lags 0-7 the ANN outperforms the LSTM. Further research confirms that forecasting using variables like volume is ineffective in any case. In [32], As machine learning techniques, the authors used ARIMA, *FBProphet*, and XGBoosting for time series analysis. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R2 are the parameters they used to evaluate these models. They tested these three techniques, but time series analysis revealed that ARIMA was the best model for forecasting Bitcoin price in the crypto market, with an RMSE score of 322.4 and an MAE score of 227.3.

III. METHODOLOGY

We use a new approach that achieves two convergent tasks: forecasting the price and classifying the price movement. The intuition is that the focus is not only on the next point's prediction value, but also on the price movement direction whether it's going up or falling down. The determination of whether prices will rise, or fall is critical, particularly for traders. The proposed method focuses on utilizing LSTM layers' ability to extract useful knowledge by learning the internal representation of cryptocurrency features as well as correlation with external factors features. LSTM layers are used to identify short and long-term dependencies between time-series data, which are then used to predict the next point value and answer the question of whether price will fall or rise. Next sections discuss the underlaying hypothesis, the data used in the experiments and the proposed model architecture.

We formulate the price movement direction problem as follows; let T be the time divided into a timestamp $t = T/N$ where N is number of observations over T. t length maybe in minutes, hours, days, or any interval. Suppose $\mu \in \mathbb{R}$ is the price change value from t_{k-1} to t_k, where μ sign indicates the direction of change. The goal is to build a data-driven model to achieve two objectives; 1) classify μ_k sign at time t_k, and 2) predict the cryptocurrency price value at tine t_k , given observations of multiple timeseries features within time interval $\{t_0 - t_{k-1}\}$.

A. Cryptocurrency Performance Metrics

For point forecast, standard RMSE and MAPE are both used. For price direction detection, standard classification metrics are used. In this study, we use roc-auc and accuracy (which could be interpreted on a time range simulation into profitability measure).

B. Raw Datasets

The top popular cryptocurrencies with the highest market capitalizations are: Bitcoin ($577.36), Ethereum ($212.79), Tether ($62.00), and Binance Coin ($47.29) in Billions of dollars "https://coinmarketcap.com/all/views/all/" at the time of this writing. **Error! Reference source not found.** presents further information about datasets in hand.

TABLE I. TOP-TRADED CRYPTOCURRENCIES

Crypto	Data date range	Samples
Bitcoin (BTC)	7/18/2010 to 7/7/2021	4000
Ethereum (ETH)	3/10/2016 to 7/7/2021	1947
Tether (USDT)	4/14/2017 to 7/7/2021	1546
Binance Coin (BNB)	11/9/2017 to 7/7/2021	1338

1) Bitcoin (BTC): Market cap: Over $641 billion. Is the first cryptocurrency, founded in 2009 under the alias Satoshi Nakamoto [33]. Bitcoin's value has soared as it has grown in popularity. Five years ago, a Bitcoin could be purchased for around $500. A single Bitcoin was worth more than $32,000 in June 2021. This equates to a 6,300 percent increase. Plotting on Fig 2 visualizes Bitcoin data since its inception in 2010. In 2018, the price rose briefly before dropping dramatically in which is now known as bitcoin bubble (also known as the Bitcoin crash and the Great crypto crash). And in 2021, a much increasing tendency emerged.

2) Ethereum (ETH): Market cap: Over $307 billion. Ethereum, which is both a cryptocurrency and a blockchain platform, is a favorite among programmers due to the potential applications it offers, such as smart contracts that execute automatically when certain conditions are met and non-fungible tokens (NFTs). Ethereum has also exploded in popularity. Its price increased by more than 22,000 percent in just five years, from around $11 to over $2,500.

3) Tether (USDT): Market cap: Over $62 billion. Tether, unlike some other types of cryptocurrency, is a stable coin, which means it is backed by *fiat currencies* such as US dollars and the Euro and theoretically maintains a value equal to one

of those denominations. In theory, this means that Tether's value should be more consistent than other cryptocurrencies, and it is preferred by investors who are wary of the extreme volatility of other coins.

4) Binance Coin (BNB): Market cap: Over $56 billion. The Binance Coin is a type of cryptocurrency that can be used to trade and pay fees on Binance, one of the world's largest cryptocurrency exchanges. Binance Coin has grown beyond simply facilitating trades on Binance's exchange platform since its inception in 2017. It is now possible to use it for trading, payment processing, and even booking travel arrangements. It can also be traded or exchanged for other cryptocurrencies like Ethereum or Bitcoin. It was only $0.10 in 2017; by June 2021, it had risen to over $350, a gain of about 350,000 percent.

Each cryptocurrency dataset comprises of 5 indicators: open, low, high, close and volume. All data are acquired from https://www.investing.com API webservices via python library "investpy". investing.com is a global platform that delivers financial market analyses and news from across the world.

Besides historical cryptocurrencies datasets, other external factors datasets are used as shown in **Error! Reference source not found.**. Factors investigated in this study include:

1) Gold and Brent oil daily prices (open and volume values).

2) NYSE Bitcoin (NYXBT) index; the bitcoin index on the New York Stock Exchange. NYXBT indicates the US dollar value of one bitcoin unit based on actual transactions on selected bitcoin exchanges that have been assessed and meet NYSE quality criteria. The NYSE Bitcoin Index integrates data from Coinbase Exchange, the largest bitcoin exchange in the United States, in which NYSE has a minority investment [34].

3) The Standard and Poor's 500 or (SPX). SPX is a stock market index that tracks the performance of 500 big firms that are listed on US stock exchanges. As of December 31st, 2020, more than $4.6 trillion has been invested in assets tied to the index's performance.

4) The Bitcoin and Ethereum Energy Consumption Index (PWR) is a measure of how much energy is consumed by the cryptocurrency mining in a certain time period available at "https://digiconomist.net". The data found online only available for Bitcoin and Ethereum.

5) Google Trends, https://trends.google.com/ . The daily historical data for two specific keywords: ("bitcoin", "Ethereum", "Tether", "Binance Coin", and "cryptocurrency") obtained from Google API webservices via the "pytrends" python module. Google Trends examines the popularity of top search queries in Google Search across several countries and languages. The website makes use of graphs to compare the search volume of various queries over time.

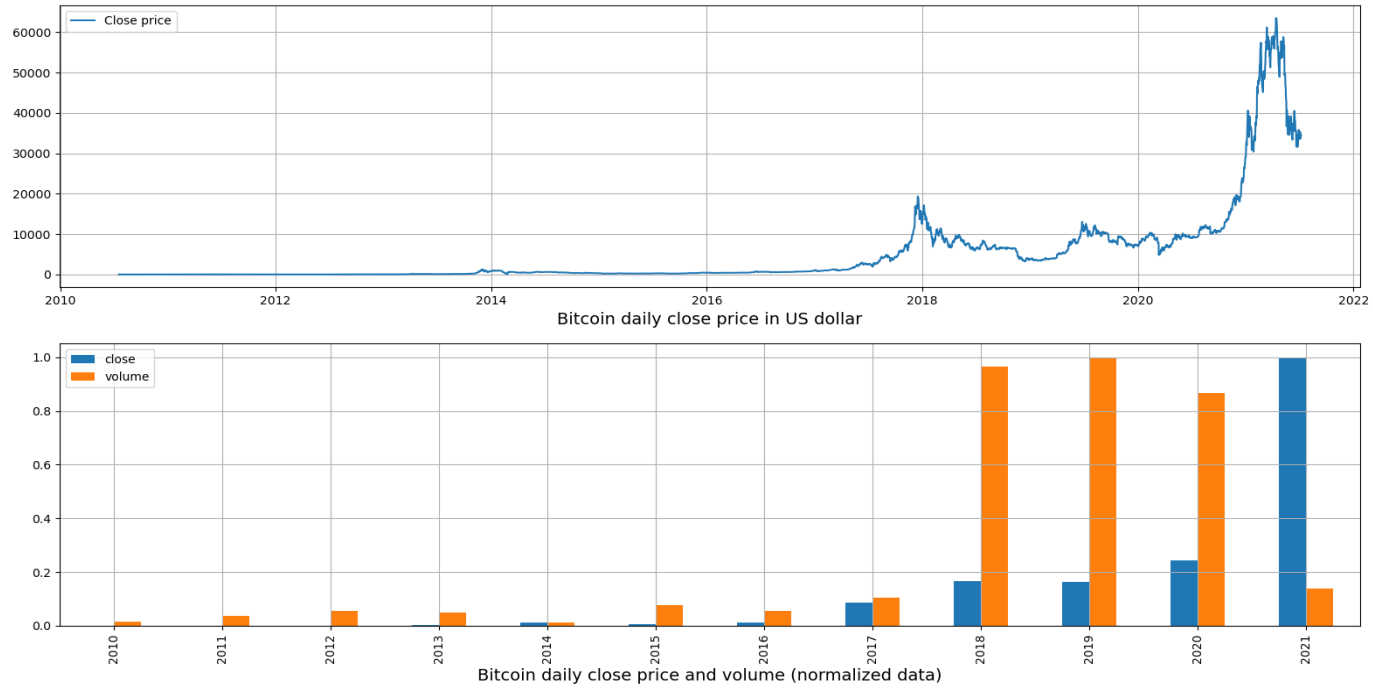


Fig. 2. Bitcoin Dataset Properties.

TABLE II. EXTERNAL FACTORS

Features	Description
gold_close	Gold commodity material daily price
oil_close	Brent oil daily price
nyxbt	NYSE Bitcoin index
spx	Standard and Poor's 500
pwr, pwr_min	Bitcoin Energy Consumption Index (expected estimation and minimum).
gtrend_bitcoin	Google Trends for the cryptocurrency name keyword
gtrend_crypto	Google Trends "cryptocurrency" keyword

C. Features Engineering

In this work, the strategy is to leverage the availability of various data sources as predictors for cryptocurrency price fluctuations and forecasting. In terms of which features are chosen for training, the features selection method is adaptive. The most important fundamental requirements are to find the right balance between training time and prediction accuracy. The proposed work dynamically infers the prediction power for each feature by using *predictive power scoring (PPS)* and *correlation coefficient* methods. PPS is an asymmetric, data-type agnostic tool for detecting linear and non-linear correlations between two columns. Values range from 1 (full predictive power) to zero (complete absence of predictive power). Asymmetric here in the sense that correlation between A-to-B is not the same as B-to-A. PPS uses a decision tree regression algorithm.

1) Features selection (Phase I): We use PPS, as a first phase filtration, to disregard features that have not much significance to the target prediction as PPS has the ability to detect non-linear relations.

2) Data preprocessing: We begin data preprocessing by normalizing the features. Normalization method is done by squashing the data from its original numerical scale to $[0, 1]$ range. We make use of "*MinMaxScaler*" found in *sklearn* library. Normalization is essential step for neural networks innerworkings as unscaled data with higher values tend to wrongly dominate when calculating derivates by the optimizer. Also, normalization prevents some activation functions from getting numerically saturated. For detailed information please check [35]. Second, we devise a set of new fabricated features as following:

Crossing original features.

Encoding date minutes information.

Time lagging features.

3) Features Selection (Phase II): We continue the feature selection process by computing the linear statistical relationship between the target and candidate predictor variables with measures such as Pearson, Kendall, and Spearman. Pearson Correlation Coefficient (PCC) is used, which produces a value between [-1, 1], where an absolute value of one shows perfect correlation, the sign indicates whether the correlation is positive or negative, and zero indicates no linear dependency at all [36]. Fig 3 illustrates PPS and PCC analysis for bitcoin external features datasets. The target is the "close" price.

We think that the forecasting possibility may changes over time in a market as volatile and inefficient as cryptocurrency trading. To overcome this difficulty, it's necessary to adopt flexible features selection approach. We use an adaptive features selection strategy in where it selects the prospective features that have the best chance of being a good predictor throughout time.

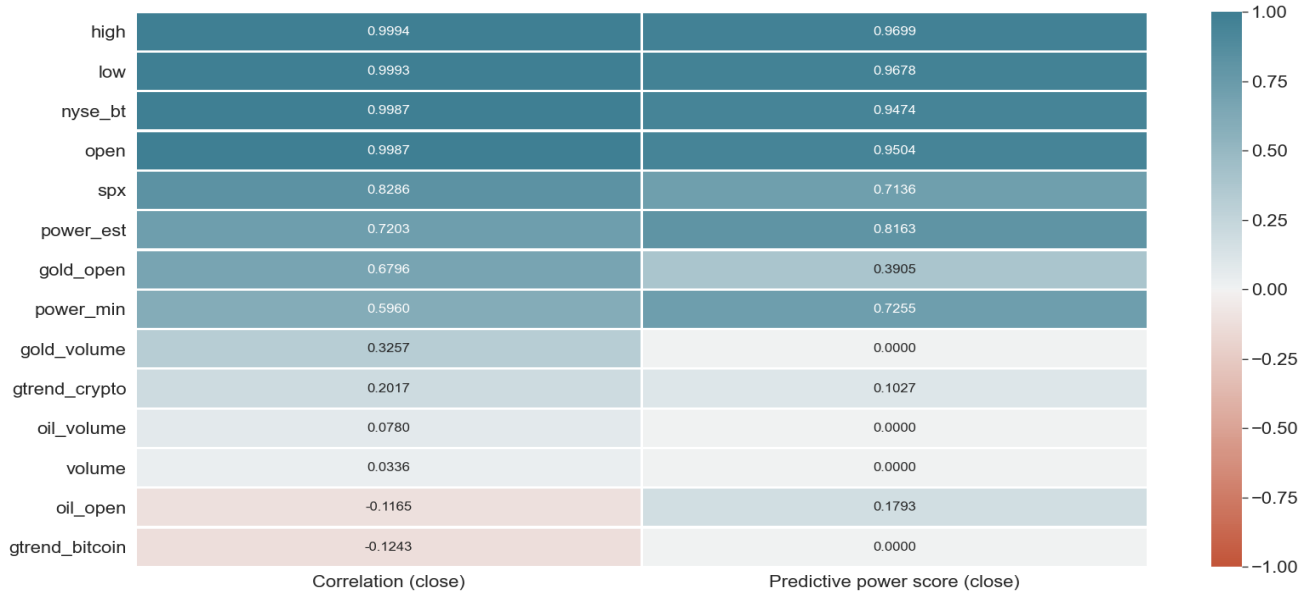


Fig. 3. Heat Map shows Correlation Coefficient and Predictive Power Scoring (Bitcoin External Features).

D. Proposed Model Architecture

The proposed model is built around a deep learning engine that receives two tensors from a preprocessed multivariate timeseries datasets as input ($X_p and\ X_d$) where $X_p = \{x_1^p, x_2^p, \dots, x_n^p\}$ and $X_d = \{x_1^d, x_2^d, \dots, x_n^d\}$, n is the number of training samples. The target variables ($Y_p and\ Y_d$) are the cryptocurrency closing price value and direction μ which is encoded as (1, -1) for up and down respectively. The prediction output vector $o = \{\hat{y}^p, \hat{y}^d\}$ where $\hat{y}^p \in \mathbb{R}$ is inferred from the learned reference distribution $p(Y^p)$. $\hat{y}^p$ is a value which when scaled back to its original scale represents the forecasted closing price. $\hat{y}^d$ is further marginalized by a threshold ε to label the closing price direction:

$$\mu^{sign} = \begin{cases} up \ \ if\ \hat{y}^d \geq \varepsilon \\ down\ if\ \hat{y}^d < \varepsilon \end{cases}$$

The proposed model consists of two stacks of layers for each input tensors $X_p and\ X_d$: 1) X_p stack consists of three GRU layers with 256 neurons each. Followed by MLP layer and an MLP output layer. 2) X_d stack consists of three LSTM layer with 256 neurons each. Followed by MLP layer, dropout layer and an MLP output layer. As shown in Fig 4, output of X_p stack GPU layer concatenates with X_d stack LSTM layer and feeds the X_d MLP layer. Layer counts, number of neurons per layer, batch sizes, training epochs, and other model hyper-parameters are set empirically.

The training algorithm works as the following steps:

1) Step 1: the PPS values are computed for all raw inputs to identify n factors with the largest predictability score. The target variable here is price "close".

2) Step 2: preprocessing features by normalizing data in range $[0, 1]$ and devising new cross features.

3) Step 3: computing PCC values and further filter features to the selected n features.

4) Step 4: train the neural network model on data.

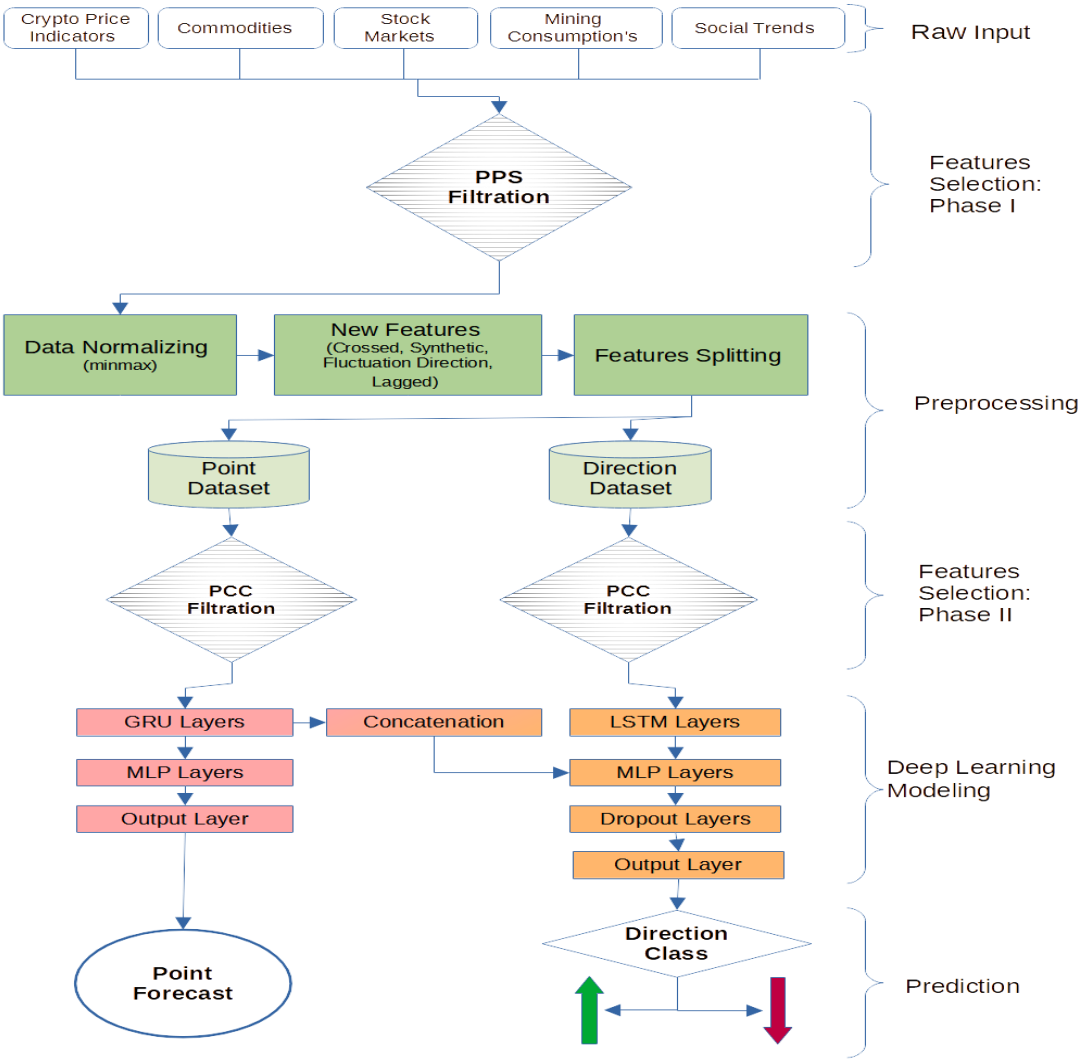


Fig. 4. System Diagram for the Proposed Cryptocurrency Model.

IV. Experiments and Results

A. Experiments Setup

We implemented the proposed model on python 3.8 platform. The deep learning library is Keras on top of TensorFlow. Hardware consists of nVidia GTX 1660Ti (6GB dedicated RAM), 16GB RAM and i7 10th generation 2.6GHz processor. Results shown are the highest of multiple trials with different combination of configurations. Average training time within range 65^{+}_{-} seconds with GPU accelerated computations. We split datasets in training /testing as 80% ratio.

For regression task, the model forecasts the price value of the next point. The regression branch of the model consists of two cascaded GRU layers of 256 neurons. Experiments show no benefit from applying recurrent dropouts. For the classification task of price movement direction, we trained the model with a list of two types of features: original and devised features.

B. Results and Findings

Results about bitcoin attained 92% accuracy score. We define accuracy as the overall proportion of correct predictions among the entire number of instances investigated. For regression, the proposed model achieved a score of 2.4 MAPE for bitcoin data. The overall performance metrics results are summarized in **Error! Reference source not found.**. Best classification achieved occurred with bitcoin data. As shown, the direction is successfully classified for all cryptocurrencies. Tether regression metrics are on top.

We can deduce no unique patterns or circumstances that are connected to the direction classification from the data. Fig 5 compares the classification of price movement direction to real data. The test data is shown in sub-figure (a) for the time period 3-2019 to 7-2021, with hits in blue circles and misses in red crosses. The data in sub-figure (b) is zoomed in to reveal one-month details. The true pricing data is shown by the green line.

TABLE III. Results Summary in Terms of MAPE and PR-AUC

Cryptocurrency		Regression		Classification	
	Mean ($)	**RMSE ($)**	**MAPE (%)**	**Accuracy**	**PR-AUC**
Bitcoin (BTC)	6690.87	850.8	2.28	92.1%	0.981
Ethereum (ETH)	425.6	82.03	3.17	90.26%	0.930
Tether (USDT)	0.999	0.0045	0.230	85.3%	0.853%
Binance Coin (BNB)	56.4	26.7	3.100	63.29%	0.743%

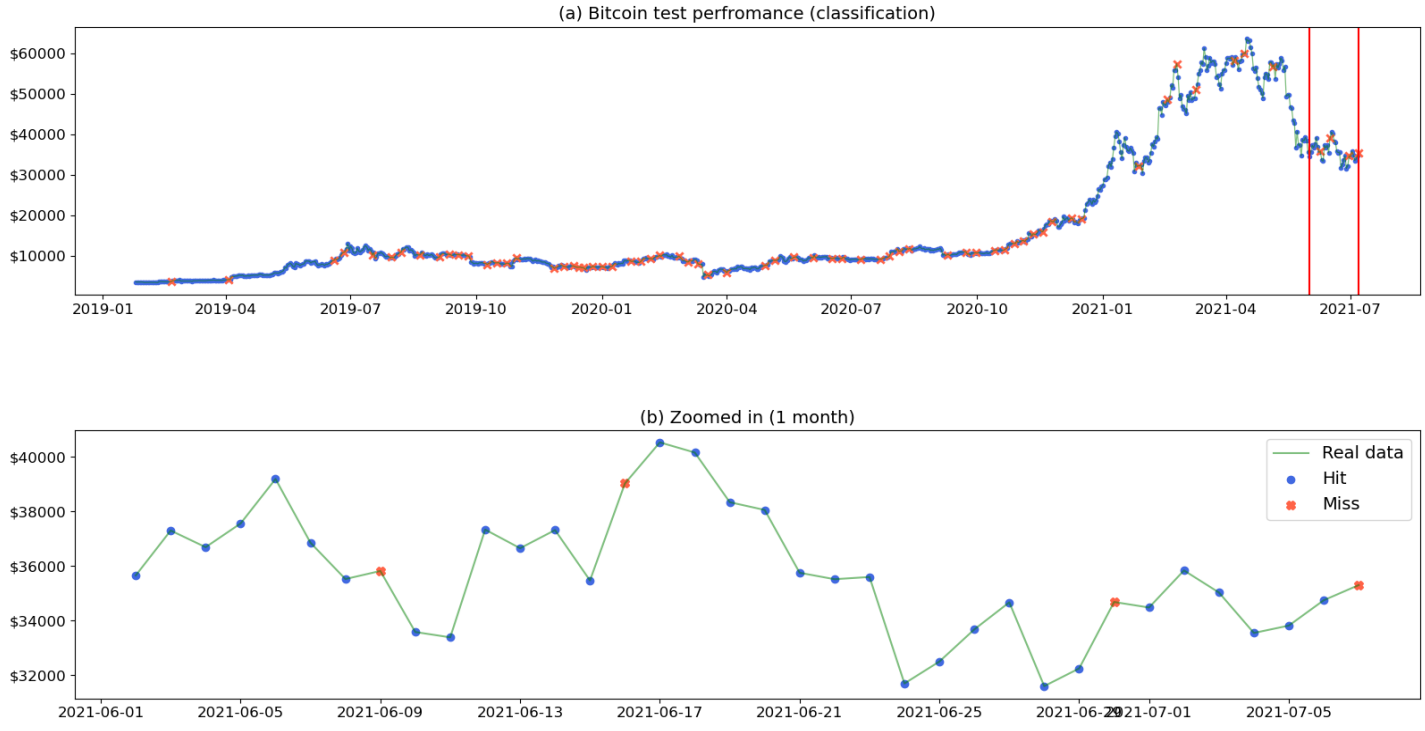


Fig. 5. Cryptocurrency Price Movement Classification Test Results (Hits and Misses).

Regression results shown in Fig 6 are in orange color. RMSE for bitcoin is 850.8. considering the scale of the bitcoin price which is extremely volatile from several thousands to tenth of thousands of dollars. RMSE don't reveal much about how good the regression results are. We can claim that MAPE is better representing the performance. In sub-figure (b), a one month zoom-in plotting illustrates the proposed model which performs a good job generalizing on unseen test data. In addition to forecast the next point, we forecast the 90% and 10% quantiles as a confidence boundaries (in sub-figure (b) as grey lines enclosing the real and forecasted values).

C. Comparison with Related Work

A Comparison with a three related deep learning work to the problem of interest reveals that the proposed model achieves highest accuracy in terms of classifying the next day price movement. From our perspective, adaptive features selection is a major contributor the results.

D. Profitability Simulation

We understand that traders must earn a profit at the end of the day. We use a basic trading algorithmic method to determine the projected profitability. We also use python code to implement it. The simulation method is naive since we simply want to illustrate how to utilize the anticipated results to prove the suggested model's profitability. Simulation results are shown as a waterflow chart in Fig 7. The simulation is based on the results of the tests (895 samples for bitcoin). The trade epochs were divided into 60 days period. The number of epochs is 15, with the first one beginning on March 3, 2019, and ending on July 7, 2021. The simulation resulted in 9 profits and 6 losses. The profit was 199038 dollars, with a return of 68.24 percent, on a maximum investment of 291690 dollars. Over time, the volume of profit and losses represents the rising and diminishing volatility of cryptocurrency trading prices and volumes, with the biggest earnings and losses accompanying the most spectacular jumps and falls, respectively.

E. Final Discussion

The predictive power score (PPS) method is used in Phase I of the feature selection procedure to exclude features that have little potential for cryptocurrency prediction. When using PPS, there are two variables to consider: 1) the threshold, which is the PPS score value used to determine if a feature is acceptable or not. 2) The effective sample size, which is the number of samples needed to compute the PPS score, or the cryptocurrency time period over which we are looking for the best predictors. The effect of these two factors on MAPE and accuracy is depicted in Fig 8. When we ignore predictors with a score of less than 0.3, as shown in sub-figure (a), the least regression MAPE value happens. In sub-figure (b), a score of 0.5 results in the most accurate price movement direction classification. When we limit the effective sample size to the last 500 samples of the training data in (c), we get the lowest MAPE value, and when we increase the size to 1000 samples, we get the maximum accuracy as shown in (d).

TABLE IV. COMPARISON WITH RELATED WORK

Model	Technique	Accuracy
RNN-based ensemble [37]	RNN-based ensemble with decision tree classifier	62.91%
CNN-BiLSTM [38]	Convolutional-based bidirectional LSTM	55.43%
Bayesian optimized LSTM [39]	Bayesian optimization LSTM	76.83%
The proposed model	RNN-based with adaptive features selection	**93.1%**

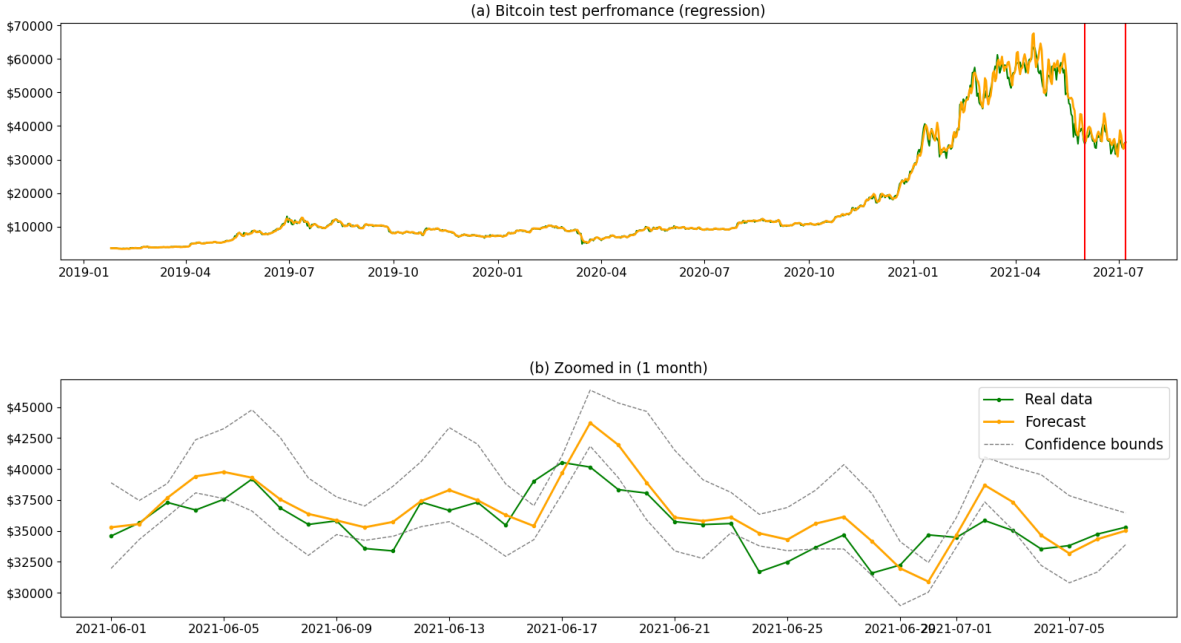


Fig. 6. Cryptocurrency Price Forecast Results (Next Point Forecast).

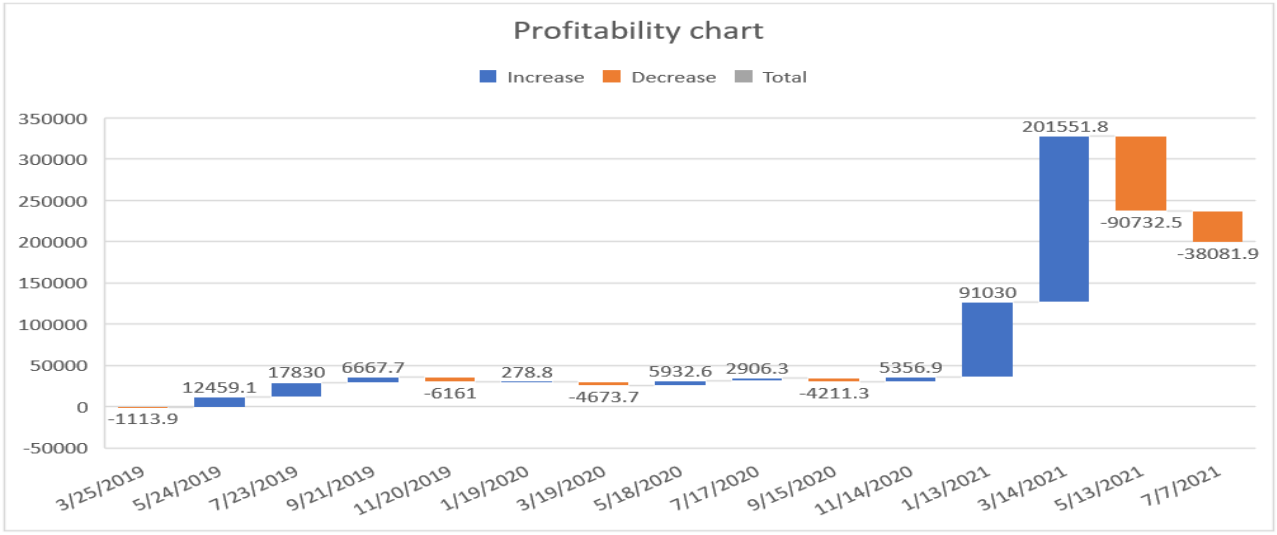


Fig. 7. Profitability Simulation Results.

The model's ability to forecast and classify accurately is influenced by the values of recurrent timesteps. Formally, the model needs to learn the probability of a sample y_{i+1} at time $i+1$, given the context $x_{i-t:i}$ of earlier t input features samples: $p(y_{i+1}|x_{i-t:i})$. RNN (LSTM, GRU, and other recurrent neural network variants) incorporate past data in addition to the current data point, which is the timestep parameter t passed to the RNN layers. Contextual-based prediction increases model performance. We may consider that point inside the problem of interest as follows: we use the earlier n days to properly forecast the price value and direction of the next day. $\widehat{y_i} = model(X_{i-t}^{i})$, where $\widehat{y_i}$ is the model prediction at point i, X_{i-t}^{i} input vector of t number of samples prior to i time, $t \in \mathbb{N}$ is the timesteps parameter. Sub-figures (a) and (b) shows model behavior with different timesteps values. Apparently, a timestep value of "7" gives lowest MAPE and highest accuracy.

The predictors vector is shifted back into the time domain using time lags. $lag_l = x_{t-l}$, where $l \in \mathbb{N}$ is the lag value, t is the time point. As an example, $lag_1$is using the previous day $t-1$ to predict the coming day t. $lag_2$is using the day before $t-2$ to predict the coming day t. Usually in seasonal data with high stationarity, autocorrelation function (ACF) analysis computes which time lag is best to be used as a predictor. In addition, in timeseries forecasting lags could be looked in as the forecast horizon. However, in case in hand, the situation is different since cryptocurrency timeseries data is a random walk and non-stationary, i.e., no seasonality, no steady statistical moments values and high variability. Augmented Dickey Fuller Test (ADF Test) on bitcoin "close" data give a p-value of "0.738698" which can be interpreted as *strong non-stationarity* property ($p-value > 0.05$). Fig 9 shows time lags from "1" to "7", lag_1 gives best performance. also, we notice accuracy keeps above 60% till lag_5, that means profitability still possible till forecasting horizon of 5-days (we limit the scope to daily data, forecasting long-term periods such as months requires a different data frequencies and different strategies).

In deep learning, hyper-parameters tuning task requires a lot of hands on efforts, and sometimes the task is automated [40]. Fig 10 depicts two model hyper-parameters: neurons per recurrent layers and dropouts. As we can notice, for regression layers, a "128" gives least MAPE and for classification, a "256" neurons give highest accuracy. For both tasks, a dropout of 0.1 is enough to have a positive impact on the predictions.

: 1) Are cryptocurrency prices predictable? And if so, is that due to inherent features of the pricing data? or as a result of external influences? 2) Which is more important in terms of profitability: predicting the next price value or predicting the future price direction (increase or drop) over a forecasting horizon? 3) Can deep learning algorithms accurately forecast the price of cryptocurrencies?

We have presented applied research to address both problems in cryptocurrency prices fluctuations, the next point forecast and next price movement direction classification. We have included several features grouped into cryptocurrency indicators ("open", "close", "low", "high" and "volume") and external factors gathered from various sources, including "SPX", "NYCE_BT", "GTrends_CRYPTO", and others.

The results of the experiments and the profitability simulation done in this study show that we can develop predictive model for two tasks using deep learning methods, particularly, recurrent neural networks algorithms: regression of the future pricing values and classification of the direction of price movement. A careful features selection strategy is mandatory to successfully identify good predictors and how to include them in learning process. Such results answer the third research question: "Can deep learning methods accurately forecast the price of cryptocurrencies and classify its movement direction?"

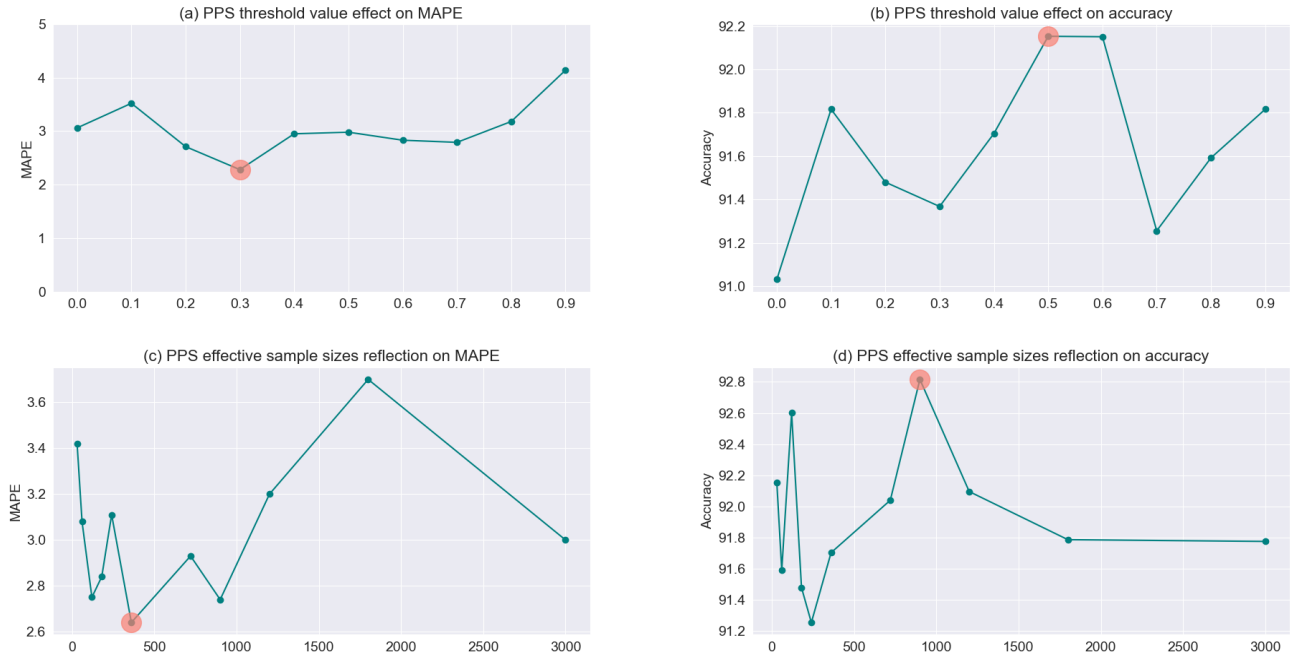


Fig. 8. PPS Settings Effect on Accuracy and MAPE.

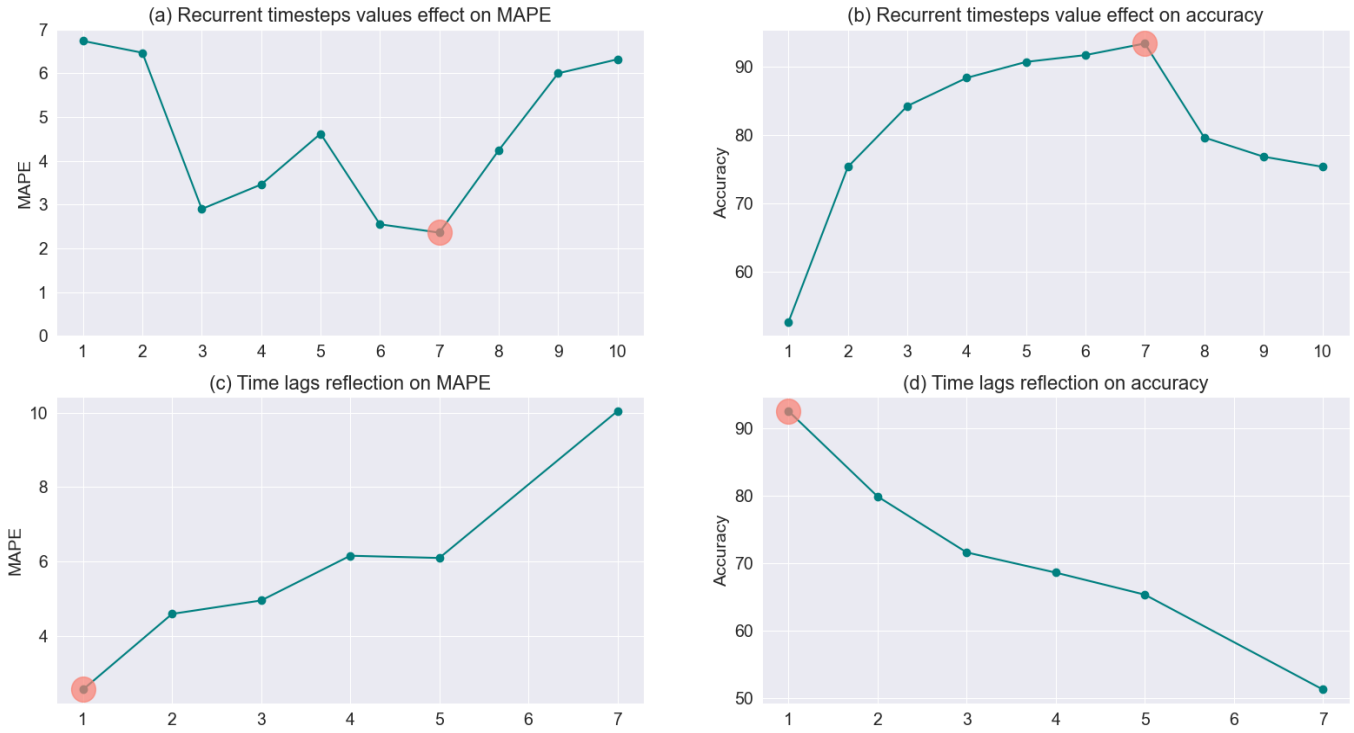


Fig. 9. Timesteps and Forecasting Lags Effect on Performance.

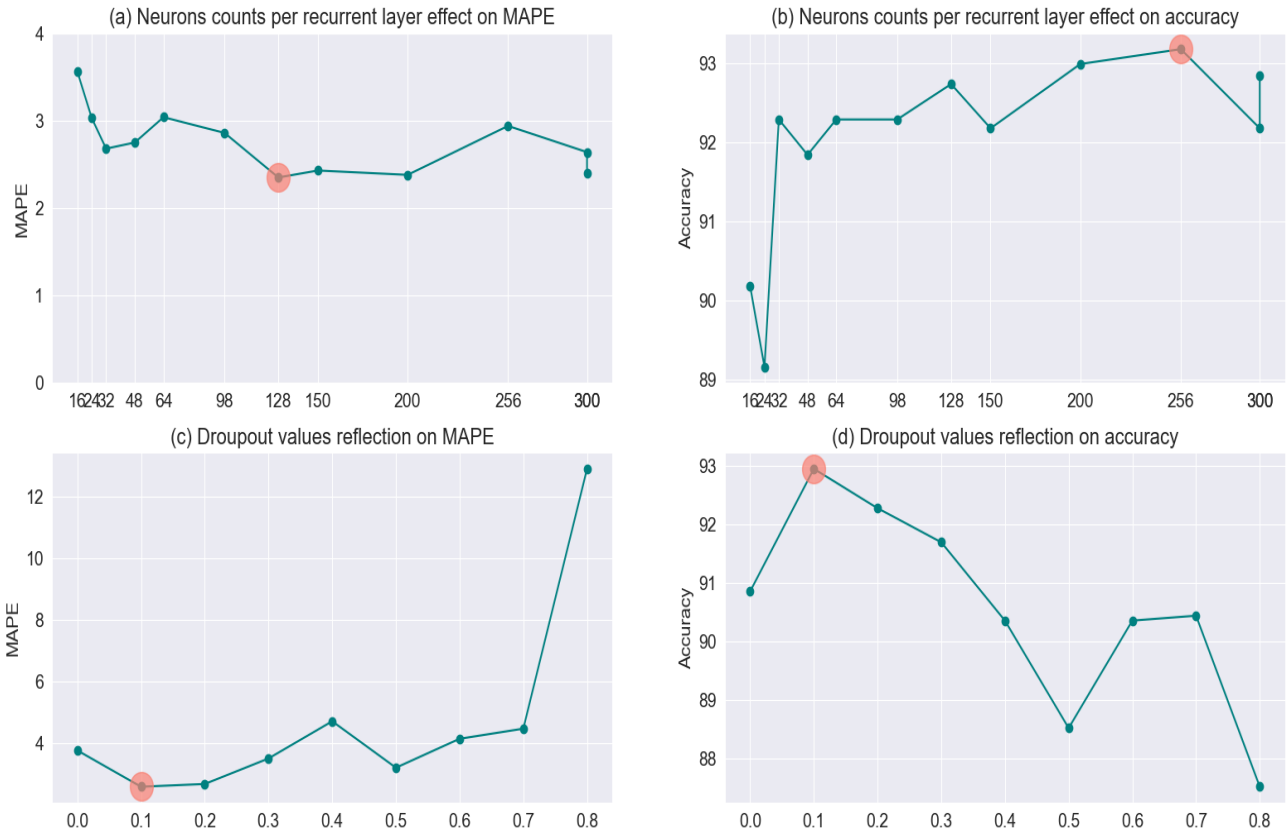


Fig. 10. Units and Dropouts Parameters Effect on Performance.

V. Conclusions and Future Work

This research presents a deep neural network model for bitcoin price/movement prediction based on a multi-input architecture. To determine which variables to include, a two-phase adaptive feature selection process is applied. As a result of this method, the data provided to the neural network has the maximum predictive power for the price. The results suggest that using an adaptive feature selection technique to improve classification performance has a significant influence. Future study could consider using more advanced deep learning architectures, such as attention mechanisms, to focus on key patterns in data. Another idea is to look at generative adversarial architectures for various speculative scenarios with varying levels of confidence. In addition, other external elements that correlate or have a degree of predictability with the bitcoin price movement should be included.

References

[1] J. Chu, S. Chan, S. Nadarajah, and J. Osterrieder, "GARCH Modelling of Cryptocurrencies," J. Risk Financ. Manag., vol. 10, no. 4, p. 17, 2017, doi: 10.3390/jrfm10040017.

[2] P. Katsiampa, "An empirical investigation of volatility dynamics in the cryptocurrency market," Res. Int. Bus. Financ., vol. 50, pp. 322–335, 2019, doi: 10.1016/j.ribaf.2019.06.004.

[3] A. S. Hayes, "Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin," Telemat. Informatics, vol. 34, no. 7, pp. 1308–1321, 2017, doi: 10.1016/j.tele.2016.05.005.

[4] B. G. Malkiel, "The efficient market hypothesis and its critics," J. Econ. Perspect., vol. 17, no. 1, pp. 59–82, 2003.

[5] S. Palamalai, K. K. Kumar, and B. Maity, "Testing the random walk hypothesis for leading cryptocurrencies," Borsa Istanbul Rev., 2020, doi: 10.1016/j.bir.2020.10.006.

[6] C. Peng and G. Yichao, "Cryptocurrency Price Analysis and Time Series Forecasting," no. April, 2020.

[7] R. K. Malladi and P. L. Dheeriya, "Time series analysis of Cryptocurrency returns and volatilities," J. Econ. Financ., vol. 45, no. 1, pp. 75–94, 2021, doi: 10.1007/s12197-020-09526-4.

[8] B. M. Blau, "Price dynamics and speculative trading in bitcoin," Res. Int. Bus. Financ., vol. 41, pp. 493–499, 2017.

[9] N. A. Bakar and S. Rosbi, "Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction," Int. J. Adv. Eng. Res. Sci., vol. 4, no. 11, pp. 130–137, 2017, doi: 10.22161/ijaers.4.11.20.

[10] C. G. Akcora, A. K. Dey, Y. R. Gel, and M. Kantarcioglu, "Forecasting bitcoin price with graph chainlets," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10939 LNAI, pp. 765–776, 2018, doi: 10.1007/978-3-319-93040-4_60.

[11] C. G. Akcora, M. F. Dixon, Y. R. Gel, and M. Kantarcioglu, "Bitcoin risk modeling with blockchain graphs," Econ. Lett., vol. 173, pp. 138–142, 2018, doi: 10.1016/j.econlet.2018.07.039.

[12] A. K. Dey, C. G. Akcora, Y. R. Gel, and M. Kantarcioglu, "On the role of local blockchain network features in cryptocurrency price formation," Can. J. Stat., vol. 48, no. 3, pp. 561–581, 2020, doi: 10.1002/cjs.11547.

[13] V. Derbentsev, V. Babenko, K. Khrustalev, H. Obruch, and S. Khrustalova, "Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices," Int. J. Eng. Trans. A Basics, vol. 34, no. 1, pp. 140–148, 2021, doi: 10.5829/IJE.2021.34.01A.16.

[14] M. M. Patel, S. Tanwar, R. Gupta, and N. Kumar, "A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions," J. Inf. Secur. Appl., vol. 55, no. May, 2020, doi: 10.1016/j.jisa.2020.102583.

[15] K. Wołk, "Advanced social media sentiment analysis for short-term cryptocurrency price prediction," Expert Syst., vol. 37, no. 2, pp. 1–16, 2020, doi: 10.1111/exsy.12493.

[16] J. Abraham, D. Higdon, and J. Nelson, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," SMU Data Sci. Rev., vol. 1, no. 3, p. 22, 2018, [Online]. Available: https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/1.

[17] D. Shen, A. Urquhart, and P. Wang, "Does twitter predict Bitcoin?," Econ. Lett., vol. 174, pp. 118–122, 2019, doi: 10.1016/j.econlet.2018.11.007.

[18] S. Mohapatra, N. Ahmed, and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments," Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019, pp. 5544–5551, 2019, doi: 10.1109/BigData47090.2019.9006554.

[19] S. Bhambhwani, S. Delikouras, G. M. Korniotis, and others, Do fundamentals drive cryptocurrency prices? Centre for Economic Policy Research, 2019.

[20] H. Bystrom and D. Krygier, "What Drives Bitcoin Volatility?," SSRN Electron. J., 2018, doi: 10.2139/ssrn.3223368.

[21] P. Giudici and I. Abu-Hashish, "What determines bitcoin exchange prices? A network VAR approach," Financ. Res. Lett., vol. 28, pp. 309–318, 2019.

[22] L. dos Santos Maciel and R. Ballini, "On the predictability of high and low prices: The case of Bitcoin," Rev. Bras. Finanças, vol. 17, no. 3, pp. 66–84, 2019.

[23] V. Derbentsev, N. Datsenko, O. Stepanenko, and V. Bezkorovainyi, "Forecasting cryptocurrency prices time series using machine learning," CEUR Workshop Proc., vol. 2422, pp. 320–334, 2019.

[24] A. Greaves and B. Au, "Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin," pp. 1–8, 2015.

[25] N. A. Hitam, A. R. Ismail, and F. Saeed, "An Optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO) for Cryptocurrency Forecasting," Procedia Comput. Sci., vol. 163, pp. 427–433, 2019, doi: 10.1016/j.procs.2019.12.125.

[26] P. Mohanty, D. Patel, P. Patel, and S. Roy, "Predicting Fluctuations in Cryptocurrencies' Price using users' Comments and Real-time Prices," 2018 7th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2018, no. August, pp. 477–482, 2018, doi: 10.1109/ICRITO.2018.8748792.

[27] A. Mittal, V. Dhiman, A. Singh, and C. Prakash, "Short-term bitcoin price fluctuation prediction using social media and web search data," in 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1–6.

[28] G. S. Atsalakis, I. G. Atsalaki, F. Pasiouras, and C. Zopounidis, "Bitcoin price forecasting with neuro-fuzzy techniques," Eur. J. Oper. Res., vol. 276, no. 2, pp. 770–780, 2019, doi: 10.1016/j.ejor.2019.01.040.

[29] I. A. Hashish, F. Forni, G. Andreotti, T. Facchinetti, and S. Darjani, "A hybrid model for bitcoin prices prediction using hidden Markov models and optimized LSTM networks," in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2019, pp. 721–728.

[30] A. Radityo, Q. Munajat, and I. Budi, "Prediction of Bitcoin exchange rate to American dollar using artificial neural network methods," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2017, pp. 433–438.

[31] M. Maiti, Y. Vyklyuk, and D. Vuković, "Cryptocurrencies chaotic co-movement forecasting with neural networks," Internet Technol. Lett., vol. 3, no. 3, pp. 1–6, 2020, doi: 10.1002/itl2.157.

[32] M. Iqbal, M. S. Iqbal, F. H. Jaskani, K. Iqbal, and A. Hassan, "Time-Series Prediction of Cryptocurrency Market using Machine Learning Techniques," EAI Endorsed Trans. Creat. Technol., p. e4, 2021.

[33] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Bus. Rev., p. 21260, 2008.

[34] Treasurers.org, "NYSE launches Bitcoin Index," https://www.treasurers.org/, 2021. https://www.treasurers.org/hub/treasurer-magazine/nyse-launches-bitcoin-index (accessed Jul. 07, 2021).

[35] X. Wan, "Influence of feature scaling on convergence of gradient iterative algorithm," J. Phys. Conf. Ser., vol. 1213, no. 3, 2019, doi: 10.1088/1742-6596/1213/3/032021.

[36] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in Noise reduction in speech processing, Springer, 2009, pp. 1–4.

[37] D. C. A. Mallqui and R. A. S. Fernandes, "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques," Appl. Soft Comput. J., vol. 75, pp. 596–606, 2019, doi: 10.1016/j.asoc.2018.11.038.

[38] P. Pintelas, T. Kotsilieris, I. Livieris, E. Pintelas, and ..., "Fundamental research questions and proposals on predicting cryptocurrency prices using DNNs," 2020, [Online]. Available: http://nemertes.lis.upatras.gr/jspui/handle/10889/13296.

[39] M. Rahmani Cherati, A. Haeri, and S. F. Ghannadpour, "Cryptocurrency direction forecasting using deep learning algorithms," J. Stat. Comput. Simul., 2021, doi: 10.1080/00949655.2021.1899179.

[40] A. S. Elberawi and others, "A Deep Learning Approach for Forecasting Global Commodities Prices," Futur. Comput. Informatics J., vol. 6, no. 1, p. 4, 2021.

User-centric Activity Recognition and Prediction Model using Machine Learning Algorithms

Namrata Roy, Rafiul Ahmed, Mohammad Rezwanul Huq, Mohammad Munem Shahriar
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh

***Abstract*—Human Activity Recognition has been a dynamic research area in recent years. Various methods of collecting data and analyzing them to detect activity have been well investigated. Some machine learning algorithms have shown excellent performance in activity recognition, based on which many applications and systems are being developed. Unlike this, the prediction of the next activity is an emerging field of study. This work proposes a conceptual model that uses machine learning algorithms to detect activity from sensor data and predict the next activity from the previously seen activity sequence. We created our activity recognition dataset and used six machine learning algorithms to evaluate the recognition task. We have proposed a method for the next activity prediction from the sequence of activities by converting a sequence prediction problem into a supervised learning problem using the windowing technique. Three classification algorithms were used to evaluate the next activity prediction task. Gradient Boosting performs best for activity recognition, yielding 87.8% accuracy for the next activity prediction over a 16-day timeframe. We have also measured the performance of an LSTM sequence prediction model for predicting the next activity, where the optimum accuracy is 70.90%.**

***Keywords*—Machine learning algorithms; activity recognition; gradient boosting; next activity prediction; LSTM sequence prediction model**

I. Introduction

Rapid advancement in machine learning addresses a significant area of research, recognition, and human activity prediction. Predicting the next activity ahead of time can have a substantial impact on shaping and designing future technologies. A system needs to know the daily activities of a human to predict the next activity, which requires activity recognition. Recognition of an activity depends on capturing the movements and gestures made by different body parts. It is pretty challenging to correctly detect daily activities from a whole bunch of body movements. Again, the same activity can be performed in different ways. So, activity recognition has gained increasing interest in research in the past years. Activity prediction is the next step into the advancement of technology. From the recognized activities, a system would be able to predict what is going to be the next activity. It would be a massive leap towards reshaping the future, and the promise of such systems and technologies motivated this work to contribute to this field.

Activities are parts of Human behavior. An Activity consists of Actions. Three important terminologies are to be considered: Action, Activity, and Behavior [1]. Actions are a more straightforward form of conscious body movements like moving hands up and down, which form a specific action such as eating, running, walking, etc. All the actions a human performs in daily life contribute to creating his behavior, a complex structure of activities with a hidden pattern. Action data is needed to be captured by using different sensors for recognition of Activity. It can also be done from a video feed. Thus, we have two approaches for activity recognition- vision-based and sensor-based [1]. The vision-based activity recognition approach often turns into privacy concerns, so sensors get more attention in research and thus our area of interest.

In recent years, the consumer electronics industry has made a considerable investment in wearable technology. Companies produce many different wearable devices: fitness trackers, smartwatches, connected headsets, smart glasses, wrist bands, etc. Despite wearable devices not being new, the development of mobile technologies and the quantified-self movement related to fitness and sports activities have led to their explosion [2], [3]. Among the wide variety of wearable devices, wrist-wearables such as smartwatches and wrist bands seem to have become mainstream. Estimations indicated that by 2019 [4], wrist wearable devices would reach 1 million sold units, while all the other wearable devices together will achieve just 7.3 million units. In addition to other features associated with their reduced size and comfortable use, wrist-wearable devices include many sensors providing continuous data about vital signs (e.g., heart rate, skin temperature, acceleration) and environmental variables (e.g., movements). Such advancements in wearable technology create opportunities to study further by analyzing the data and developing new applications, technologies, and solutions.

Different context-aware systems such as personalized assistants, home automation, health monitoring, security management systems, etc., can benefit from activity data. The anticipation of the following activity will empower such systems to interact and perform more efficiently with users to improve context-aware experiences.

Predicting the next activity of a user requires previous activity data. So, a system that recognizes activity and stores activity information (e.g., activity name, timestamp, etc.) can be used for activity prediction. For activity recognition, it requires data for training a machine learning model. Our work includes data collection using a wearable device to build a recognition model to recognize the activity. Then the predictor model predicts the next activity from the sequence of recognized activities. We will focus on the following facts:

1) Collecting sensor data for building a machine learning model.

2) Deploying a machine learning model that best recognizes activity from collected data.

3) Storing activity information in a log file for creating an activity sequence.

4) Build another machine learning model to learn patterns in the activity sequence for predicting the next activity.

We intend to record sensor data generated from body movements and use the data to train Machine Learning Models to recognize activities. Further, the detected activities will be used for predicting the next activity for a specific user. There are different sensors available to collect required data, and these sensors can be positioned in various body parts. Sensor position is an essential factor in HAR (Human Activity Recognition) [5]. They can also be attached to home entities (e.g., bedroom doors, kitchen doors, refrigerators, washing machines, etc.) in a smart home setting. Some sensors are integrated into smartphones and smartwatches/fitness trackers. We propose to record data from a wrist wearable device because wrists are engaged in most activities in daily life, and the position is ideal for collecting data for activity recognition purposes [5]. Choosing the suitable machine learning model for this task requires effort [6]. In our proposed work, the prediction of the next activity depends on the recognized activity sequence. Though activity recognition is a widespread research interest, activity prediction is still moderately new and challenging. Several approaches have been adopted to predict activities which include Hidden Markov Models (HMM) [7], Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) [5], etc. In this work, we intend to collect data from sensors positioned at the wrist, recognize activity and predict the next possible activity of a specific user in the nearest future by implementing machine learning models. This paper contributes to the following sectors:

1) We derive three new features from collected data and build an Activity Recognition Model that performs moderately well based on a reduced dataset.

2) We propose a novel approach to predict activity by converting the sequence prediction problem into a supervised learning problem.

3) We also explore an LSTM sequence prediction approach for the next activity prediction.

The study schemes to propose a model to recognize human activity from the data collected by the sensors of a wrist-wearable device and then predict the next possible activity from the sequence of previous activities. Section 2 contains a brief discussion about Activity Recognition and Activity Prediction and prior works related to these fields. The Architecture and System Workflow of our proposed model is described in Section 3. The evaluation of our work and results are covered in Section 4. Section 5 includes an insightful discussion of our findings.

II. Literature Review

Activity recognition is the process of recognizing an activity performed by a human. It is a machine learning approach to detect activity by analyzing data given as input to a machine learning model built on a machine learning algorithm. It is a way to teach a machine to recognize an activity.

It is a fundamental and challenging problem to track and understand agents' behavior through videos taken by various cameras—the primary technique employed in computer vision. Vision-based activity recognition has found many applications such as human-computer interaction, user interface design, robot learning, and surveillance. In vision-based activity recognition, a great deal of work has been done. Researchers have attempted many methods such as optical flow, Kalman filtering, Hidden Markov models, etc., under different modalities such as single-camera, stereo, and infrared. In addition, researchers have considered multiple aspects of this topic, including single pedestrian tracking, group tracking, and detecting dropped objects. Recently some researchers have used RGBD cameras like Microsoft Kinect to detect human activities. Depth cameras add an extra dimension, i.e., the depth which a regular 2d camera fails to provide. Sensory information from these depth cameras has been used to generate real-time skeleton models of humans with different body positions. This skeleton information provides meaningful information that researchers have used to model human activities, which are trained and later used to recognize unknown activities.

Despite the remarkable progress of vision-based activity recognition, its usage for most actual visual surveillance applications remains a distant aspiration. Conversely, the human brain seems to have perfected the ability to recognize human actions. This capability relies not only on acquired knowledge but also on the aptitude of extracting information relevant to a given context and logical reasoning. Based on this observation, it is proposed to enhance vision-based activity recognition systems by integrating commonsense reasoning and contextual and commonsense knowledge.

Sensor-based activity recognition integrates the emerging area of sensor networks with novel data mining and machine learning techniques to model a wide range of human activities. Mobile devices (e.g., smartphones) provide sufficient sensor data and calculation power to enable physical activity recognition to estimate energy consumption during everyday life. Sensor-based activity recognition researchers believe that these computers will be better suited to act on our behalf by empowering ubiquitous computers and sensors to monitor agents' behavior (under consent). Sensor-based activity recognition is a challenging task due to the inherently noisy nature of the input. Thus, statistical modeling has been the main thrust in this direction in layers, where the recognition at several intermediate levels is conducted and connected. At the lowest level where the sensor data are collected, statistical learning concerns how to find the precise locations of agents from the received signal data. At an intermediate level, a statistical inference may be concerned about recognizing individuals' activities from the inferred location sequences and environmental conditions at the lower levels. Furthermore, at the highest level, a significant concern is to find out the overall goal or sub-goals of an agent from the activity

sequences through a mixture of logical and statistical reasoning.

Activity Recognition can be done in various ways. Some AR works include only accelerometer data. Fernando G.D Silva [8] designed a recognition system for simple human body movements using a tri-axial accelerometer sensor integrated with a sports watch. Min-Cheol Kwon and Sunwoong Choi built a system for recognizing activity using accelerometer and location data generated from a wrist-worn smartwatch using an Artificial Neural Network. Their approach is location-specific; a user can only perform certain actions in a specific location. An activity recognition model based on a wavelet using one or more accelerometers was proposed by Mannini and Sabitini [9]. Casale et al. [10] used a wearable device for collecting accelerometer data for human activity recognition. Some works include multiple accelerometers. Chung, Lim, Noh, Kim, and Jeong [11] built a testbed to collect motion data using a tri-axial accelerometer, gyroscope, and magnetometer by attaching eight Inertial Measurement Units (IMU) devices on different parts of the human body. They trained that dataset using the Long Short-Term Memory (LSTM) network to recognize a few Activities of Daily Living (ADLs). Foerster and Fahrenberg [12] collected data using five accelerometers and proposed a hierarchical model to classify different body postures and movements. Beddiar, Nini, Sabokrou and Hadid [13] surveyed numerous vision-based human activity recognition research papers to describe the method of HAR. They featured three essential components of this approach, which are video frame segmentation for activity recognition, action representation of the body postures and motions and ML algorithms to recognise activities by learning process. Bao and Intille [14] used five biaxial accelerometers worn on other body parts to monitor 20 types of activities using C4.5 and Naive Bayes classifiers. Wallace [15] Ugulino proposed another ML-based recognition classifier to detect five different activities (sitting, standing, sitting down, standing up, and walking) using body-worn accelerometer data collected from 4 participants. Krishnan et al. [16] collected data from ten participants using three accelerometers to detect lower body motions. Samad Zabihi [17] used transformation of the accelerometer data (x, y, z) to a spherical coordinate system (r, Ø, θ) for activity recognition and extracted features from transformed data. Zhen-Yu [18] used tri-axial accelerometer data to build an autoregressive model to detect human activity. Different activities (run, still, jump, and walk) were classified using AR coefficients feature extraction. Huawei Wang [19] used Principal component analysis to reduce the dimensionality and selected 3 out of the 12 features of a dataset. Magnetic-induction based communication system is used for sensing and transmitting data generated by every movement of the body part to recognize a physical activity in [20] Acceleration data of a smartphone is investigated in [21]. Twenty-nine users participated in data collection, and each of them carried an Android phone in their pocket. They were asked to perform six activities: Walking, Standing, Sitting, Jogging, Stairs-Up and Stairs-Down. They used Logistic Regression, J48, and Multilayer Perceptron for the model evaluation. The accuracy was above 90%. They found it a little hard to differentiate between Stairs-Up versus Stairs-Down. For activity prediction purposes, the most popular approaches are to use Recurrent Neural Networks (LSTM) and Hidden Markov Model (HMM) [1]. This paper [22] shows the importance of prediction in intelligent environments. Most of the Prediction tasks are carried out as a sequence prediction.

Activity Prediction is the process of predicting an activity ahead of time that will be performed in the nearest future. It is a way to teach a machine to predict an activity by using machine learning models. Most of the prediction tasks carried out in the past are either in a Smart Home scenario to identify the next sensor that would generate the next event or from a video to infer what will happen next. Personalized activity prediction is still a new concept and is merely investigated as a research topic. An activity can be predicted for a specific user by learning from the pattern of activities performed previously by that user. So, the activity prediction problem can be formulated as a sequence prediction problem. Sequence prediction is a problem that includes using historical sequence information to predict the next value or values in the sequence. Various methods are available for sequence prediction, but Recurrent Neural Networks, especially LSTMs, have been the best in use.

Predicting future activity empowers different applications like personal assistants and context-aware systems to interact more efficiently with the user. The problem of next activity prediction is often addressed as sequence prediction, which can be adapted to predict the label of the activity that will occur next in the sequence. Du, Lim, and Tan [23] performed activity recognition on some ADLs to generate a series of activities and then implemented LSTM and Naive Bayes to find the accuracy of predicting the next activity. The active LeZi algorithm is implemented in this [24] work to identify the sensor in a home that would generate the next event. There have been works on predicting the next location [25] and user's location-based mobility [26]. Location-based human behavior is focused in these papers and is subject primarily to using the Markov models. Markov models lack the flexibility to explore past activities instead of making a prediction based on only the most recent previous state. This feature restricts Markovian models from getting high-level insight into the data. Another popular method is Sequence mining that can be used to address such problems [27]. A dataset consisting the name of activities was generated from a collection of human actions using mapping and word embedding using LSTM algorithms to predict the future activity was implemented in [28]. Some activity prediction works are also found to be vision-based [7], [29], [2]. Alfaifi and Artoli [30] evaluated recent improvements in activity prediction and proposed a 3D-convolutional neural network model that extracted features and classified them to predict the action by LSTM.

III. Architecture of Proposed Solution

In this paper, we intend to propose a conceptual model for activity recognition and next activity prediction. The model complements the existing wearable technology architecture [31]. As illustrated in Fig. 1, the current architecture has three main components: a wrist-worn device (i.e., fitness tracker/smartwatch), an intermediate medium (i.e., Smartphone, computer), and a server/cloud service. It follows

a Proprietary system. Wearable vendors use this system for data collection and analytics and send the analytical result to the user and the authorized third parties [31]. A wrist-worn device is used to collect data using the sensors integrated into it. Theoretically, the device should be capable of sending data directly to the server for permanent storage, but it is not in use [31].

It sends data to the server/cloud service via the intermediate platforms. The server performs analytical operations to generate insight from data and sends it back to the middle layer as an interface for the wearable device. Our work will enhance the analytical capability by adding the feature to recognize an activity with a machine learning model. The model will take a chunk of sensor data from the device as input and generate a label for that movement (i.e., activity). After successfully recognizing activities and storing them as an activity sequence for a certain period, another machine learning model will take the sequence of activities as input and predict the next possible activity.

The first task of our proposed work requires data for building an activity recognition model. A fitness tracker or smartwatch is the best solution for collecting such data, integrated with different sensors. Still, we have used a smartphone for data collection and storage simplicity, tying it on the participants' hands like a tracker/smartwatch. This approach has collected the desired data but with more straightforward storage options as we could save it directly as CSV files. We have used an android app named 'AndroSensor'[1] available on the Google Play Store to collect data. We have recorded accelerometer, gyroscope, and sound level data and also kept timestamps. The data recording rate has been set to: 0.125 s/per data (instance).

Table I shows the value of the day-of-the-week and corresponding value. One is a weekday, and another is the weekend. We have considered Sunday, Monday, Tuesday, Wednesday, Thursday as weekdays while Friday and Saturday as weekends according to local holidays. Table II illustrates time-of-the-day data. Furthermore, the datasets have used in this study can be categorized into two types: recognition dataset and prediction dataset. We have selected ten activities (Table III). Here, we have held a common assumption: A user will not perform multiple activities simultaneously. The recognition dataset consists of 330993 rows and ten columns. It has contained acceleration values of x, y, z-axis, orientation values of x, y, z-axis, sound level (in DB), day-of-the-week (Weekday/Weekend), time-of-the-day (Morning, Noon, Evening, Night). The last two features have been generated from a timestamp—the acceleration data of x, y, and z-axis have depicted hand movement through the corresponding axis. We have also considered the orientation data of the x, y, z-axis because sometimes there were activities that did not generate insightful acceleration data. Some activities have been performed in specific places, e.g., we have taken transport in noisy places rather than a closed room. So, the sound level has played an essential role in recognizing activities accurately.

In the data pre-processing task, we have converted timestamp data into day-of-the-week (Table I) and time-of-the-day (Table II). Sound level, day-of-the-week, and time-of-the-day are three novel features we have introduced in our research. A human annotator has labeled the collected data according to activity labels (Table III). There have been three participants; two of them were male and one female. Each of them has collected data for three consecutive days consisting of two weekdays and one weekend. This dataset has been used to detect activity.

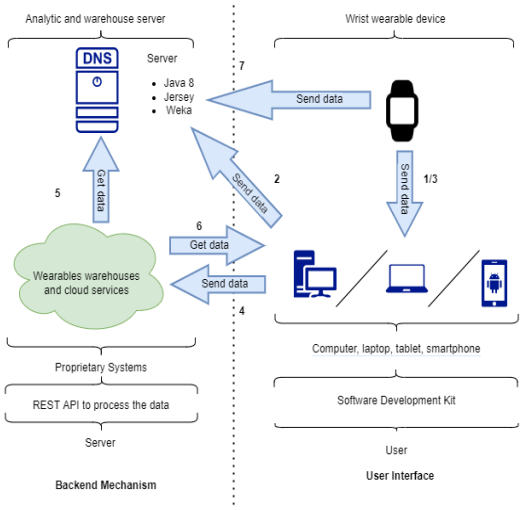


Fig. 1. Wearable Technology Architecture.

TABLE I. VALUE OF DAY-OF-THE-WEEK

Day-of-the-week	**Value**
Weekday (Sunday, Monday, Tuesday, Wednesday, Thursday)	1
Weekend (Friday, Saturday)	2

TABLE II. VALUE OF TIME-OF-THE-DAY

Time-of-the-day	Value
Morning (5:00 AM- 11:59 AM)	1
Noon (12:00 PM- 16:59 PM)	2
Afternoon (17:00 PM- 19:59 PM)	3
Night (20:00 PM- 4:59 AM)	4

TABLE III. ACTIVITY LABELS

Activity	Label
Brushing teeth	1
Drinking tea or coffee	2
Eating	3
Walking	4
Taking Transport	5
Working on PC	6
Using Mobile Phone	7
Reading	8
Cooking	9
Cleaning Utensils	10

[*1]https://play.google.com/store/apps/details?id=com.fivasim.androsensor&hl=en&gl=US

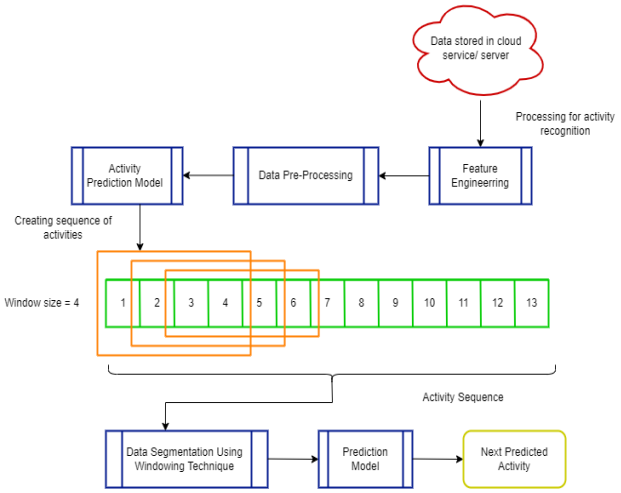


Fig. 2. System Workflow.

Fig. 2 depicts the workflow of our approach in this study. Our proposed model is initiated after data is collected by a wearable device and sent to the server. Feature engineering and data pre-processing are performed on the stored data for efficient modeling.

In our conceptual system architecture, data is then sent to the recognition model. In this study, we have investigated six machine learning algorithms: Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB) to recognize activities. When the model has started identifying activities, the labels were sent to a pipeline to create a sequence of activities. We have implemented a windowing technique to extract data from the pipeline to prepare the dataset for the next activity prediction.

Activity prediction is a user-centric process. When the activity recognition algorithm is implemented in the wearable technology architecture, it starts detecting the user's activities. So, we propose that the sequence of activities will be stored in a log file and used as the dataset for prediction purposes. We have collected data for some consecutive days to make a sequence of activities for demonstrating the activity prediction problem.

The prediction dataset used in this paper has been prepared by recording activities for eight consecutive days and sixteen days to evaluate the model performance regarding time. As prediction of the next activity depends on a specific user's behavior, data has been collected from one participant. The prediction model has taken previous activities and the current activity into consideration for predicting the next activity. Table IV shows a glimpse of the sequence of activities performed by the participant. Hence, it has a various number of attributes depending on the window size we choose. Depending on different window sizes concerning 8 days and 16 days, the number of rows has differed in the prediction dataset.

TABLE IV. SEQUENCE ACTIVITIES

1	6	4	5	10	2	6	3	1	…..

TABLE V. TRANSFORMED DATA FOR NEXT ACTIVITY PREDICTION (WINDOW SIZE = 3)

	Prev_act2 (Feature 1)	**Prev_act1 (Feature 2)**	**Curr_act (Feature 3)**	**Next_act (Target)**
1	1	6	4	5
2	6	4	5	10
3	4	5	10	2
4	5	10	2	6
5	10	2	6	9

We have introduced a novel approach for the next activity prediction. We have transformed the sequence of activities into a feature and target-shaped data frame by implementing the windowing technique. It creates the opportunity to use the transformed sequential dataset in a supervised learning problem predicting human activities. Table V shows a sample of the dataset for the next activity prediction, where the window size is 3, which features were used as input and the target as output for the prediction model.

We propose another approach to predict the next activity, as shown in Fig. 3. An LSTM sequence prediction approach is adopted to predict the next value in the activity sequence.

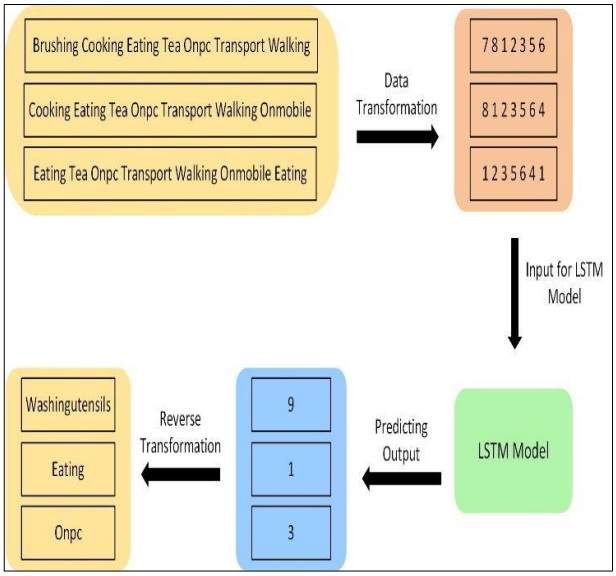


Fig. 3. LSTM Sequence Prediction Workflow.

IV. RESULT AND DISCUSSION

We have used python with scikit-learn ML packages for the Machine Learning algorithms implemented in our study [32]. Keras and Tensorflow are used to implement LSTM. Some modifications have been made in the default hyperparameters of the ML algorithms to achieve better accuracy. The codes run in the Anaconda Jupyter Notebook (Version3).

Fig. 4 illustrates the accuracy comparison of the six algorithms for activity recognition. For the Artificial Neural Network, the number of hidden layers has played an important role. The accuracy for the 3-layer network was 98.99% and for the 7-layer network was 99.01%. However, we have

considered the outcome of five hidden layer networks, as it performed the best for this dataset with an accuracy of 99.1%. On the other hand, the K-Nearest Neighbor yielded an accuracy of 99.2%. Where the accuracy of the Random Forest was 99.3%, for the Support Vector Machine, it was 98.7%. Though the Naïve Bayes performed poorly, having an accuracy of 61.9%, among all the six ML algorithms, the Gradient Boosting showed the highest accuracy of 99.7%.

As mentioned earlier, in this study, we have introduced three novel features: sound level, day-of-the-week, and time-of-the-day. These trio features help to detect the activity more accurately. As a comparison, we have run all six algorithms on the same dataset without these trio features.

Fig. 5 clearly describes that all six algorithms' accuracy had reduced compared to Fig. 4. Yet, Gradient Boosting had the best accuracy of 98.8%, and Naïve Bayes yielded the lowest accuracy of 54.4%. In comparison, the accuracy differed by around 2% for all the algorithms except for Naïve Bayes. Naïve Bayes accuracy difference was almost 7%.

After studying the accuracy score of the six implemented ML algorithms in the previous section for activity recognition, we can see that GB gave the best result [99.7%], where the nearest result was shown by KNN and the bagging algorithm Random Forest with an accuracy rate of 99.2% and 99.3% respectively. As we know, both bagging and boosting are ensemble methods and perform better compared to other ML algorithms. So, as expected, both of them have the highest accuracy among all six algorithms.

Fig. 6 explains the confusion matrix of our best model for detecting human activity, Gradient Boosting. Almost all the instances of the actual class have been predicted accurately by the model. Specifically, the activity 'Walking' is labeled as activity no. 4. Of all the instances of 'Walking,' data has been predicted correctly as 'Walking' with an accuracy of 99.9%. Also, the instances of other activities have been predicted correctly with varying accuracy from 98.8% to 99.8%.

For predicting the next activity, we have used two timeframes. The first one was a prediction on an 8 days activity sequence and the second one was on 16 days. We have used 3 ML algorithms to check the accuracy, e.g., ANN, KNN, GB. We have experimented with different window sizes to observe the effect on accuracy.

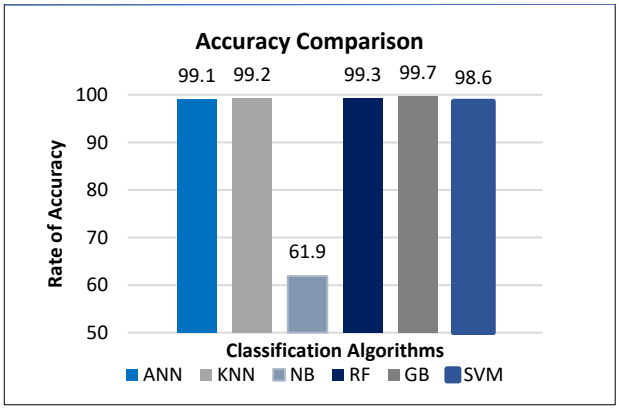


Fig. 4. Accuracy Comparison for Activity Recognition.

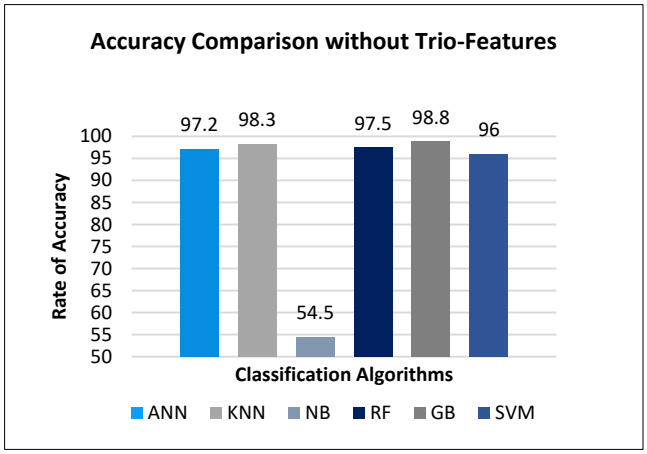


Fig. 5. Accuracy Comparison without Trio-Features.

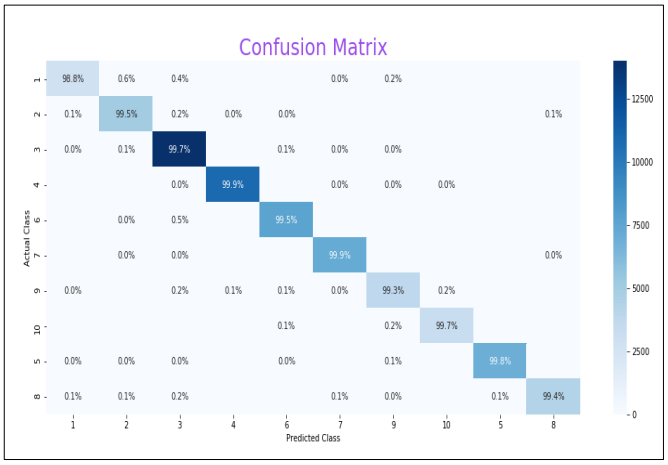


Fig. 6. Confusion Matrix of the Best Model (Gradient Boosting).

Here we have kept both training and test accuracy in determining if the models have overfitting or underfitting issues. For an 8-days timeframe, window size (W) was taken from 2 to 6 with an interval of 1 to explore the effect of window size on accuracy comparison. Fig. 7, Fig. 8, Fig. 9 show the accuracy comparison of training and test for 8 days on ANN, KNN, GB for predicting the next activity, respectively.

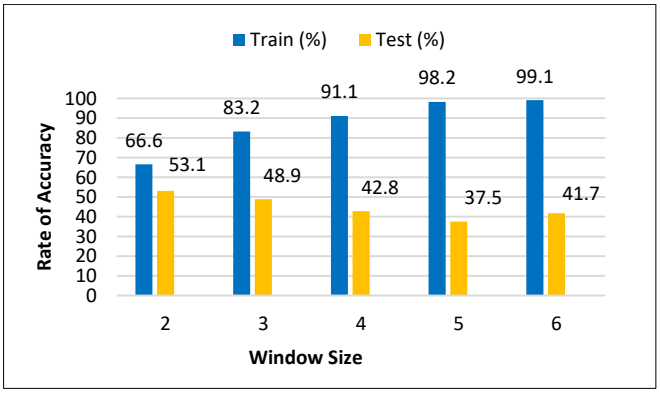


Fig. 7. Accuracy Comparison (8 Days) with varied Window Size for ANN.

Fig. 7 illustrates that an increase in window size increases training accuracy but decreases test accuracy. At W = 2, ANN yielded the highest test accuracy of 53.1% but kept lowering with the window size increase.

Similarly, in Fig. 8, training accuracy increases with the increase of window size, but there is a sudden drop at W = 6. However, the highest test accuracy for KNN was 46.9% at W = 2, 3. The test accuracy gradually kept decreasing with the change of window size. KNN showed similar behavior as ANN to predict the next activity from the activity sequence of 8 days.

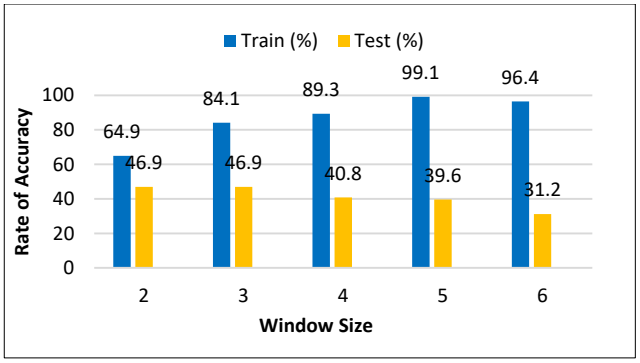


Fig. 8. Accuracy Comparison (8 Days) with varied Window Size for KNN.

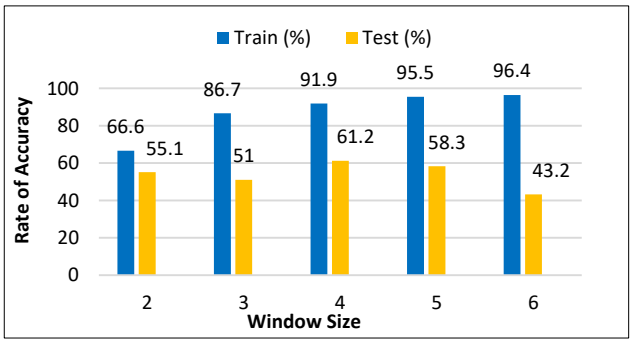


Fig. 9. Accuracy Comparison (8 Days) with varied Window Size for GB.

Fig. 9 displays that with the change of window size, training accuracy kept increasing. In the case of testing accuracy, there was some fluctuation of performance for varied window sizes. Notably, for W = 4, GB showed the best result of 61.2% accuracy.

We can sense a trend in training and test accuracy with varied window sizes. In most cases, for all the algorithms, an increase in window size increases training accuracy but results in a gradual decrease in test accuracy. So, there is a hint of slight overfitting of data here. Now we will be analyzing the same for a more extended timeframe of 16 days of activity sequence.

For a 16-days timeframe, the window size varied from 2 to 8. For predicting the next activity, the accuracy comparison of training and test for 16 days on ANN, KNN, and GB is illustrated in Fig. 10, Fig. 11, and Fig. 12, accordingly.

Analyzing Fig. 10, it can be said that, while training accuracy increases with the increase of window size, testing accuracy varies significantly. The highest test accuracy was yielded at W = 6, which was 81.6%.

Similarly, in Fig. 11, for KNN, training accuracy increased with the change of window size. KNN also varied for test accuracy, but like the previous one, it performed best at W = 6, 82.7%.

The accuracy comparison for GB on 16 days dataset is graphed in Fig. 12. With the increase of window size, training accuracy kept increasing. GB also yielded the highest accuracy of 87.8% at W = 6.

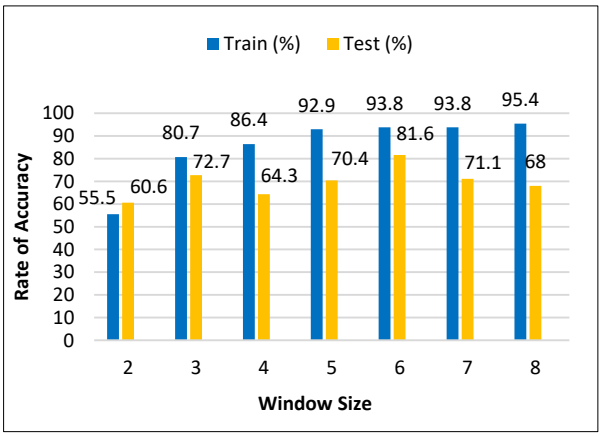


Fig. 10. Accuracy Comparison (16 Days) with varied Window Size for ANN.

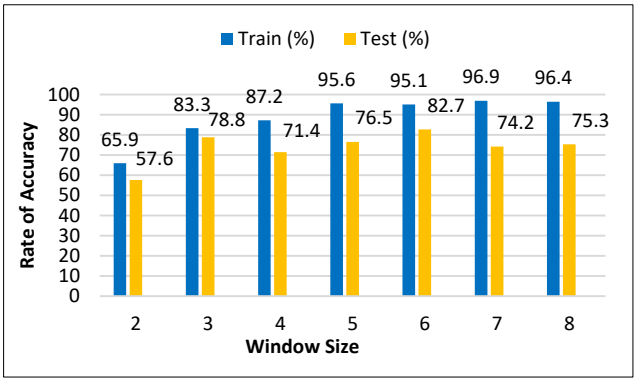


Fig. 11. Accuracy Comparison (16 Days) with varied Window Size for KNN.

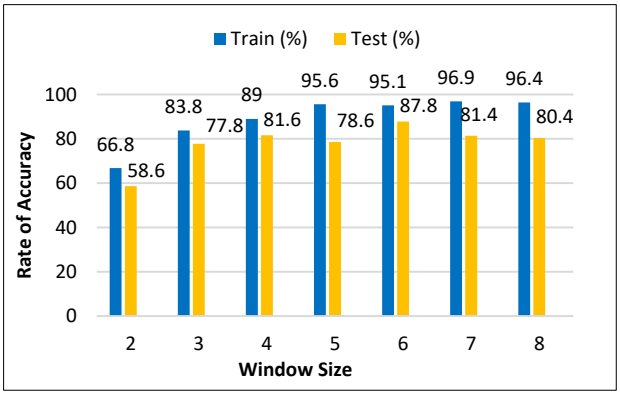


Fig. 12. Accuracy Comparison (16 Days) with varied Window Size for GB.

After vivid exploration of the above cases, a longer timeframe of activity sequence yields better performance for

predicting the next activity. Also, we can say that the algorithms are prone to overfitting with smaller timeframe datasets.

Observing Table VI, it can be said that increasing the window size gives a good test accuracy compared to the training accuracy for a specific limit. The training accuracy increases with the increase of window size. Just a small exception in the case of KNN when the window size was changed from 5 to 6, it decreased instead of increasing. But in the case of testing, while changing the window size from 4 to 5, the test accuracy started to decline and continued for all the algorithms. On the other hand, Table VII indicates the accuracy comparison between the training and testing dataset for 16 days. Unlike Table VI, after increasing the window size with an interval of 1, GB's testing accuracy was quite good, though the accuracy decreased for ANN and KNN on some points.

For training accuracy, we can say that the accuracy increased gradually with the change of the window size. Though after size 5, it decreased for some, at size 7, it rose again. It can be determined that for ANN and KNN the test accuracy varied while increasing the window size after analyzing Table VII. But for GB, the accuracy increased up to window size 6, and after that, it started to decrease even though the window size was increasing. This was the purpose of considering a higher range of window size for16 days compared to 8 days.

TABLE VI. TRAINING AND TESTING ACCURACY (8 DAYS) WITH VARIED WINDOW SIZE

Window Size	ANN		KNN		GB	
	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
2	66.6	53.1	64.9	46.9	66.6	55.1
3	83.2	48.9	84.1	46.9	86.7	51.0
4	91.1	42.8	89.3	40.8	91.9	61.2
5	98.2	37.5	99.1	39.6	95.5	58.3
6	99.1	41.7	96.4	31.2	96.4	43.2

TABLE VII. TRAINING AND TESTING ACCURACY (16 DAYS) WITH VARIED WINDOW SIZE

Window Size	ANN		KNN		GB	
	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
2	55.5	60.6	65.9	57.6	66.8	58.6
3	80.7	72.7	83.3	78.8	83.8	77.8
4	86.4	64.3	87.2	71.4	89.0	81.6
5	92.9	70.4	95.6	76.5	95.6	78.6
6	93.8	81.6	95.1	82.7	95.1	87.8
7	93.8	71.1	96.9	74.2	96.9	81.4
8	95.4	68.0	96.4	75.3	96.4	80.4

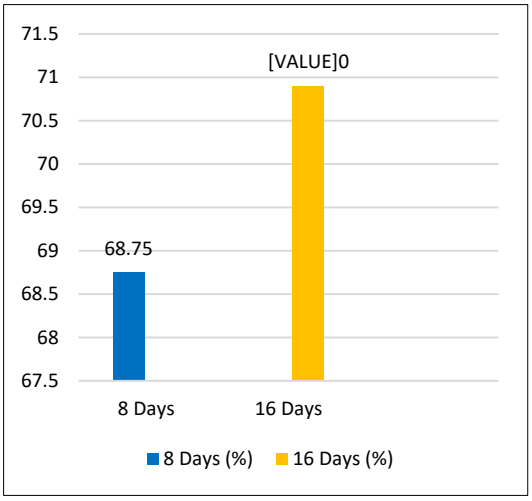


Fig. 13. Accuracy Comparison (8 Days vs. 16 Days) LSTM Activity Prediction.

Fig. 13 shows an accuracy comparison for the LSTM model used for predicting the next activity. Here we have assessed the performance of the model in a limited way. The model is assessed under timeframes of 8 and 16 days but considered a fixed window size for each timeframe, unlike the previous method that used varied window sizes. For the 8 days' timeframe, window size 6 was taken, and it was 8 for 16 days. It showed slightly better performance for a longer sequence of activities of 16 days than 8 days. For the 8 days' timeframe, the model yielded an accuracy of 68.75% and 70.90% for 16 days. We observe a similar behavior of the LSTM model performing with better accuracy for a longer timeframe, just like the previous method.

Both the supervised model and LSTM had performed well for predicting the next activity for an individual when there was a long activity sequence. Fig. 14 shows a performance comparison between LSTM and the supervised model used for predicting next activity. For 8 days timeframe and window size 6, LSTM outperforms our best-supervised model. But in the longer timeframe of 16 days and window size 8, the supervised model performs better than LSTM.

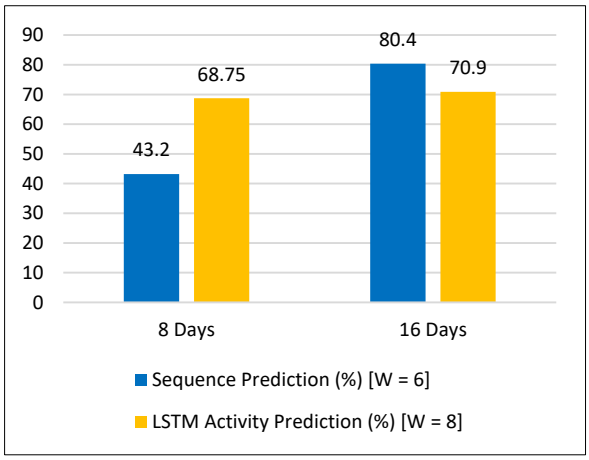


Fig. 14. Accuracy Comparison between Supervised Model and LSTM Activity Prediction (8 Days vs. 16 Days).

Hence, we can conclude that both the methods used for predicting the next activity perform better with an extended timeframe of activities.

Compared to the current work, our proposed model includes three new novel features: sound level, day-of-the-week, and time-of-the-day while collecting data. Moreover, the activities chosen to be recognized shows deviation quantitatively. Various machine learning algorithms have been applied, and their accuracy is as good as the current work. Furthermore, we have converted a sequential dataset into a supervised learning model, which was not conducted in previous work. The accuracy rate of the LSTM model for activity prediction performed better than the existing surveys.

To summarize, our proposed models show a significant outcome. The Gradient Boosting produced 99.7% accuracy out of all six algorithms for recognizing activity. In addition, a longer timeframe with a longer day count performs better than others for predicting next activity. Sequence and LSTM activity predictions provide 80.4% and 70.9% accuracy respectively for a 16-day timeframe with window size 8.

V. Conclusion

Due to resource constraints, we could not use a smartwatch or any wearable device to collect data; instead, we used a smartphone tying it on the wrist, which worked just like a wrist wearable. For data collection, we had to depend on only 3 participants. The work can be best understood and explained by deploying it in real-time. But we had to work with batch processing due to the lack of high-end machines and components.

This paper proposes a conceptual model for activity recognition and prediction, which can be extended with the existing wearable architecture. We have shown how sound level, day-of-the-week, and time-of-the-day can improve activity recognition model accuracy. We adopted a straightforward but effective approach to convert a sequence prediction problem into a supervised learning problem to predict the next activity. The accuracy was moderate, considering the dataset we have. We have also used a Long-Short-Term Memory (LSTM) model to explore the prediction process. The model we developed has shown good performance, yet the result is not always accurate.

Our models have shown promise for both activity recognition and prediction. Wearable technologies and home automation, security systems, and health monitoring systems can significantly leverage this concept. By integrating our model in a smartwatch or a wrist-worn wearable architecture, this unique technology can be readily available to common people. Predicting the next human activity can create a significant impact on existing and future technologies.

The next activity predictor model predicts the next activity based on previously detected activities. Prediction model accuracy depends mainly on the timeframe. The comparison between the 8 days dataset and 16 days dataset clearly shows the difference. If the dataset is sufficiently large and the hyperparameters are appropriately tuned, we expect a far better result. We have noticed that both prediction models yield better accuracy with a larger window size when the time frame increases from 8 days to 16 days. Though it is difficult to say which model can predict the next activity of a person more accurately, for different timeframes with different window sizes, both the supervised learning model and LSTM perform differently. We have a plan to extend our research with more participants.

Moreover, we want to determine if it is possible to predict which approach performs more precisely using even more timeframes with different window sizes. We will also investigate the fact that if there is any correlation between timeframe and window size. We intend to further mature this proposed model by implementing it in real-time with streaming data. In the future, we will try to make predictive modeling more efficient by incorporating timestamps in the model and the activity to create time series forecasting. We also intend to introduce some more valuable features and recognize more new activities. Furthermore, a recommendation system can be designed based on the predicted activity. The recommender system will use the next activity data indicated by the model to make corresponding recommendations.

References

[1] A. Almeida and G. Azkune, "Predicting Human Behaviour with Recurrent Neural Networks," in DeustoTech-Deusto Foundation, University of Deusto, Av. Universidades 24, 487 Bilbao, Spain, 2018.

[2] Y. Fu and Y. Kong, "Human Activity Recognition and Prediction," JOURNAL OF LATEX CLASS FILES, VOL. 13, NO. 9, SEPTEMBER 2018, pp.23-48.

[3] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearable devices: Has the Age of Smart watches Finally Arrived?," Communications of the ACM, 2015, pp. 45-47.

[4] M. Swan, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self," 2.0. Journal of Sensor and Actuator Networks, 2012, pp. 217-253.

[5] S. Muhammad, S. Bosch, O. D. Incel, H. Scholten and P. Havinga, "Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors," Sensors, 2016.

[6] A. Moraru, M. Pesko, M. Porcius, C. Fortuna and D. Mladenic, "Using machine learning on sensor data," Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, Cavtat, 2010, pp. 573-578.

[7] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 1036-1043.

[8] F. G. da Silva and E. Galeazzo, "Accelerometer based intelligent system for human movement recognition," 5th IEEE International Workshop on Advances in Sensors and Interfaces IWASI, Bari, 2013, pp. 20-24.

[9] A. Mannini and A. Sabatini, "Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers," Sensors, Basel, Switzerland, 2010, pp. 1154-1175.

[10] P. Casale, O. Pujol, and P. Radeva, "Human Activity Recognition from Accelerometer Data Using a Wearable Device," Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011, pp. 289-296.

[11] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning," Sensors, vol. 19, no. 7. MDPI AG, p. 1716, Apr. 10, 2019.

[12] F. Foerster and J. Fahrenberg, "Motion pattern and posture: Correctly assessed by calibrated accelerometers," Behavior research methods, instruments, & computers: a journal of the Psychonomic Society, Inc. 32. 450-7, 20.

[13] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, “Vision-based human activity recognition: a survey,” Multimed. Tools Appl., vol. 79, no. 41–42, pp. 30509–30555, 2020.

[14] L. Bao and S. S. Intille, "Activity Recognition from User-Annotated Acceleration Data," In A. Ferscha & F. Mattern (eds.), Pervasive, 2014, pp. 1-17.

[15] W. Ugulino, D. Cardador, K. Vega and E. Velloso, "Wearable computing: accelerometers data classification of body postures and movements." Advances in Artificial Intelligence-SBIA 2012. Springer Berlin Heidelberg, 2012, pp. 52-6.

[16] N. C. Krishnan, D. Colbry, C. Juillard and S. Panchanathan, "Real time human activity recognition using tri-axial accelerometers, Sensors, signals and information processing workshop," Sensors Signals and Information Processing Workshop, Sedona, AZ, 28,2018, pp.915-922.

[17] W. T. D. Souza, and R. Kavitha, 'Human Activity Recognition Using Accelerometer and Gyroscope Sensors," International Journal of Engineering and Technology, 2017, pp.1171-1179.

[18] Zhen-Yu He and Lian-Wen Jin, "Activity recognition from acceleration data using AR model representation and SVM," 28 International Conference on Machine Learning and Cybernetics, Kunming, 28,2008 pp. 2245-2250.

[19] H. Wang and J. Wu, "Classification of Human Posture and Movement Using Accelerometer Data," in International Conference on Innovations in Computing & Networking (ICICN16), CSE, RRCE, ISSN: 0975-0282.

[20] N. Golestani and M. Moghaddam, “Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks,” Nat. Commun., vol. 11, no. 1, p. 1551, 2020.

[21] J. R. Kwapisz, G. M. Weiss and S. A. Moore, "Cell phone-based biometric identification," 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington, DC, 2010, pp. 1-7.

[22] D. J. Cook and S. K. Das, "How smart are our environments? An updated look at the state of the ar" t. Pervasive Mobile Comput. 27, 3, pp.53–73.

[23] Y. Du, Y. Lim, and Y. Tan, "A Novel Human Activity Recognition and Prediction in Smart Home Based on Interaction," Sensors, vol. 19, no. 20. MDPI AG, p. 4474, Oct. 15, 2019.

[24] K. Gopalratnam and D. J. Cook, "Online Sequential Prediction via Incremental Parsing: The Active LeZi Algorithm," IEEE Intelligent Systems, vol. 22, no. 1, Jan.-Feb. 27, pp. 52-58.

[25] S. Sigg, S. Haseloff and K. David, "An Alignment Approach for Context Prediction Tasks in UbiComp Environments," IEEE Pervasive Computing, vol. 9, no. 4, October-December 2010, pp. 90-97.

[26] J. Mcinerney, S. Stein, A. Rogers, Alex and N. R. Jennings, "Breaking the habit: Measuring and predicting departures from routine in individual human mobility," Pervasive and Mobile Computing, 2013, pp. 808–822.

[27] O. Brdiczka, N. S. Makoto and J. Begole, "Temporal task footprinting: Identifying routine tasks by their temporal patterns," International Conference on Intelligent User Interfaces, Proceedings IUI, 2010, pp. 281-284.

[28] J. Septiadi, B. Warsito, and A. Wibowo, “Human activity prediction using long Short Term Memory,” E3S Web Conf., vol. 202, p. 15008, 2020.

[29] L. Draschkowitz, C. Draschkowitz and H. Hlavacs, "Using Video Analysis and Machine Learning for Predicting Shot Success in Table Tennis," EAI Endorsed Transactions on Creative Technologies, 2015.

[30] R. Alfaifi and A. M. Artoli, “Human action prediction with 3D-CNN,” SN Computer Science, vol. 1, no. 5, 2020.

[31] A. P. Francisco, M. C. Rodriguez, J. M. Santos, "Collection and Processing of Data from Wrist Wearable Devices in Heterogeneous and Multiple-User Scenarios," Sensors, 2016, pp. 1538.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res. 2011, 12, 2825–2830.

Study of Haar-AdaBoost (VJ) and HOG-AdaBoost (PoseInv) Detectors for People Detection

Nagi OULD TALEB[1], Mohamed Larbi BEN MAATI[2]
Mohamedade Farouk NANNE[3], Aicha Mint Aboubekrine[4], Adil CHERGUI[5]
Department of Computer Sciences, CSSE Laboratory, University of Abdelmalek Essaadi, Tetouan, Morocco[1, 2]
Department of Computer Sciences, University of Nouakchot Al Asriya, Nouakchot, Mauritania[3]
Department of Computer Sciences, LIROSA Laboratory, University of Abdelmalek Essaadi, Tetouan, Morocco[4]
Department of Computer Sciences, CSSE Laboratory, ENSAM-Casablanca, Tetouan, Morocco[5]

***Abstract*—Object detection in general and pedestrians in particular in images and videos is a very popular research topic within the computer vision community; it is an issue that is currently at the heart of much research. The detection of people is a particularly difficult subject because of the great variability of appearances and the situations in which a person may find themselves (a person is not a rigid object; it is articulate and unpredictable; its shape changes during its movement). The situations in which a person may find themselves are very varied: They are alone, near a group of people or in a crowd, obscured by an object. In addition, the characteristics vary from one person to another (color of the skin, hair, clothes, etc.), the background simple, clear or complex, the lighting or weather conditions, the shadow caused by different light sources, etc. greatly complicate the problem. In this article, we will present a comparative study of the performance of the two detectors Haar-AdaBoost and HOG-AdaBoost in detecting people in the INRIA images database of persons. An evaluation of the experiments will be presented after making certain modifications to the detection parameters.**

***Keywords*—Pedestrian detection; learned-based methods; Haar-like features; HOG descriptor; AdaBoost; behavior analysis**

I. Introduction

The detecting of people in images is a very important subject in the field of computer vision. The pedestrians' detection is therefore a main concern of several researchers in the field of computer vision. These applications, ranging from surveillance, retail data mining and automatic pedestrian detection in the automotive industry, have fueled research over the past decade, leading to a growing number of approaches on the subject [1].

Many factors can influence the human figure, such as the constantly changing appearance, crowds, obscuration by objects, the type of environment and the unpredictability of pedestrians [2, 3].

In the literature, we find techniques that require segmentation or subtraction of the background and others directly detect the person without such preprocessing. These techniques use many characteristics to describe human appearance (shape, color, movement) in order to build shape models used on explicit or learning-based detection techniques.

Several systems have been developed in this context with dynamic methods such as Phantom [4] and Pfinder [5]. Other methods have been conducted [6, 7, 8, 9] for the detection of people with a measure of their activity in video sequences. Shooting with a fixed camera allows background subtraction to reduce search space. Finally, we find the system that performs fast and accurate human detection by integrating the cascade approach with histograms of gradient directions [10, 11].

Among these approaches, we find a so-called global one that has a principle of using the shape of the whole body as a source of information without taking into account local characteristics [10, 12, 13]. Viola and Jones [14, 15] [16] also proposed a detector based on Haar filters and the boosting algorithm. There are some aspects of this algorithm, based on infrared vision to detect a human in a room and provide a history of occupancy of the room.

Another so-called local approach uses local characteristics. Here we extract the characteristics from the image base and then we build the discriminating model, for example, Papageorgiou [17] that proposed a detector based on the Haar wavelet.

The latest so-called hybrid approach combines local and global characteristics to improve recognition performance [18].

The research work proposed in this article aims to contribute to the shapes (or objects) recognition modeling methods and more particularly of pedestrians by descriptors classification containing the most relevant information of an object and applying the models found to the human silhouette (people or pedestrians) detection in images or multimedia streams (video).

II. Learning Methods for Human Detection

The main approaches based on discriminate learning train different types of classifiers on a large number of samples of negative and positive images, where humans are well framed.

Each method must extract the appropriate characteristics and the main information captured from the training data is the spatial recurrence of local shape events. If the trained classifier does not detect an object (misses the object) or mistakenly detects the absent object (false detection), it is easy to make an adjustment by adding the corresponding positive or negative samples to the learning set.

However, due to the complexity of articulated human poses and variable visualization conditions, training data becomes very large (especially positive samples) therefore the generalization ability of the trained classifier may be compromised.

The Fig. 1 shows an illustration of the data formation process that is common basic in all detectors.

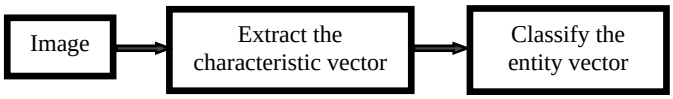


Fig. 1. Common Learning Process.

III. Study of Haar-AdaBoost and HOG-Adaboost Detectors for the People Detection

The study which we carried out in the paper [19] of 14 traditional techniques resulting from the literature and representing the state of the art allowed us to choose the two most popular methods in the detection of the objects Haar-AdaBoost (VJ, Viola and Jones) [14, 15] and HOG-AdaBoost (PoseInv, Pose-Invariant) [20] which constitutes a variant of HOG-SVM [12] (SVM design a Support Vector Machine classifier [21]) to study their feasibility for detecting people.

Note that the detectors Haar-AdaBoost and HOG-AdaBoost uses respectively the Haar-like features (or Haar wavelets) [14, 15, 17] and HOG (Histograms of Oriented Gradients) descriptor [12] to extract the characteristic vector from an image of person and they are based on the same classifier AdaBoost [21] to classify this vector. AdaBoost is one of the most powerful binary classification methods in supervised learning, its uses an iterative selection of weak classifiers based on a distribution of learning examples. Each example is weighted according to its difficulty with the current classifier. The main motivation for boosting was to form a procedure which combines the output of several weak binary classifiers to produce a powerful binary classifier [22].

In this approach we will present the experimental results carried out in the Computer Science and Systems Engineering Laboratory of the Faculty of Sciences of Tetouan for evaluing people detection in images using the two detectors Haar-AdaBoost and HOG-AdaBoost.

The performance analysis of these two detectors was carried out on the people images database from INRIA Person Dataset (http://pascal.inrialpes.fr/data/human/). This database provides 460 color bitmap (BMP) images of people at 640 × 480 resolution.

The study thus made is based on the plotting of the TP-IoU, FP-IoU, FN-IoU, IoU-Recall curves and the evaluation of the AR (Average Recall) metric. Plotting the Precision-Recall curve and evaluating the AP (Average Precision) metric cannot be performed in this study because the Haar-AdaBoost and HOG-AdaBoost detectors do not return a confidence score, but rather an indication whether an object detected belongs to the desired class or not.

Here is the meaning of the TP, FP, FN, Precision, Recall and IoU metrics:

- TP: True Positive, also called detection, is a correctly detected person (or object).
- FP: False Positive, also called false detection, occurs when the predicted box provided by the classifier does not contain any person to be detected.
- FN: False Negative, denotes a case where a person is missed.
- Recall: is the number of true positives divided by the sum of true positives and false negatives, this last sum is just the number of ground-truths boxes: $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$. This metric measures the rate of true positives detected among all positives, so it is a measure of detector performance in finding positives.
- Precision: is the number of true positives divided by the sum of true positives and false positives: $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$. This metric measures the rate that the detection is correct, so it is used to measure the accuracy of the detection.
- IoU: Intersection over Union, is the ratio of the area of the intersection between the predicted bounding box and the ground-truth bounding box on the area of their union.

The OpenCV library offers a list of classifiers in XML format already trained to respectively detect faces, eyes, profile heads, human bodies, etc. These classifiers are located in the opencv \data\haarcascades\folder.

Among the classifiers provided by OpenCV, we have chosen to study the performance of two of them which are already trained for detecting people in images, haarcascade_fullbody.xml and hogcascade_pedestrians.xml which provide two models for detecting people in the images obtained respectively by the implementations under OpenCV of the cascade classifier Haar-AdaBoost of Viola and Jones and HOG-AdaBoost of Lin and Davis (a variant of the Dalal and Triggs detector [12]).

IV. INRIA Person Dataset Image Database and Manual Images Labeling

Among the 460 images in the INRIA Person Dataset, we have manually annotated only 187 images from the first images of this dataset using the objectmarker annotation program, resulting in a total of 481 ground-truth bounding boxes. But, it was better to annotate all the images in the database, which requires more effort and time. Also note that during the annotation, we ignored some images containing a single person very close (or on a very large scale) that we did not consider interesting, the number of this last images was very little.

The Fig. 2 shows four images from the INRIA Person Dataset labeled using the objectmarker program. Each person presented in these images is manually framed using a rectangle, called a ground-truth bounding box. These boxes give precise positions of the people in the images; they are presented by the coordinates (x, y) of the upper left point of the rectangle, its width and its height.

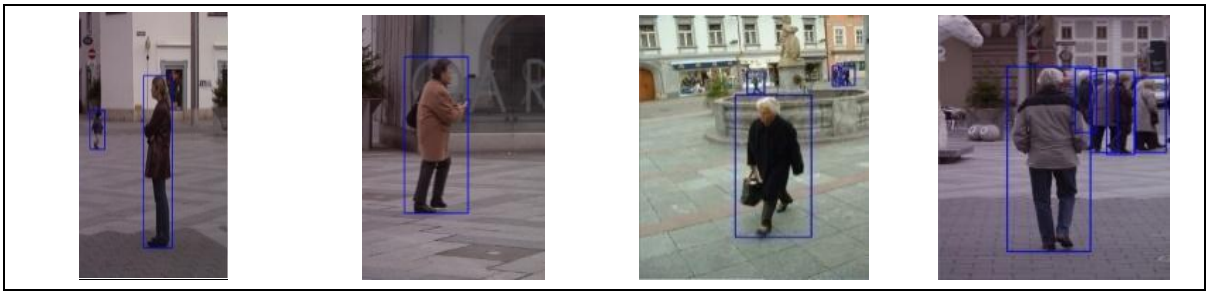


Fig. 2. People Labeled Manually in the INRIA Person Dataset Images using the Objectmarker Program and they are Framed with a Blue Ground Truth Bounding Box. The Four Images show Respectively the Labeling of One Person, Two Persons and Four Persons at different Scales and then One Person with a Crowd of People.

V. Comparison of the Two Detectors Haar-AdaBoost and HOG-AdaBoost using Examples of People Detection in the INRIA Person Dataset

In this paragraph, we will present a preliminary comparative study of the two detectors Haar-AdaBoost and HOG-AdaBoost. This comparison will be based on the application of these two detectors on the first 187 images that we were precedently annotated in the INRIA Person Dataset images. We will discuss the strengths of each of these two detectors as well as their failing.

Table I shows some examples of people detection obtained respectively by the application of the two detectors Haar-AdaBoost and HOG-AdaBoost on the 187 images tagged in the INRIA Person Dataset. The first column corresponds to the application of the Haar-AdaBoost detector, while the second column corresponds to the application of the HOG-AdaBoost detector.

To compare the two detectors and discuss their performance, we have chosen to show some of the most significant detection results obtained on a sample of well-selected INRIA Person Dataset images. In the images below, the blue frame corresponds to the ground truth bounding box produced by manual labeling using the objectmarker program. The boxes in green correspond to the boxes predicted respectively by the two detectors Haar-AdaBoost and HOG-AdaBoost.

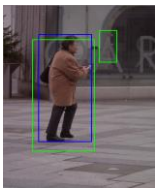
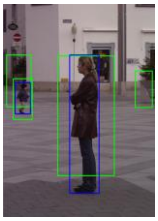
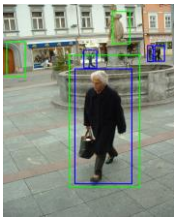
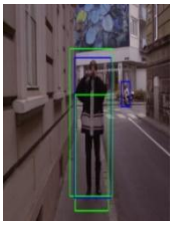
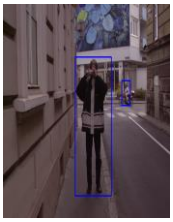
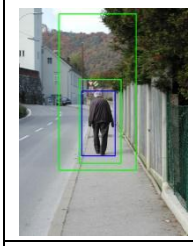
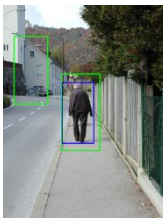
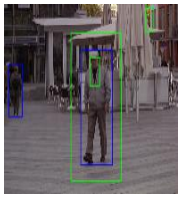
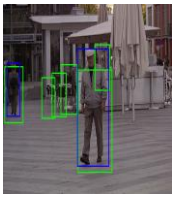
Experimentation with Haar-AdaBoost and HOG-AdaBoost detectors on images from the INRIA Person Dataset allowed the following conclusions to be drawn:

- The two detectors generally fail to detect people on a small scale (or very far away).
- Likewise, on a very large scale or when people are very close and fill almost the entire image, the two detectors generally do not succeed in detecting them or sometimes generate, in particular by the HOG-AdaBoost detector, small predicted framing boxes whose IoU with their associated ground truth boxes is of small value.
- Sometimes the shape of the clothes (especially if a person is wearing a coat or a dust jacket) can also cause a person on a medium scale to not be detected by the Haar-AdaBoost and HOG-AdaBoost detectors.
- The HOG-AdaBoost detector is overall better than the Haar-AdaBoost detector for detecting people on a medium scale (i.e., people who are slightly close) and large scale (i.e., people who are close), but unfortunately it generates a lot of false detections (or false positives) than the Haar-AdaBoost detector.
- On a medium and large scale, the Haar-AdaBoostr detector sometimes sends two predicted bounding boxes corresponding to the detection of the same person. This does not happen with the HOG-AdaBoost detector which does a good job of eliminating duplicates and typically returns a single predicted box for each person detected.
- The choice of the IoU threshold is very important so as not to miss some correct detections. We have observed that with the IoU threshold set at 0.5, the Haar-AdaBoost detector sometimes returns detections which are correct, but which have an IoU lower than 0.5, this leads to an erroneous interpretation of the detections result obtained. This situation rarely happens with the HOG-AdaBoost detector where the IoU of detection is often greater than 0.5.

We have also found that the minimum value of the IoU threshold that must be set depends on how to label people, in fact, if the ground-truth bounding boxes are manually drawn too tight to the person that they frame, the detector can sometimes generate an IoU with the predicted bounding box less than 0.5.

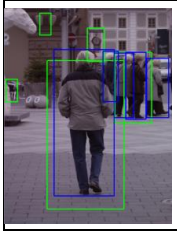
Based on the analysis of the detection results obtained by the Haar-AdaBoost and HOG-AdaBoost detectors, we found that for the minimum IoU threshold value set at 0.4, almost all detections that give rise to true positive are correctly determined. Therefore, to study the performance of the two detectors, it will be preferable to vary the minimum threshold of the IoU between 0.4 and 1 instead of 0.5 and 1, this is what we used in the plotting of the True Positive as a function of the IoU (TP-IoU), False Positive as a function of the IoU (FP-IoU), False Negative as a function of the IoU (FN-IoU) and Recall as a function of the IoU (Recall-IoU).

TABLE I. THIS TABLE SHOWS SOME PEOPLE DETECTION RESULTS OBTAINED RESPECTIVELY BY APPLYING THE TWO HAAR-ADABOOST AND HOG-ADABOOST DETECTORS TO THE IMAGES IN THE INRIA PERSON DATASET

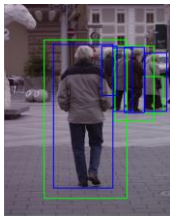
Haar-AdaBoost detector	**HOG-AdaBoost detector**
 The person on a large scale (or at a close distance) is not detected. This is therefore a false negative.	 Here, the same large-scale person is indeed detected with an IoU equal to 0.755259, this is a true positive. The image also shows a false detection that matches the green frame on the side of the display case, so this is a false positive.
 The medium-scale child is well detected (IoU equals 0.580247). The large-scale lady goes undetected. There is a true positive (the child) and a false negative (the lady) here.	 Here, the medium-scale child and the large-scale lady are well detected (the respective IoUs are 0.613888 and 0.482886). Despite the lady being detected, the IoU between the green and blue frame is 0.482886 which is less than 0.5. In this case, if the IoU threshold is set to 0.5, then the lady's blue and green frames will be considered as a false negative and a false positive respectively, which is incorrect. There are also two false detections (or false positives).
 The large-scale lady is well detected (IoU is 0.636414), but the small-scale people are not detected. There is therefore 1 true positive and 3 false negatives.	 The same thing here, the large-scale lady is detected (IoU equals 0.70401), but the small-scale people are not detected. There is therefore 1 true positive, 3 false negatives and 1 false positive (the statue).
 The same person is detected twice, the two predicted bounding boxes in green have respectively for IoU 0.622019 and 0.793165. Only the box with the maximum IoU, that is 0.793165 should be counted as a true positive, the other should be removed and it should not be counted. The small-scale person on a motorcycle is not detected, so in this example there is 1 true positive (one of the two predicted boxes is not counted) and one false negative (the person on the motorcycle).	 Both large scale and small-scale motorcycle people are not detected. There are 2 false negatives here. Apparently, here is the shape of the jacket worn by the person who trained it to go undetected by the HOG-AdaBoost detector.
 Same thing as the previous example, the same person at medium scale is detected twice with two predicted bounding boxes having respectively for IoU of 0.623512 and 0.181492. In this case the box having the IoU of 0, 623512 will be considered as a true positive, however the one with an IoU of 0.181492 will be considered as a false positive.	 The medium scale person is well detected with an IoU equal to 0.653686. There is 1 true positive and one false positive here.
 The medium-scale person is detected with a predicted bounding box having an IoU equal to 0.457869. If the IoU threshold is taken equal to 0.5, then the ground-truth bounding box in blue will be considered as a false negative and the predicted bounding box will be considered as a false positive, which is wrong.	 The medium scale person is well detected with an IoU equal to 0.580978. There is therefore 1 true positive here.
 A large-scale person is detected twice with the green boxes predicted having respectively IoU equal to 0.48906 and 0.0806955, the box with IoU 0.0806955 will be rejected and considered as a false positive. Likewise, the box with the IoU equal to 0.48906 will also be rejected if the IoU threshold is set to 0.5 and it will also be considered as a false positive.	 The large-scale person is detected twice with the predicted bounding boxes in green which have respectively for IoU 0.7074 and 0.153458.The box with the IoU of 0.153458 will be rejected and it will be considered as a false positive. On the other hand, the box with the IoU of 0.7074 will be accepted and considered as a true positive. The medium-scale lady is also detected with the predicted bounding box which has the IoU equal to 0.612489.

Both persons are detected, but they have respectively for IoU 0.437131 (person on the left) and 0.518307 (person on the right). If the IoU threshold is set to 0.5, then only the box predicted for the person on the right with the IoU of 0.518307 will be considered a true positive. On the other hand, the predicted box and the ground truth bounding box for the person on the left will be respectively considered as a false positive and a false negative. So, for the IoU threshold set at 0.5, there is 1 true positive, 1 false positive and 1 false negative, which is not correct.

Both persons are well detected with respectively IoU of 0.543696 (person on the left) and 0.5461 (person on the right). There are therefore 2 true positives here.



Two people are detected, the large-scale man and a lady in the medium-scale crowd. The IoUs obtained are 0.674091 and 0.522472 respectively. Note that the crowd side predicted bounding box overlaps with multiple ground truth framing boxes, but only the ground truth framing box having the highest IoU with predicted box will be taken, the others will be considered false negatives. In this example there are 2 true positives, 3 false negatives and 3 false positives.



Here, three people are detected, the large-scale man and two ladies in the medium-scale crowd, the obtained IoUs are 0.650545 and 0.449743 and 0.48198, respectively. If the IoU threshold is set to 0.5, the two detections in the crowd will be considered false positives and the corresponding ground truth framing boxes will be considered false negatives. With the IoU threshold set at 0.5, the detection in this example gives 1 true positive, 4 false negatives and 2 false positives.

VI. Simple Version of the TP, FN and FP Metrics Evaluation Algorithm

To evaluate the TP (True Positives), FN (False Negatives) and FP (False Positives) metrics, we will start by presenting a first simple version of an algorithm for calculating these values.

For simplicity's sake, let's assume that each person detected in an image is located only once with a predicted bounding box. In other words, there is a single predicted bounding box associated with the ground-truth bounding box framing a detected person.

Algorithm: *Evaluate the number of True Positives, False Negatives and False Positives.*

Input:

- *Database of labeled images.*
- *For each image in the database, we have the list of ground truth bounding boxes and the list of predicted bounding boxes.*
- *The minimum threshold of the IoU.*

Output: *TP, FN et FP.*

We initialize: TP = 0 and FP = 0.
For each image of the database:
    For each detection (or predicted frame box) in the current image:
        Choose from all the ground-truth bounding boxes labeled in the image, the one that has the highest IoU with the predicted bounding box.
        If all the ground-truth bounding boxes in the current image have an IoU below the minimum IoU threshold (typically 0.5), then :
            Detection is a false positive and increments FP : FP = FP + 1
        else:
            The detection is a true positive and we increment TP : TP = TP + 1
Since each predicted bounding box corresponds to one and only one ground-truth bounding box (or a person in the image), one can easily calculate FN by:
FN = Total number of ground-truth bounding boxes in the database – TP

This simple algorithm has the advantage of quickly calculating TP, FN, and FP metrics, but unfortunately it is only suitable if the detector effectively returns a single predicted bounding box for each object detected in an image. In our case, the considered object is a person labeled manually using a ground truth framing box.

This algorithm is therefore suitable for the HOG-AdaBoost detector, but not for the Haar-AdaBoost detector, since this last one sometimes returns two predicted bounding boxes for the same person and therefore this box will be counted twice as a true positive. Whereas normally only one predicted bounding box should be counted as a true positive and the other should be ignored.

Subsequently, we will propose a general algorithm making it possible to correctly calculate the TP, FN and FP metrics. This second version of the algorithm is unfortunately slower in computing time than the previous algorithm, but it has the advantage of working regardless of the number of predicted bounding boxes returned by a detector for the same object (or person) labeled in an image using a ground truth-bounding box.

VII. General Version of the TP, FN and FP Metrics Evaluation Algorithm

It is assumed that the same person can be detected more than once, that is, there are several predicted bounding boxes which may correspond to the ground-truth bounding box framing that person.

Here are two problems that can arise when it comes to associate predicted bounding boxes with ground-truth bounding boxes (or detected person):

- Several predicted bounding boxes can correspond to the same person if they have, together with the ground-truth bounding box framing this person, an IoU greater than a certain minimum threshold of the IoU (typically 0.5). In this case, only a one predicted bounding box should be counted as a true positive, others if not associated with other nearby people should be ignored.

- For people located side-by-side in an image, the ground-truth bounding boxes can usually overlap with each other. In this case, the predicted bounding boxes may also overlap with each other and with several ground-truth bounding boxes. These predicted bounding boxes can therefore have an IoU greater than the minimum threshold with several ground-truth bounding boxes (or several labeled people). We must therefore determine how to correctly associate each predicted bounding box with the ground-truth bounding box it represents (or the person detected).

To overcome these two difficulties and correctly evaluate the TP, FN and FP metrics, that is to say, to avoid repeatedly counting the same person detected with several predicted bounding boxes, which will distort the calculation of TP and FP, we propose a general algorithm whose idea is based on the principle of Non Maximum Suppression (NMS) [23]. We have used this late one to associate each ground-truth bounding box (or detected person) with the predicted bounding box that maximizes the IoU with it and eliminate other predicted bounding boxes that do not maximize the IoU provided that they do not match other people in the vicinity.

Typically on a sliding detection window, the exhaustive search for a person (or an object in general) in an image carried out by certain detectors such as Haar-AdaBoost and HOG-AdaBoost, for example, test all the possible detection windows at all scales and locations. For each of these detection windows, a decision on whether or not it belongs to the desired class is obtained by the detector.

For a person in the initial image, there is a window framing it in the most precise way. However, windows that are spatially close or in scale may also give a positive classification. We then obtain a constellation of positive detection windows around the same detected person, see Fig. 3.

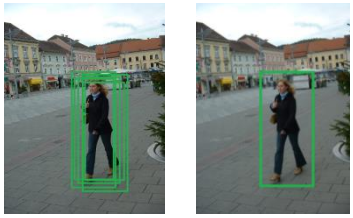


Fig. 3. For the Same Person in the Image, a Multitude of Windows are Detected (the Image on the Left). We must Determine which one Best Frames the Person (the Image on the Right). The Confidence Score is used by the non-maximum Elimination Technique to Find the Window that Maximizes it and to Eliminate the others that do not.

The Non Maximum Suppression technique is one of the methods used during the object detection phase to eliminate neighboring windows that do not maximize the confidence score for a detected object. The confidence score is a value between 0 and 1 generally predicted by a classifier, it represents the probability that a detection window contains an object. The confidence score is used as a comparison value between neighboring detection windows. The principle consists in keeping for a detected object only the detection window which maximizes the confidence score and to eliminate the others which do not maximize it.

In our case, we use the principle of No Maximum Suppression after the phase of the people detection, we based it on the comparison of the IoU between the predicted bounding boxes and those of ground-truth and not on the confidence score. This choice to use the NMS with the IoU was made for the following two reasons:

- The Haar-AdaBoost detector can sometimes generate for the same person detected two predicted bounding boxes that correspond to it.
- The Haar-AdaBoost and-HOG-AdaBoost detectors are respectively based on the binary classifier AdaBoost which do not return a confidence score, but rather the values -1 (or a negative value) for non-membership of the object class to be detected or 1 (or a positive value) to indicate membership of the object class.

The principle of the general evaluation algorithm for TP, FN and FP metrics that we have developed is as follows:

Algorithm: *Evaluate the number of True Positives, False Negatives, and False Positives.*
Input:

- *Database of labeled images.*
- *For each image in the database, we have the list of field truth bounding boxes and the list of predicted bounding boxes.*
- *The minimum threshold of the IoU.*

Output: *TP, FN and FP.*
We initialize TP = 0, FN = 0 and FP = 0.
For each image of the database perform the following processing:

- *Mark all bounding boxes predicted as not being assigned to a ground-truth bounding box.*
- *Associate with each ground-truth bounding box (or a labeled person) in the current image the list of predicted bounding boxes that have an IoU with it that exceeds a certain threshold (typically 0.5). The list of predicted bounding boxes is sorted in descending order of IoUs and all predicted bounding boxes in the list are marked as affected.*
  If the list of predicted bounding boxes is empty, that is, there is no predicted bounding box associated with the ground-truth bounding box, and then this last one is a false negative or a missed person. In this case, we increment the FN metric.
  The final goal of the algorithm is to determine for a ground-truth bounding box framing a person detected in the image one predicted bounding box that maximizes the IoU with it. In this case, only this predicted box will be counted as a true positive, the other predicted bounding boxes on the list if they are not associated with other people located side by side will be ignored.
- *Evaluate the FP metric: it corresponds to the number of predicted bounding boxes that are not marked as assigned to a ground-truth bounding box.*
- *If several detected bounding boxes correspond to the same ground-truth bounding box framing a person (or object in general), the Non Maximum Suppression principle is applied to keep only the detected bounding*

box having a maximum IoU with the ground-truth bounding box. This operation is necessary to properly calculate the TP number, as it avoids counting the predicted boxes for a detected person several times.
For each ground-truth bounding box b1 in the current image:
If the list of predicted bounding boxes associated with box b1 is not empty, then:
The predicted bounding box p1 at the beginning of the list has the maximum IoU. We then take this box p1.
For each ground-truth bounding box b2 in the current image:
If the box p2 at the beginning of the list of predicted bounding boxes associated with box b2 is the same as p1:
If the IoU of p2 with b2 is greater than that of p1 with b1 then it can be confirmed that the predicted bounding box p1 is not associated with the box b1.
Otherwise (the IoU of p2 is smaller than that of p1), we remove p2 from the beginning of the list of the predicted bounding boxes associated with the box b2.
If in the previous iteration it was determined that the predicted bounding box p1 was not associated with b1, then in this case p1 is removed from the beginning of the predicted bounding boxes list associated with box b1.

- *Evaluate the TP metric: it corresponds to the number of ground-truth bounding boxes with a list of predicted bounding boxes associated with them non-empty (these ground-truth bounding boxes therefore correspond to detected people).*

VIII. Analysis of the Two Detectors Haar-AdaBoost and HOG-AdaBoost Performance on the Images of the INRIA Person Dataset

After having implemented the general algorithm for evaluating TP, FN and FP metrics in C ++ language using the OpenCV library, we used it to evaluate the performance of the two detectors Haar-AdaBoost and HOG-AdaBoost.

The following two tables show the results of analyzes obtained by applying respectively the two detectors on 187 first images taken from the 460 bitmap color images of people in the INRIA Person Dataset. Manual labeling of the people in the 187 images resulted in a total of 481 ground-truth bounding boxes framing these people.

TABLE II. Result Obtained by Applying the Haar AdaBoost-Detector on 187 Images from INRIA Person Dataset Containing 481 People Labeled with Ground-Truth Bounding Boxes

Haar-AdaBoost					
IoU threshold	TP	FN	FP	Precision	Recall
0,4	179	302	90	0,665428	0,372141
0,5	144	337	127	0,531365	0,299376
0,6	73	408	203	0,264493	0,151767
0,7	18	463	261	0,064516	0,037422
0,8	2	479	277	0,007168	0,004158
0,9	0	481	279	0	0

TABLE III. Result Obtained by Applying the HOG-AdaBoost Detector on 187 Images from the INRIA Person Dataset Containing 481 People Labeled with the Ground-Truth Bounding Boxes

HOG-AdaBoost					
IoU threshold	TP	FN	FP	Precision	Recall
0,4	258	223	368	0,412141	0,536383
0,5	191	290	441	0,302215	0,397089
0,6	103	378	532	0,162205	0,214137
0,7	31	450	604	0,048819	0,064449
0,8	2	479	633	0,00315	0,004158
0,9	0	481	635	0	0

Since the two detectors are applied to the same images in the INRIA Person Dataset, we will start by making a simple comparison by plotting the curves of the True Positives as a function of the Intersection over Union (TP-IoU), False Positives as a function of the Intersection on the union (FP-IoU) and False Negatives as a function of the Intersection on the union (FN-IoU) (see these curves in Fig. 4).

It can be seen from the TP-IoU curve in Fig. 4 that the HOG-AdaBoost detector (curve in red) is more efficient than the Haar-AdaBoost detector (curve in blue), since it allows to detect more positives (or the persons labeled) than Haar-AdaBoost.

Likewise, the FP-IoU curve also shows that there are fewer false negatives or misses' people with HOG-AdaBoost than with Haar-AdaBoost.

On the other hand, the HOG-AdaBoost detector is less efficient than Haar-AdaBoost with regard to false detections, since in return for its efficiency in detecting positives, it has the disadvantage of generating a lot of false detections or false positives.

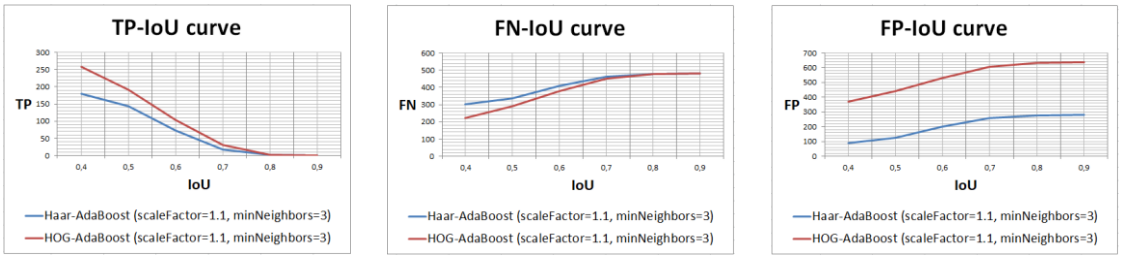


Fig. 4. These Curves show that the HOG-AdaBoost Detector (in Red) is more efficient to Detect People than Haar-AdaBoost (in Blue), but it Generates more False Detections than the Latter. These Results were Obtained for the Values of the Detection Parameters ScaleFactor and MinNeighbors respectively Equal to 1.1 and 3.

Since the two detectors are based respectively on the AdaBoost binary classifier which does not return a confidence score, but rather a response indicating whether the detected object is part of the required class or not, it will therefore not be possible to use the curve Precision-Recall that can be used to calculate the AP (Average Precision) metric. We will therefore use in its place the Recall-IoU curve which makes it possible to calculate the AR (Average Recall) metric.

Subsequently, we will complete the comparisons made by the curves in Fig. 4 by plotting the Recall-IoU curve (Fig. 5). This is more general than the previous curves, it is often used to study the efficiency in detecting true positives; in addition it allows evaluating the average recall metric AR (Average Recall) which is used to compare detectors even if they are applied to different image databases.

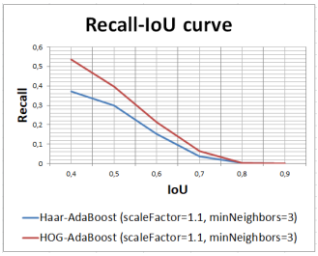


Fig. 5. This Curve shows that the HOG-AdaBoost Detector (in Red) is more Efficient at Detecting People than the Haar-AdaBoost Detector (in Blue).

Knowing that the AR metric corresponds to the area of the region below the Recall-IoU curve between the values of IoU 0.5 and 1 and, it is given by equation (1). To evaluate this metric, we will approximate the integral (1) using the rectangle method that is given by the equation (2):

$$AR = 2\int_{0,5}^{1} Recall(IoU) \quad (1)$$

$$AR = 2\sum_{i=1}^{n-1}(IoU_{i+1} - IoU_i)\, Recall(IoU_{i+1}) \quad (2)$$

Here, IoU_1 is equal to 0.5 and IoU_n is equal to 1. The interval [0.5, 1] is divided into n intervals of the same length equal to $IoU_{i+1} - IoU_i = \frac{1-0,5}{n} = \frac{0,5}{n}$, $1 \le i \le n - 1$.

Since in our case, we have taken n = 5 and the IoU variable between 0.5 and 1, we can deduce that the step of the variation will be fixed at $\frac{1-0,5}{n} = \frac{0,5}{5} = 0,1$, equation (2) will become:

$$AR = 2 \times 0,1 \times \sum_{i=1}^{n-1} sensibility(IoU_{i+1}) \quad (2)$$

With IoU1 = 0,5 and IoUi+1 = IoUi + 0,1, $1 \le i \le n - 1$.

Based on the detections results obtained by the Haar-AdaBoost and HOG-AdaBoost detectors and which are presented in Tables II and III, respectively, we evaluated the AR metric for each of the two detectors which made it possible to obtain the following result:

- Haar-AdaBoost : AR = 0,0985446.
- HOG-AdaBoost : AR = 0,1359666.

From the plot of the Recall-IoU curve and the evaluation of the AR metric for both detectors, the HOG-AdaBoost detector is more efficient to detect people than Haar-AdaBoost, because the HOG-AdaBoost AR is larger than the Haar-AdaBoost AR. But unfortunately, according to the FP-IoU curve in Fig. 4, the HOG-AdaBoost detector has the disadvantage of generating a lot of false detections than the Haar-AdaBoost detector.

IX. Experimenting by Changing Certain Detection Parameters

To perform the detection of people, we used the detectMultiScale method of the CascadeClassifier class. It admits seven parameters, the most important that can be varied to study the detection of people or objects in general are the following two parameters:

- scaleFactor : Allows to define how much the size of the detection window will be reduced with each iteration. The default value for this parameter is 1.1.
- minNeighbors : Defines the minimum number of neighboring detections that a candidate area must have to be retained. The default value for this parameter is 3.

The other parameters are: the image matrix, the flags (not used in detection), minSize (minimum size of the object, the default value is size 0×0) and maxSize (maximum size of the object, the value by default is the full size of the image) are not important for detection.

The results of the analyses presented in paragraph 8 above were obtained using the values of scaleFactor and minNeighbors parameters respectively equal to the default values 1.1 and 3.

We repeated these experiments by assigning to the scaleFactor parameter the fixed value 1.1 and by varying the value of the minNeighbors parameter by assigning it the successive values 2, 3, 4 and 5.

The results of the analyses obtained by the two detectors Haar-AdaBoost and HOG-AdaBoost are respectively shown in Fig. 6 and Fig. 7.

The TP-IoU, FN-IoU and Recall-IoU curves show that when the value of the minNeighbors parameter decrease from 5 to 2, the detection of people (or true positives) improves by both detectors, but, in return, the number of false detections (or false positives) increases (see the FP-IoU curve).

It can be seen that the lower the value of the minNeighbors parameter, the better the detection of people and the higher the number of false detections. A compromise between good detection and false detections can be achieved by the intermediate value of minNeighbors equal to 3 (values 4 and 5 also give a suitable result).

To complete this study, we also assigned other values to the parameters (scaleFactor,minNeighbors), such as (1.01, 5), (1.01, 4), (1.01, 3), (1.01, 2), (1.05, 5), (1.05, 4), (1.05, 3), (1.05, 2), (1.1, 5), (1.1, 4), (1.1, 3), (1.1, 2), (1,15,5),(1.15, 4),(1.15, 3), (1.15, 2), (1.2, 5), (1.2, 4), (1.2, 3),(1.2, 3).

The curves in Fig. 8 (Haar-AdaBoost) and Fig. 9 (HOG-AdaBoost) were obtained for the values of (scaleFactor,minNeighbors) equal to (1.01, 3), (1.05,3), (1.1,3), (1.15,3) and (1.2,3), they give an idea for comparing the

detections that we obtained by varying the values of the scaleFactor and minNeighbors parameters as shown above.

The analyses we performed for the scaleFactor parameter value varying from 1.01 to 1.2 and the minNeighbors parameter value fixed at 3 and which are illustrated by Fig. 8 and 9 allowed us to deduce the following conclusions:

- When the value of the scaleFactor parameter decreases, there is an overall improvement in the people detection due to an increase in the number of true positives (see the TP-IoU, FN-IoU and Sensitivity-IoU curves).

Unfortunately, this improvement is achieved in detriment of an increase in false detections (or the number of false positives, see the FP-IoU curve) and also in the time of detections calculation.

Tables IV and V also confirm the previous results, they give an overview of the detection rates obtained by the two detectors when the IoU value is set at 0.5, that of the minNeighbors parameter is set at 3 and by varying the scaleFactor parameter value which successively takes the values 1.01, 1.05, 1.1, 1, 15 and 1.2.

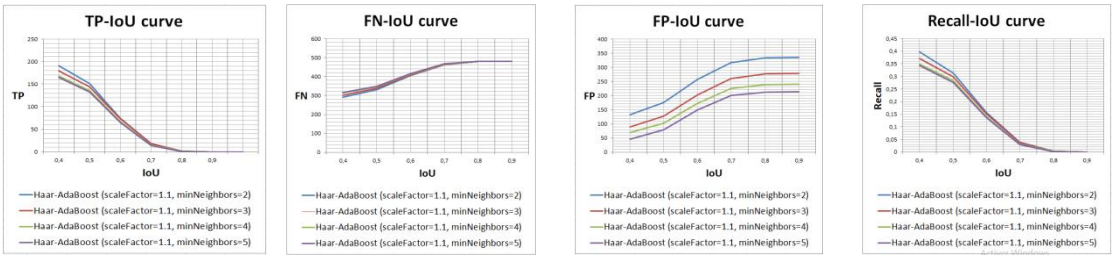


Fig. 6. Haar-AdaBoost Detector.These Curves show the Results of the Detections Analysis Obtained by the Haar-AdaBoost Detector by Setting the Value of the ScaleFactor Parameter to 1.1 and Varying the Value of the minNeighborsen Parameter Successively Assigning it the Values 2, 3, 4 and 5.

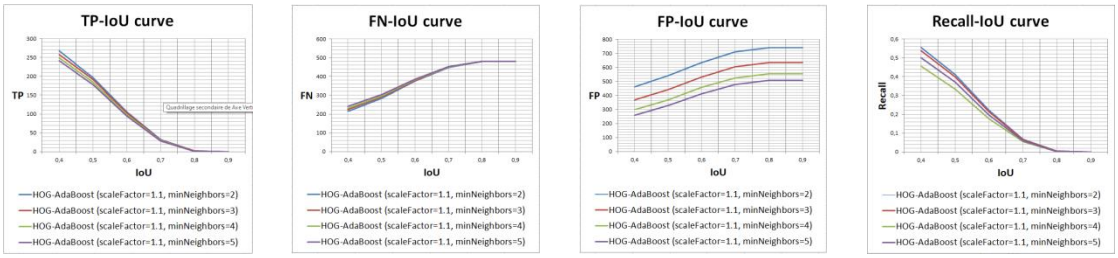


Fig. 7. HOG-AdaBoost Detector. These Curves show the Results of the Detection Analysis Obtained by the HOG-AdaBoost Detector by Setting the Value of the ScaleFactor Parameter to 1.1 and Varying the Value of the minNeighbors Parameter by Successively Assigning it the Values 2, 3, 4 and 5.

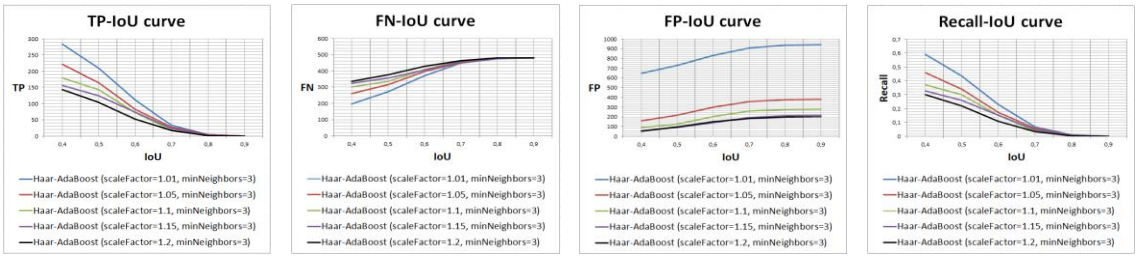


Fig. 8. Haar-AdaBoost Detector. These Curves show the Results of Detections Analysis Obtained by the Haar-AdaBoost Detector by Varying the Value of the ScaleFactor Parameter Assigning it the Successive Values 1.01, 1.05, 1.1, 1.15 and 1.2 and Setting the Value of the minNeighbors Parameter to 3.

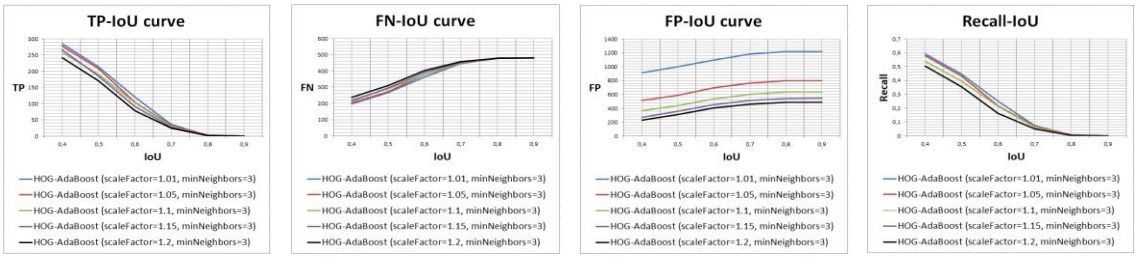


Fig. 9. HOG-AdaBoost Detector. These Curves show the Results of Detections Analysis Obtained by the HOG-AdaBoost Detector by Varying the Value of the ScaleFactor Parameter Assigning it the Successive Values 1.01, 1.05, 1.1, 1.15 and 1.2 and Setting the Value of the minNeighbors Parameter to 3.

Knowing that in the first 187 images of the INRIA Person Dataset, we have labeled 481 people by ground-truth bounding boxes, in this case, the detection rate (equal to the recall metric) will therefore be equal to the number of true positives detected in all 187 images divided by 481, that is, equal to $\text{Recall} = \frac{\text{TP}}{481}$.

It can be seen from Tables IV and V that the detection rate obtained by the two detectors is overall less than 50%, it increases when the scaleFactor parameter decreases from 1.2 to 1.01.

TABLE IV. DETECTION RATES OBTAINED BY THE HAAR-ADABOOST DETECTOR BY SETTING THE VALUE OF THE IOU TO 0.5, THAT OF THE MINNEIGHBORS PARAMETER TO 3 AND BY VARYING THE VALUE OF THE SCALEFACTOR PARAMETER

Haar-AdaBoost detector					
scaleFactor	1.01	1.05	1.1	1.15	1.2
TP	210	165	144	125	105
Detection rate in %	43,66	34,30	29,94	25,99	21,83

TABLE V. DETECTION RATES OBTAINED BY THE HOG-ADABOOST DETECTOR BY SETTING THE IOU VALUE TO 0.5, THAT OF THE MINNEIGHBORS PARAMETER TO 3 AND VARYING THE VALUE OF THE SCALEFACTOR PARAMETER

HOG-AdaBoost detector					
scaleFactor	1.01	1.05	1.1	1.15	1.2
TP	215	208	191	186	171
Detction rate in %	44,70	43,24	39,71	38,67	35,55

- The respective default values 1.1 and 3 of the two parameters scaleFactor and minNeighbors are central; they allow obtaining a suitable detection result that provides a compromise between true detections and false detections and a reasonable calculation time.
- The parameters (scaleFactor,minNeighbors) pairs of values (1.15,3) and (1.2,3) also provide a suitable detection result, since the TP-IoU, FN-IoU and FP-IoU and Recall-IoU curves obtained for these two pairs of values are very close to those obtained for the pair value (1.1,3). In addition, these two pairs of values make it possible to carry out detections with a lower calculation time than that obtained for (1.1, 3).

X. TRAINING THE HAAR-ADABOOST DETECTOR BY SUPERVISED LEARNING ON THE INRIA PERSON DATASET IMAGES

Training a classifier is a long step. It requires gathering and annotating a large number of images containing the object to be detected (positive images) and images not containing the object to be detected (negative images).

In our case, we used the first 187 images taken from 460 images in the INRIA Person Dataset. The manual people labeling in these 187 images resulted in 481 ground-truth bounding boxes that we will use to train the Haar-AdaBoost detector by supervised learning.

The training database therefore consists of 187 positive images and 273 negative images (also called background images). All these images are taken from the INRIA Person Dataset. The positive images are labeled in 481 people who will be used jointly with the negative images during the learning process as training examples of the Haar-AdaBoost detector.

The aim of this experiment is to test whether we can improve the detection of people on a medium and large scale by injecting into the learning database examples of people images on medium and large scales. During the learning phase, the training of the detector with the opencv_traincascade program takes a lot of time depending on the number of positive and negative images and the size w×h. In our case, the number of positive and negative images was set at 481 and 273 respectively. The training time of the Haar-AdaBoost detector increases according to the used size w×h.

For example, this time takes 1 hour and 50 minutes for the size 24×24, 3 days and 21 hours for the size of 32×32 and more than 5 days for the sizes 64×64, 24×60 and 32×80 on the Intel (R) Core (TM) microprocessor having the frequency of 1.8 GHz and a RAM memory of 4 GB.

After training the detectors for sizes 24×24, 32×32, 64×64, 24×60 and 32×80, we applied them to the INRIA images for analyzing the results obtained.

Fig. 10 shows some images of people detections obtained by Haar-AdaBoost detectors formed with sizes 64×64, 24×60 and 32×80. The frames in blue are the ground-truth bounding boxes, while the green frames correspond to the predicted bounding boxes.

These detections were obtained with the values 1.1 and 3 assigned respectively to the scaleFactor and minNeighbors parameters of the detectMultiScale method defined in CascadeClassifier class provided by OpenCV.

The detection results obtained by Haar-AdaBoost detectors formed with sizes 24×24, 32×32 are very bad, there is practically no detection of people and generate a very high number of false detections (see the curves in blue and red that are often confused in Fig. 11).

On the other hand, the results obtained by the detectors formed with sizes 64×64, 24×60 and 32×80 are suitable, they resemble practically to the detection results obtained with the detector provided by OpenCV.

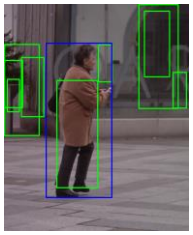
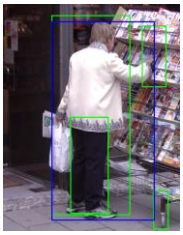
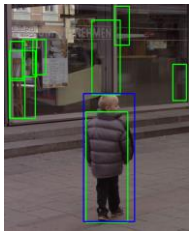
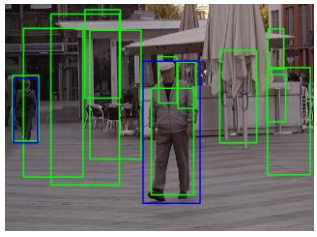
In addition, since the learning examples of these detectors contained many people on a large scale, they thus made it possible to slightly exceed the OpenCV detector in terms of detecting people on a large and medium scale (see Fig. 10 and 11).

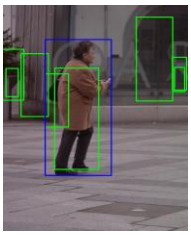
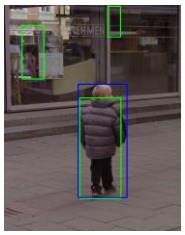
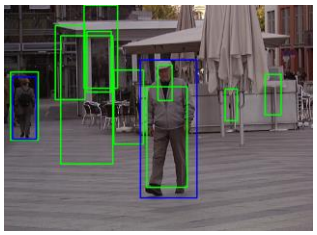
In addition, the TP-IoU, FN-IoU and Recall-IoU curves in Fig. 11 also show that detectors trained for sizes 24×60 and 32×80 provide a better detection result with regard to the number of true positives that is larger and the number of false detections that are smaller than those provided by the detector formed for size 64×64. This result comes from the fact that the aspect ratio, that is, the ratio of width to height, chosen for the

detectors 24×60 and 32×80 is equal to 0.4 ($\frac{24}{60} = \frac{32}{80} = 0.4$) which generally corresponds to the aspect ratio of people standing.

Unfortunately, the disadvantage of these detectors thus formed is that they generate a very high number of false detections compared to those generated by the OpenCV detector (see Fig. 10 and 11), this is most likely due to the number of negative (273 images) and positive (481 positive images of persons) examples of learning which is very low.

Normally, to train correctly a detector, it actually takes thousands of positive and negative examples, which requires gathering a very large number of positive images containing people to be labeled and negative images not containing people. In this case, the training of the detector will require a very high learning time.

64×64			
IoU = 0.43435	IoU = 0.725217	IoU = 0.700581	(man) IoU = 0.558974 (lady) IoU = 0.860678
			

24×60			
IoU = 0.496905	IoU = 0.694407	IoU = 0.737387	(man) IoU = 0.524845 (lady) IoU = 0.79023
			

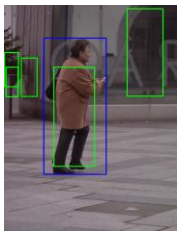
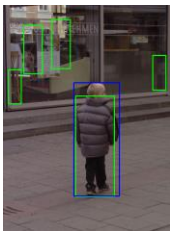
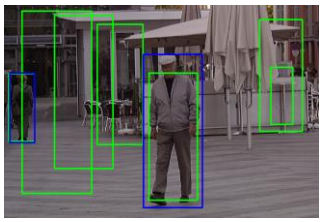
32×80			
IoU = 0.477611	IoU = 0.711113	IoU = 0.748899	(man) IoU = 0.681874 (lady) IoU = 0.940507
			

Fig. 10. People Detection Obtained by Haar-AdaBoost Detectors Formed respectively with Sizes 64×64, 24×60 and 32×80.

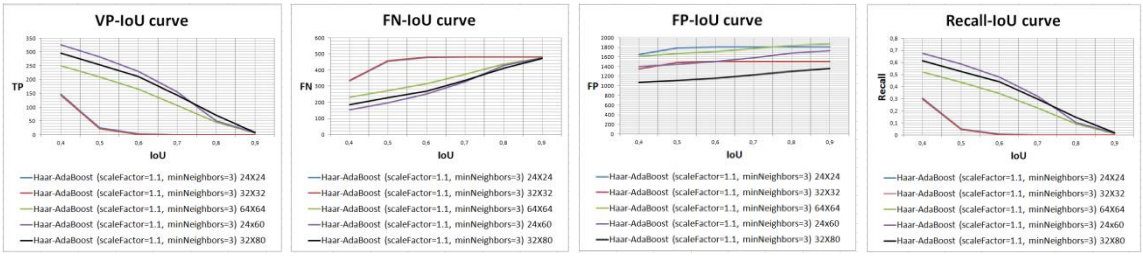


Fig. 11. These Curves show the Results of Detections Analysis Obtained by the Haar-AdaBoost Detector Formed for Sizes 24×24, 32×32, 24×60 and 32×80.The ScaleFactor and minNeighbors Parameters have Values of 1.1 and 3, respectively.

XI. CONCLUSION

In this article, we first studied the two detectors Haar-AdaBoost (VJ) and HOG-AdaBoost (PoseInv) following the study that we did before in the paper [19].

After having studied the two methods we made a comparison of two approaches Haar-AdaBoost and HOG-Adaboost which constitutes a variant of HOG-SVM [12]. Secondly and after modifying certain detection parameters, we carried out an evaluation of the experiments found to have more performance.

The application of these two detectors on the images taken from the INRIA Person Dataset enabled us to draw the following conclusions:

- The HOG-AdaBoost detector is more efficient at detecting people on a medium scale (or nearby) than the Haar-AdaBoost detector, but on the other hand, it generates much more false detection than the latter.
- Generally, the two detectors studied do not correctly detect people on a small scale (or distant) and on a very large scale (or very close). This is most likely due to the training examples that were used to train these two detectors which contained very few examples of people on a small and on a very large scale.
- Sometimes the shape of the clothing, people close together, crowds, etc. can prevent these detectors from properly detecting people in images.
- The detection rate of people obtained by the two detectors Haar-AdaBoost and HOG-AdaBoost is less than 50%.

In an attempt to improve the detection of people at medium and large scale, five Haar-AdaBoost detectors were formed for the respective image sizes 24×24, 32×32, 64×64, 24×60 and 32×80 on an image database containing many examples of medium and large scale people.

There are practically no detection results provided by 24×24 and 32×32 detectors. In contrast, 64×64, 24×60 and 32×80 detectors have improved the performance of detecting people at medium and large scale compared to the detector provided by OpenCV, but on the other hand, they generate a very high number of false detections. This disadvantage is probably due to the reduced number of positive and negative images that we used to train these detectors.

Unfortunately, it is not possible to apply the finetuning operation to the Haar-AdaBoost and HOG-AdaBoost detectors. This operation consists of re-training a detector already trained on new examples in order to readjust it so that they can adapt to the recognition of these new examples, such as for example in our case, the detection of small, medium and large scale people.

In practice, the fine-tuning operation is preferable to training a new detector on a new sample database which is a very computationally expensive operation. The fact that this operation is not supported by Haar-AdaBoost and HOG-AdaBoost, this is a disadvantage of these detectors, as it will be difficult to expand the capacity of these detectors to new examples.

Another disadvantage of the Haar and HOG descriptors is that they only allow to process grayscale images and only take into account the shape of the objects.

An alternative to the Haar-AdaBoost and HOG-AdaBoost detectors is to use deep convolutional neural network models. Indeed, the latter have made it possible to obtain great performances by their training for the detection of objects [24, 25, 26] and in particular of people [27, 28, 29, 30, 31, 32].

In addition, it is very easy to expand the capacity of an already trained deep convolutional neural network to new learning examples through the fine-tuning operation.

Deep convolutional neural networks also have the advantage of being applied to color images, which gives them the ability to take into account not only the shape of objects, but their texture and color as well.

REFERENCES

[1] Ming-ShiWang and and Zhe-Rong Zhang, “FPGA implementation of HOG based multi-scale pedestrian detection”, Proceedings of IEEE International Conference on Applied System Innovation,ISBN={978-1-4503-5614-5}, 2018.

[2] Jia Xiang Zhao and Jun Li, “RPN+ Fast Boosted Tree: Combining deep neural network with traditional classifier for pedestrian detection”, ACM Digital Library, Volume 47 Issue C, Sep 2016.

[3] Xiaowei Zhang Li Cheng, Hai-Miao Hu, “Too far to see? Not really! pedestrian detection with scale–aware localization policy”, University Library, Volume {abs/1709.00235}, 2017.

[4] J. Leskovec, Sentjost, and Slovenia, “Detection of human bodies using computer analysis of a sequence of stereo images”. 11th European Union Contest for Young Scientists, 1999.

[5] B. Wu and R. Nevatia, “Tracking of multiple, partially occluded humans based on static body part detection” in IEE Proc. of the IEEE International Conference on Computer Vision, pp. 886–893, 2005.

[6] P. Viola, M. J. Jones and D. Snow, “Detecting pedestrians using patterns of motion and appearance”, in International Journal of Computer Vision, pp. 153–161, 2005.

[7] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, “Humain presence detection and charactezation of activity”, In XXIIe conference GRETSI (signal and image processing), Dijon (FRA), 2009.

[8] L. Biancardini, S. Beucher, L. Letelllier, "Object extraction in motion: a mixed approach contours-regions" ORASIS 2005 9th Congress Young Researches in Computers Vision, Clermont-Ferrand, France, May 2005

[9] S. Zhang, R. Benenson, M. Omran, J. Hosang and B. Schiele, “How far are we from solving pedestrian detection”, Cornell University Library, (2016).

[10] Qiang Zhu; Mei-Chen Yeh; Kwang-Ting Cheng; S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients", Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006.

[11] S. Zhang, R. Benenson, B. Schiele, “Filtered feature channels for pedestrian detection”, in ‘CVPR’, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1751-1760, (2015).

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, , pp. 886-893 vol. 1. (2005).

[13] S. Singh, R. Mittal, “Image segmentation using edge detection and poincare mapping method”. European Journal of Advances in Engineering and Technology, Volume2, Issue4, p. 81-83. (2015).

[14] P. Viola and M. Jones. “Rapid object detection using a boosted cascade of simple features”, in IEE Proc. Of the conference on Computer Vision and Pattern Recognition, pp. 511-518, 2001.

[15] P. Viola and M. Jones "Robust real-time face detection". International Journal of Computer Vision, Spring Journal, pages 137–154, May 2004.

[16] P. Viola, M. J. Jones and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance", in International Journal of Computer Vision, pp. 153–161, 2005.

[17] C. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection", in Proc. of the IEEE International Conference on Computer Vision, pp. 555-562,1999.

[18] B. Wu, R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors", International Journal of Computer Vision 75, no. 2 (2007): 247-266. CA 90089- 0273, 2007.

[19] N. Ould Taleb, A. Chergui, M. L. Ben Maâti, M. F. Nanne, "Overview on automatic detection of human body", IEEE Xplore, 24 April 2017, ISSN: 2472-7652. DOI: 10.1109/ICMCS.2016.7905638.

[20] Z. Lin, L. S. Davis, A Pose-Invariant descriptor for human detection and segmentation. In Computer Vision – ECCV 2008. Lecture Notes in Computer Science, vol 5305. Springer, pp. 423-436. Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88693-8_310.

[21] V. Vapnik, The nature of statistical learning theory. Springer-Verlag, 1995.

[22] Yoav Freund and Robert Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, vol. 55, no 1, 1997, p. 119-139.

[23] Navaneeth Bodla, Bharat Singh, Rama Chellappa, Larry S. Davis, "Soft-NMS – Improving object detection with one line of code" ,Computer Vision (ICCV), IEEE International Conference, 2017

[24] Y. LeCun, B. Boser, J.S. Denker, D. Henderson R.E. Howard, W. Hubbard, and L.D Jackel (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, MIT Press.1(4):541–551.

[25] S. Ren, K. He, R. Girshick. and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, June 2016.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed4, Cheng-Yang Fu1, Alexander C. Berg (2016). SSD: Single shot multiBox detector, European Conference on Computer Vision, pages 21-37, Oct 2016.

[27] L. Zhang, L. Lin, X. Liang., K. He, Is faster R-CNN doing well for pedestrian detection?. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_28.

[28] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang and Bernt Schiele. How far are we from solving pedestrian detection. Cornell University Library, (2016).

[29] Nagi Ould Taleb, Adil Chergui, Mohamed Larbi Ben Maâti, Mohamedade Farouk Nanne, Mohamed O. M. Khelifa, Aicha Mint Aboubekrine, Pedestrian detection and the effect of diverse benchmarks. International Journal of Scientific & Engineering Research, Volume 9, Issue 9, pages 922-927. September 2018.

[30] Ujjwal, Aziz Dziri, Bertrand Leroy, Francois Bremond (2018). Late fusion of multiple convolutional layers for pedestrian detection. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). DOI: 10.1109/AVSS.2018.8639083.

[31] ZHANG, X., CHENG, L., LI, B., AND HU, H.-M (2018). Too far to see? not really! A pedestrian detection with scale-aware localization policy. IEEE transactions on image processing 27, 8 (2018), 3703–3715.

[32] Farzin Ghorban, Javier Marín, Yu Su, Alessandro Colombo, and Anton Kummert, "Aggregated channels network for real-time pedestrian detection", Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017), 106960I (13 April 2018); https://doi.org/10.1117/12.2309864.

Towards Stopwords Identification in Tamil Text Clustering

M.S. Faathima Fayaza[1]
Department of Information Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka

F. Fathima Farhath[2]
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka

***Abstract*—Now-a-days, digital documents have become the primary source of information. Therefore, natural language processing is widely utilized in information retrieval, topic modeling, document classification, and document clustering. Preprocessing plays a significant role in all of these applications. One of the critical steps in preprocessing is removing stopwords. Many languages have defined their list of stopwords. However, a publicly available stopwords list isn't available for the Tamil language since it is under-resourced. This study identified 93 general and some domain-specific stopwords for sports, entertainment, local and foreign news by analyzing more than 1.7 million Tamil documents with more than 21 million words. Also, this study shows that removing stopwords improves the accuracy of a Tamil document clustering system. It showed an improvement of 2.4%, 0.95% in the F-score for TF-IDF with one pass algorithm and FastText with the one-pass algorithm, respectively.**

Keywords*—Stopwords; Tamil; pre-processing; TF-IDF; clustering

I. Introduction

Usage of digital text information is growing exponentially in today's world, not only in English but also in other regional languages. Managing this data and extracting the relevant information has become a challenge. Henceforth, Natural Language Processing (NLP) has emerged as a new research field. Yet, the text is more challenging to manipulate and process than numerical data since it is unstructured, ambiguous, and difficult to manipulate. Information retrieval, document clustering, question and answering, document classification, sentiment analysis, and text summarization are some trivial applications of NLP. In every such application, the very first step is data preprocessing. Typically, data preprocessing includes tokenization, stemming, lemmatization, and stopwords removal [1]. Preprocessing eliminates the noise in the data. Further, preprocessing improves the performance of applications. More than 70% of the total text classification process comprises of preprocessing of text alone [2].

Stopwords are the frequently occurring words in a language containing very little or no meaning when used alone. They influence the syntax of a language rather than the semantics of a language [3]. "Are, is, be, a, the, an, of" are some examples of stopwords in English [3]. Therefore, removing stopwords shrinks the size of the text corpus by nearly 35-45% [4] by leaving only the semantically significant words. Also, this aids in improving the accuracy and efficiency of application as the emphasis is given to the semantic of the text. For instance, in document classification, the corpus size reduction reduced the time needed to train a model [5].

In the literature, there are several studies conducted for stopwords identification for many languages such as English, Hindi, Arabic, Chines [4], [6]. Yet, there is no publicly available list of stopwords for Tamil as less research has been done. This paper focuses on identifying stopwords for the Tamil language. Tamil is one of the highly agglutinative languages. The language is used in Sri Lanka, India, Canada, Malaysia, and many other parts of the world. Also, its use in the online platform has shown a hike in the recent past. Therefore, there is a notable need for NLP applications to make use of these available online resources. Therefore, developing and defining stopword lists will benefit the preprocessing step for Tamil language NLP applications. In this research, the authors identify 93 general stopwords for Tamil. Apart from that, the authors identify some domain-specific stopwords for the domain of sports, entertainment, international and local news. Further, by incorporating the preprocessing step of stopwords removal in Tamil text clustering, an improvement of 2.4 and 0.95 in F-score were noted for TF-IDF with one pass algorithm and FastText with one pass algorithm accordingly.

The remaining part of the paper is laid out as follows. Section II elaborates on the related work. Section III details the methodology of stopwords generation for the Tamil language. Section IV reports on the evaluation conducted. Section V analyzes the result and discusses the findings. Section VI conveys the conclusion of the research.

II. Related Work

Fox[4] created a stopword list for English using Brown Corpus of 1,014,000 words. Here the author manually added the words that appeared more than 300 times in the corpus in a list and finalized it by manually analyzing. The stopword list contained 421 words. This approach is domain-independent and is widely used in retrieval systems. Hao et al. [6] generated a stopword list using the weighted Chi-squared statistic technique for the Chinese language. Researchers observed that the suggested methodology effectively improved the F1 classification score by nearly 7%.

Raulji et al. [7] proposed a dictionary-based approach to remove the stopwords for the Sanskrit Language. They used a predefined word list, compared it with the targeted text, and removed the stopwords. The researchers stated that from

87,000 words corpus, 11,200 words were removed as stopwords. This reduced the corpus size by 13% and reduced the feature space and CPU cycle.

Saif et al. [8] generated stopwords for the twitter sentimental analysis using a semantic approach. The researchers concentrated on word semantics and contextual semantics of the words. For this study, six different datasets were used. It was noticed that using semantically found out stopwords improved the accuracy by 0.42% and F-score by 0.94% than using a classic stopword list. Further stopwords removal reduced the classification features by 48.34% and size by 1.17% compared to traditional approaches. Miretie and Khedkar[9] generated a stopword list for the Amharic language by applying aggregated term frequency, entropy and inverse document frequency. El-Khair [10] conducted a comparative study to determine the effect of stopwords removal for the Arabic language. The notion of this study is to combine statistical and linguistic approaches. This study used three stopword lists: general list, corpus-based list, and combined list. Also, they used inverse document frequency, probability weight, and statistical modeling approaches. It was concluded that the general stopword list performed better than the latter two.

Bouzoubaa et al. [11] standardized the Arabic stopword list. They [12] used a statistical approach to find stopwords in the Arabic language. They concluded that this approach increased the performance of the Artificial Neural Network (ANN) classifier than when the general stopword list was used.

Ghag and Shah [5] studied the consequences of stopwords elimination in sentiment analysis and reported that the traditional classifier accuracy improved from 50% to 58.6% when stopwords were removed. However, when applying for "Average Relative Term Frequency Sentiment Classifier," "Senti-Term Frequency Inverse Document Frequency," and "Relative Term Frequency Sentiment Classifier," the improvement was insignificant.

Gunasekara and Haddela [13] created a domain-specific stopword list for the Sinhala language. Reported improved precision, recall, F-score, and accuracy when removed the stopwords in Naïve Bayes and Maximum Entropy-based classifier to classify the Sinhala news articles.

Ladani and Desai [3] surveyed the available stopwords removal techniques for Indian and Non-Indian Languages. They classified stopwords into two main categories as general and domain-specific. They were reported that removing stopwords reduces the size and improves the accuracy of text classification.

Jha et al. [14] proposed a Deterministic Finite Automata (DFA) based stopwords elimination algorithm for the Hindi language. Used JSON objects to implement DFA. For this study total of 200 Hindi documents were used as input, including a movie review dataset gathered from the internet. Here the accuracy and efficiency improved for text preprocessing.

Jayaweera et al. [15] proposed a dynamic approach to find Sinhala stopwords. In this study, they argued the cutoff point is subjective to the dataset. This study used 90,000 documents.

Wijeratne and de Silva [16] collected the data from patent documents between 2010-2020 and created a corpus with 540,276 words of Sinhala text and listed the stopwords using term frequency. Sarica and Luo [17] identified the stopwords in technical language.

Rakholia et al. [18] proposed a rule-based approach to detect stopwords for the Gujarati language dynamically. They developed 11 static rules and used them to generate a stopword list at runtime. They attained 98.10% accuracy for generic stopwords detection and 94.08% for domain-specific stopwords detection.

Multiple approaches have been tried in different languages in the literature to identify the stopwords. Those are mainly: a manual, dictionary-based, rule-based, statistical approach, term frequency, weighted term frequency, inverse document frequency. Most of them are static approaches and data-dependent. Further, these approaches require an intense amount of resources to gain better accuracy. Tamil is a low-resourced and highly inflected language. Therefore, applying these techniques directly to the Tamil language will not be feasible. Hence this study presents a dynamic approach for Tamil stopwords identification.

III. Stopwords Identification for the Tamil Language

A. Data Collection and Preprocessing

The data collected by Fayaza and Ranatunga [19] is utilized in this study. Datasets contain more than 1.7 Tamil documents with more than 21 million words. The datasets contain two types of data. Namely.

1) Online news data: The online news data was collected from nine news providers and covered the local, international, sport, business, and entertainment news. Each news consisted of title, body, URL, and date published. Every news domain is identified using a URL, and domain-specific groups were created as the first step. Using this data following datasets were created.

Dataset 1: International news.

Dataset 2: Sports news.

Dataset 3: Local news.

Dataset 4: Entertainment news.

The following procedures were carried out as preprocessing for all the datasets created above. Using <TITLE> and <BODY> tags, the news title and body were identified. Then the data was tokenized into distinct terms using the white space characters like space, tab, newline/carriage return, and punctuation marks. This was followed by the removal of non-Tamil characters and punctuation marks.

2) General data: It contains randomly collected data from multiple sources. The same preprocessing steps carried out on the previous data set were carried out.

B. Stopword Identification

Several approaches were implemented to identify the stopwords in different languages in the literature. However, most of these approaches are based on the term frequency of the text. In this study, term frequency (TF), inverse document frequency (IDF), and term-frequency-inverse-document-frequency (TF-IDF) are calculated for every term in the dataset. Authors select this approach since it is independent of the dataset size and domain, and it has been used in many other low-resource languages [15]. Further authors conducted multiple executions with different values to define the threshold. From that, the generated threshold value is selected by analyzing the stopword lists.

This automatic identification process consists of the following set of procedures:

- Calculate term frequency (TF) for each term in the document ($TF_{t,d}$).
- Calculate the document frequency (DF) for each term.
- Calculate the Inverse document frequency (IDF_t) : log10(N/dft) (N: total number of documents in the dataset).
- Calculate the TF*IDF for each term.
- Calculate the average TF*IDF for each term.
- List the TF*IDF in order.
- If TF*IDF is lower than threshold value term, added to stopword list.
- Identify intersect words in the lists and create a general stopword list.

It was identified that some stopwords are common among all the domain-specific stopword lists. Therefore, this study categorizes those words under general stopwords for the Tamil language.

IV. Evaluation

To date, there is no published work or publicly available stopword list for Tamil. This study created two types of stopwords lists (One general and four domain-specific). Three individuals manually evaluated these generated lists. Fleiss' Kappa statistic [20] was used to assess the agreement between the evaluators, which was 93.0.

Stopwords removal is one of the basic preprocessing steps in NLP. To evaluate the impact of stopwords removal on system performance of the NLP system, the generated lists were utilized for stopword removal in clustering for Tamil news, using ten datasets as of [19]. The incorporation was done over the same two approaches used in [19]. Those are:

1) TF-IDF with one pass algorithm [19].
2) FastText with one pass algorithm [19].

The system without removing stopwords was used as the baseline for this experiment. Then the stopwords removed datasets were tested. Generated results are compared against the manual clusters created in [19], and Pairwise F-scores [19] were calculated.

V. Result and Discussion

This study paved the way to create two types of stopwords lists, a general one and a domain-specific one. The general stopword list contains 93 words. Fig. 1 depicts the effectiveness of the clustering system with and without stopwords. Fig. 2 list downs the general stopwords. Domain-specific stopwords are listed in Fig. 3, Fig. 4, Fig. 5 and Fig. 6. Fig. 3 is the list of stopwords for the domain of International news. Fig. 4 is the same for the domain of Local news. Fig. 5 is for the Sports news, and Fig. 6 is for the Entertainment news.

Four experiments were conducted under this study. They are:

1) TF-IDF with one pass algorithm (TFIDF-OPA) clustering using the dataset with stopwords.
2) FastText with one pass algorithm (FT-OPA) clustering using the dataset with stopwords.
3) TF-IDF with one pass algorithm (TFIDF-OPA) clustering using the dataset without stopwords (stopwords removed).
4) FastText with one pass algorithm (FT-OPA) clustering with the datasets without stopwords (stopword removed.).

Table I describes the results obtained for the above 4 experiment setups. There is a significant improvement in the F-score when the dataset is used after removing stopwords with regard to that of the data set with the stopwords TFIDF-OPA increased by 2.4% and FT-OPA increased by 0.95%. This shows the impact of stopwords removal is higher in TF-IDF than that of FastText.

Tamil is a grammar-rich and highly inflected language. Even though some words are in stopwords, inflections of them are not included in the list. For example, அணி (Team – ani) is a stopword in the sports domain. It has the following inflected term: அணிகளுக்கு (for teams – Aṇikaḷukku), அணிக்கு (To the team - Aṇikku), அணியின் (Of the team- Aṇiyiṉ), அணிக்கும் (for the Team- Aṇikkum). TF-IDF fails to identify all these forms as stopwords in the clustering process. But Fasttext was able to handle this in the clustering process. Because Fasttext include representing sentences with bag-of-words and bag-of-n–grams, as well as using subword information, and sharing information across classes through a hidden representation. Further, TF-IDF considers inflected terms as different words.

Also, another challenge is some terms are written in different styles by different news providers. For example: கிரிக்கட் (Cricket -Kirikkat), கிரிக்கெட் (Cricket -Kirikket). TF-IDF considers these as two different terms.

All the following news samples reports on different arrests instances. All these news get grouped into one cluster without the stopwords removed if the clustering is performed. However, when the stopwords were removed, the system could cluster them into three different clusters based on the reason for the arrest. The first, second, and sixth reference the same instance about an arrest associated with drugs. The fourth and fifth are about an arrest related to an accident; they are

clustered together. The third news is in another cluster, which is about an arrest related to smuggling. Since the word கைது (arrested – kaitu) was in the stopword list, the stopword removal process removes it from the dataset. Therefore, it was possible to distinguish these news articles, increasing the clustering accuracy.

- 12 இலட்சம் ரூபா பெறுமதியான போதை மாத்திரைகளுடன் மூவர் கைது (12 Ilaṭcam rūpā peṟumatiyāṉa pōtai māttiraikaḷuṭaṉ mūvar kaitu) –Three arrested with drugs worth 12 lakh rupees.
- 12 லட்சம் பெறுமதியான போதை மாத்திரைகளுடன் மூவர் கைது (12 Laṭcam peṟumatiyāṉa pōtai māttiraikaḷuṭaṉ mūvar kaitu) - Three arrested with Rs 12 lakh worth of drugs.
- ஒருதொகை தங்காபரணங்களுடன் சிங்கப்பூர் பிரஜை கைது (Oru tokai taṅkāparaṇaṅkaḷuṭaṉ ciṅkappūr pirajai kaitu) - Singaporean national arrested with gold jewelry.
- மூவர் கைது (Mūvar kaitu) - Three arrested.
- 3 பேர் கைது (3 Pēr kaitu) - 3 people arrested.
- ஹெரோயின் வியாபாரத்தில் ஈடுபட்ட நபர் கைது (Herōyiṉ viyāpārattil īṭupaṭṭa napar kaitu)- The person involved in the heroin trade was arrested.

Further, the following is a set of news related to the same day. The first and the third related to a discussion on bus fare reduction, while the second is regarding a meeting between the president and the TNA parliamentarians. When they are clustered without the stopwords removal, all three are placed into one cluster. However, when classified after the stopword removal, the first and the third are allocated to the same cluster while the second is set onto another separate cluster. Since the word இன்று (today – inru) that was in the stopword list was removed in the stopword removal process, the documents similarly between first and third increased while for the second one reduced..

- பஸ் கட்டணம் குறைக்கப்படுமா ? இன்று கலந்துரையாடல் (Pas kaṭṭaṇam kuṟaikkappaṭumā? Iṉṟu kalanturaiyāṭal) - Will bus fares be reduced? Discussion today.
- தமிழ் தேசிய கூட்டமைப்பின் பாராளுமன்ற உறுப்பினர்களுக்கும் ஜனாதிபதி மைத்திரிபால சிறிசேனவுக்கும் இடையில் இன்று சந்திப்பு ஒன்று (Tamiḻ tēciya kūṭṭamaippiṉ pārāḷumaṉṟa uṟuppiṉarkaḷukkum jaṉātipati maittiripāla ciṟicēṉavukkum iṭaiyil iṉṟu cantippu oṉṟu) - Today a meeting between TNA parliamentarians and President Maithripala Sirisena.
- பஸ் கட்டணத்தைக் குறைப்பது குறித்த இறுதித்தீர்மானம் தொடர்பிலான கலந்துரையாடல் (Pas kaṭṭaṇattaik kuṟaippatu kuṟitta iṟutittīrmāṉam toṭarpilāṉa kalanturaiyāṭal)- Discussion on the final decision on reducing bus fares.

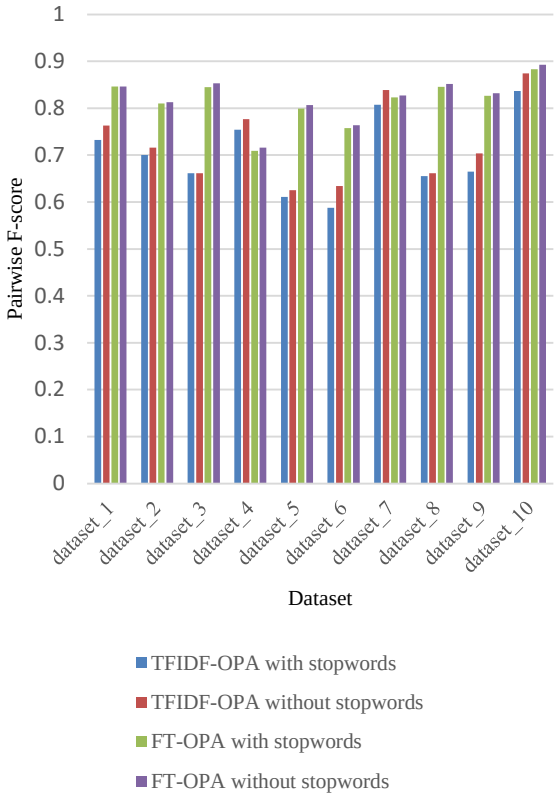


Fig. 1. Effectiveness of Clustering with and without Stopwords.

ஒரு	மற்றும்	அது	பெரும்	அனைத்து	வேண்டும்	அவர்	போல்
என்று	இல்லை	மூலம்	என்று	இன்று	மாற்றம்	போன	இங்கு
இந்த	என்று	தன்	உள்ளது	ஆண்டு	மிக	அல்ல	கூட
என	என்ற	என்ன	என்பது	ஆம்	தனது	எங்கும்	போல
இது	அந்த	என்	தான்	மீண்டும்	அவர்	இதோ	கைது
முதல்	அல்லது	ஏன்	கொண்டு	முதல்	பெற்றது	செய்து	இடையிலான
பல	வரும்	ஒரே	அதன்	என்ற	விசேட	சிறு	எதிரான / எதிராக
சில	மேலும்	யார்	நாள்	போது	ஒருவர்	மிகவும்	பலர்
இருந்து	ஆனால்	இருந்தது	பிரபல	ஊடக	இருந்து	தன	உள்ள
நாள்	வரை	பெற	தமது	கடந்த	எந்த	வந்த	காரணமாக/காரணம்
உடனான	சிறந்த	பெரும்	ஏற்பட்ட	விட	புதிய	ஆகிய	அதை
ஒன்றில்	கடந்த	தொடர்பில்	தொடர்பான	மீது			

Fig. 2. Stopwords for the Tamil Language.

தாக்குதல்	இடம்பெற்ற	தீவிரவாதிகள்	காவல்துறையினர்	பாரிய	நாட்டு
காணொளி	எச்சரிக்கை	தொடர்பான	கொல்லப்பட்டனர்.	வயது	செய்த
தலைவர்	குண்டு	நடத்தப்பட்ட	ஏற்பட்ட	பகுதி	தற்கொலை
பல்வேறு	வந்த	வகையில்	தெரிவித்துள்ளார்	கொலை	ஜனாதிபதி

Fig. 3. International News Stopwords for the Tamil Language.

அதிகரிப்பு	அமைச்சு	தெரிவித்துள்ளது	நடவடிக்கை	விசாரணை	கட்சி	திணைக்களம்	செயலாளர்
விலை	வைத்து	இடம்பெற்ற	முன்னாள்	திட்டம்	குழு	எதிர்வரும்	மோசடி
ஒன்றை	ஆர்ப்பாட்டம்	பிரதேசத்தில்	போராட்டம்	தேசிய	அதிகாரிகள்	சேர்ந்த	செய்யப்பட்டுள்ளனர்

Fig. 4. Local News Stopwords for Tamil Language.

TABLE I. STATISTICAL ANALYSIS OF OBTAINED PAIRWISE F-SCORE WITH STOPWORDS AND WITHOUT STOPWORDS

Document Representation Techniques with Clustering Algorithms	**TFIDF-OPA with stopwords**	**FT-OPA with stopwords**	**TFIDF-OPA without stopwords**	**FT-OPA without stopwords**
Mean (Average)	70.1	81.5	**72.5**	**82.0**
Median	68.2	82.5	**70.9**	**83.2**
Minimum	58.8	70.9	**62.5**	**71.6**
Maximum	83.7	88.3	**87.4**	**89.2**
Standard Deviation	7.7	4.7	**5.7**	**4.6**

In the following scenario from entertainment deomain, the first news about Deepika Ranveer Wedding Date Announcement, the second talks about the second single released in the movie “Kattrin moli,” third is about “Sarkar,” movie story released. All these sentences are clustered into one group before removing the stopword வெளியாகியுள்ளது (Released - veḷiyākiyuḷḷatu). After removing the stopword, all three news clustered into three different clusters.

- தீபிகா ரன்வீர் திருமண திகதி அறிவிப்பு ! (tīpikā raṉvīr tirumaṇa tikati aṟivippu!) - Deepika Ranveer Wedding Date Announcement!
- காற்றின் மொழி திரைப்படத்தின் 2_வது சிங்கிள் இன்று (Kāṟṟiṉ moḻi tiraippaṭattiṉ 2_vatu ciṅkiḷ iṉṟu) - second single of the “karrin moli” movie today
- சர்கார் படத்தின் கதை வெளியானது (carkār paṭattiṉ katai veḷiyāṉatu) – Sarkar film story released

அணி	வீரர்	போட்டி	ஆட்டம்	உலக
இறுதி	வெற்றி	தொடர்	சர்வதேச	

Fig. 5. Sports News Stopwords for the Tamil Language.

இயக்குனர்	நடிகை	வெளியாகியுள்ளது
நடிகர்	ரசிகர்கள்	நடிப்பில்
தற்போது	திடீர்	படங்களில்

Fig. 6. Entertainment News Stopwords for the Tamil Language.

VI. CONCLUSION

This paper presents an approach to list out the stopwords in Tamil, which is a low-resource language. So far, there is no predefined published stopword list for Tamil. The widely used technique for stopwords identification is based on term frequency. In this study, TF*IDF with threshold value is used to identify the stopwords for Tamil. The research resulted in the generation of stopword lists for general domain and domain-specific ones for local, international, sport, and entertainment domains. To evaluate its impact on Tamil NLP, it was used in document clustering using TF-IDF with one pass algorithm and FastText with the one-pass algorithm. The results revealed that the removal of stopwords at the preprocessing stage improved F-score, mean, median, and standard deviation in both the approaches.

REFERENCES

[1] M. Anandarajan, C. Hill, and T. Nolan, Cluster Analysis: Modeling Groups in Text. 2019.

[2] S. C. Satapathy, Advances in Intelligent Systems and Computing 1177 Intelligent Data Engineering and Analytics, vol. 2, no. Ficta. 2020.

[3] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020, pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

[4] C. Fox, "A Stop List for General Text," ACM SIGIR Forum, vol. 24, no. 1–2, pp. 19–21, 1989, doi: 10.1145/378881.378888.

[5] K. V. Ghag and K. Shah, "Comparative analysis of effect of stopwords removal on sentiment classification," IEEE Int. Conf. Comput. Commun. Control. IC4 2015, pp. 2–7, 2016, doi: 10.1109/IC4.2015.7375527.

[6] L. Hao and L. Hao, "Automatic identification of stop words in chinese text classification," Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 1, pp. 718–722, 2008, doi: 10.1109/CSSE.2008.829.

[7] J. K. and J. R., "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language," Int. J. Comput. Appl., vol. 150, no. 2, pp. 15–17, 2016, doi: 10.5120/ijca2016911462.

[8] H. Saif, M. Fernandez, and H. Alani, "Automatic stopword generation using contextual semantics for sentiment analysis of Twitter," CEUR Workshop Proc., vol. 1272, pp. 281–284, 2014.

[9] S. Girmaw and V. Khedkar, “Automatic Generation of Stopwords in the Amharic Text,” Int. J. Comput. Appl., vol. 180, no. 10, pp. 19–22, 2018, doi: 10.5120/ijca2018916161.

[10] I. A. El-Khair, "Effects of stop words elimination for arabic information retrieval: A comparative study," arXiv, no. December, 2017.

[11] K. Bouzoubaa, H. Baidouri, T. Loukili, and T. El Yazidi, “Arabic stop words: Towards a generalisation and standardisation,” Knowl. Manag. Innov. Adv. Econ. Anal. Solut. - Proc. 13th Int. Bus. Inf. Manag. Assoc. Conf. IBIMA 2009, vol. 3, no. November 2009, pp. 1844–1848, 2009.

[12] A. Alajmi and E. mostafa Saad, "Toward an ARABIC Stop-Words List Generation Toward an ARABIC Stop-Words List Generation," vol. 46, no. January 2012, pp. 8–13, 2018.

[13] S. V. S. Gunasekara and P. S. Haddela, "Context aware stopwords for Sinhala Text classification," 2018 Natl. Inf. Technol. Conf. NITC 2018, pp. 2–4, 2018, doi: 10.1109/NITC.2018.8550073.

[14] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," Int. Conf. Microelectron. Comput. Commun. MicroCom 2016, 2016, doi: 10.1109/MicroCom.2016.7522593.

[15] A. A. V. A. Jayaweera, Y. N. Senanayake, and P. S. Haddela, "Dynamic Stopword Removal for Sinhala Language," 2019 Natl. Inf. Technol. Conf. NITC 2019, pp. 8–10, 2019, doi: 10.1109/NITC48475.2019.9114476.

[16] Y. Wijeratne and N. de Silva, "Sinhala Language Corpora and Stopwords from a Decade of Sri Lankan Facebook," arXiv, 2020, doi: 10.2139/ssrn.3650976.

[17] S. Sarica and J. Luo, "Stopwords in Technical Language Processing," arXiv, no. June, 2020.

[18] R. M. Rakholia, and J. R. Saini, "A Rule-Based Approach to Identify Stop Words for Gujarati Language," In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, 2017, pp. 797-806, Springer, Singapore.

[19] M. S. Faathima Fayaza and S. Ranathunga, "Tamil News Clustering Using Word Embeddings," MERCon 2020 - 6th Int. Multidiscip. Moratuwa Eng. Res. Conf. Proc., pp. 277–282, 2020, doi: 10.1109/MERCon50084.2020.9185282.

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971.

Improving Chi-Square Feature Selection using a Bernoulli Model for Multi-label Classification of Indonesian-Translated Hadith

Fahmi Salman Nurfikri, Adiwijaya
School of Computing
Telkom University
Bandung, Indonesia

***Abstract*—Hadith is the foundational knowledge in Islam that must be studied and practiced by Muslims. In the Hadith, several types of teachings are beneficial to Muslims and all of mankind. Some Hadith serve as advice, while others contain prohibitions that Muslims should adhere to. There are yet others that do not belong to these categories and serve only as information. This study focuses on increasing the performance of Chi-Square feature selection to obtain relevant features for multilabel classification of Indonesian-translated Bukhari Hadith data. This study proposes a Chi-Square-based Bernoulli model to improve Chi-Square feature selection which is appropriate for short-text data such as Hadith. The findings of this study show that the proposed method can select relevant features based on data classes; thereby improving Hadith classification performance with an error value of 9.38% compared to that (9.91%) obtained using the basic Chi-Square feature selection.**

Keywords*—Bernoulli model; Chi-Square; feature selection; hadith classification

I. Introduction

Hadith is an important textual source of law, tradition, and teachings in the Islamic world [1]. Following the advancement in technology, several research studies have been conducted on Hadith including the application of natural language processing to classify Hadith based on its content. Hadith classification is a method of categorizing Hadith based on its content [2]; the structure of an Hadith is different from other textual representations. A Hadith comprises three components: Matn, Isnad, and Taraf [1]. Matn is the central text, Isnad is the chain of narrators, and Taraf is the beginning phrase(s) of the Hadith. In addition, some Hadith, for example, the Hadith provided in the book of Sahih Al-Bukhari, belong to more than one label (i.e., the data is multilabel) [1], and therefore, a multi-label classification approach is required.

Multilabel classification is a type of supervised learning where a classification algorithm needs to learn from datasets and classify data into multiple classes; in single-label classification, data can only be classified into one class. For example, a movie is multilabel data as it can simultaneously be categorized as action, crime, and/or thriller [3]. However, the generality of multilabel data makes it more difficult to classify it compared to other data.

In text classification, the features are terms or words contained in the text. A document or textual data contain a considerable number of words that can cause high computational complexity and decrease accuracy as irrelevant features may be considered during the classification [4]. To overcome this limitaion, feature reduction must be applied. One method of feature reduction is feature selection [5] wherein only relevant features to be used for classification are selected. An example of a feature selection method that has been proved to produce good results is the Chi-square [6]. However, one of the limitations of the Chi-Square is that all measured participants must be independent, i.e., one individual cannot fit into more than one class or a single label. Further, its other disadvantage is that the data must have multinomial data frequency. This is a limitation in our case because the text in the Hadith is short. The Bernoulli model has been proved to work effectively with few features [7] and therefore, it is worth exploring.

II. Related Work

A. Hadith Classification

A considerable amount of research has been conducted on Indonesian-translated Hadith, with Faraby et al. [8] being a notable work in this area. Their study categorized Sahih Al-Bukhari Hadith data into three classes: advice, prohibition, and information. The study compared the classification results using artificial neural networks (ANNs) and support vector machine (SVM), and they applied term frequency-inverse document frequency. The results of the study showed that the SVM method performed better than the ANN method, with an F1-Score of 88% to 85%.

Furthermore, Afianto et al. [9] used a dataset similar to Faraby et al. [8]; however, they used random forest as the classification method. The study obtained an F1-Score of 90%, which is better than that of previous study [8]. The most significant process in this research study was the determination of the bootstrap method used where the bootstrap sample was set to 100.

Bakar et al. [10] conducted multi- and single-label Hadith classification using 1064 data points. The multilabel classification comprised three classes (advice, prohibition, and information), while the single-label classification comprised five (faith, knowledge, ablution, prayer, and prayer times). The

study used information gain (IG) as the selection feature technique and the backpropagation neural network (BNN) as the classifier. The study obtained an F1-score of 65.275% for single-label classification, while the Hamming Loss value for multilabel classification was 0.1158. Hence, using IG as the selection feature technique significantly improved the classification performance of the model.

B. Multi-Label Classification

Classification of multilabel data can be problem as such data can be categorized into two or more classes. Research on multilabel classification is motivated by medical diagnosis and text categorization problems. Two approaches can be used for multilabel classification: problem transformation and algorithm adaptation [3].

The problem transformation approach solves multilabel problems by transforming multilabel data into single-label data, while the algorithm adaptation approach classifies multilabel data using algorithms designed for multilabel classification. Multilabel classification using the problem transformation approach achieved better performance than that using the algorithm adaptation approach [3].

Several studies on multilabel classification have been conducted; however, only a few such as those of Bakar et al. [10], Mediamer et al. [11], and Kabi et al. [12] focused on the multilabel classification of Hadith data. Liu et al. [13] conducted multilabel classification using a correlation function; this was effective for overfitted and noisy data. However, such methods are not designed to obtain optimal parameters, and therefore, this can affect classification performance. Soleimani et al. [14] used semi-supervised learning methods and latent Dirichlet allocation for learning topic classes, and used a small number of labeled training data for multilabel classification. However, this method also had a limitation; it had a high time complexity given the large amount of data used in the study. Huang et al. [15] combined a feature selection technique and a classifier for multilabel classification thereby providing an advantage for selecting relevant features of each label and training the classifier to increase the effectiveness of the model. However, the drawback of this method was that it requied a high computational time to obtain optimal parameters.

C. Feature Selection

A problem with text classification is that textual data contain a considerable number of words that can cause high computational complexity and decrease the accuracy of classification results [16], [6], [17]. One approach to tackle this problem is applying feature selection to the data. Yang et al. [6] investigated document frequency (DF), IG, Chi-Square, mutual information (MI), and term strength as feature selection methods for the Reuters corpus. The experiment found that IG and Chi-Square were the most effective feature selection methods as they could remove 98% of irrelevant features without compromising classification performance. However, Chi-Square and IG showed a limitation in that they incurred high computational cost, whereas DF had the lowest computational cost but strong correlation with Chi-Square and IG.

Forman [18] presented an extensive comparative study of feature selection metrics for text classification of high-dimensional data focusing on SVM for the two-class problems. Forman found that the new feature selection metric—Bi-normal separation—achieved better performance compared to other feature selection methods. Xu et al. [19] compared DF, IG, MI, and pointwise MI and found that MI and IG achieved the same performance. Another study used a Bernoulli model as the feature selection method [7] and found that it worked best for documents with short texts, while a multinomial model was better for handling documents with long texts.

III. Design Process

The steps followed by the proposed method (Fig. 1) are described in this section.

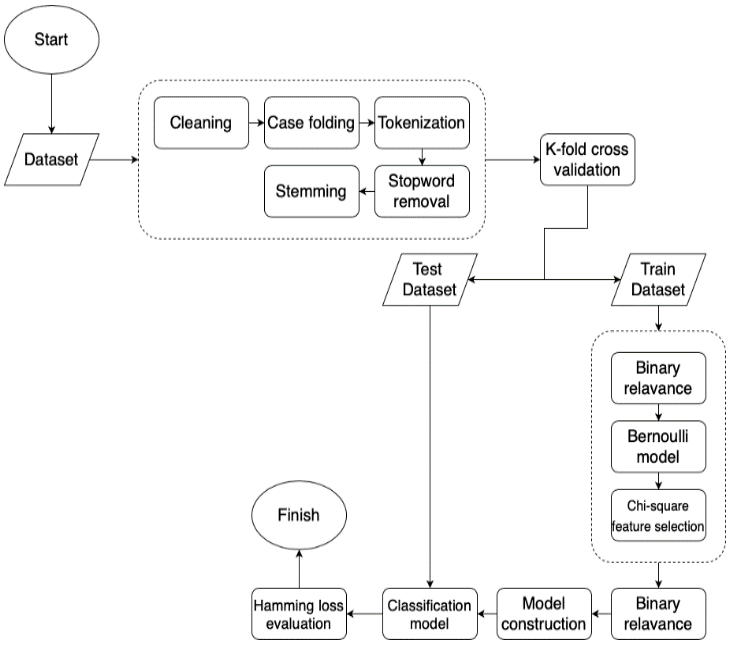


Fig. 1. System Design.

The initial stage involves collecting text-based data from the Hadith in Bahasa Indonesia from the book of Hadith Sahih Al-Bukhari; the book consists of 1066 data points and is divided into three class labels (Advice, Prohibition, and Information). An example of the data representation is listed in Table I.

TABLE I. Multi-Label Data Representation of Indonesia-Translated Hadith

Data	Class
'*Janganlah kalian berdusta terhadapku (atas namaku), karena barangsiapa berdusta atas namaku dia akan masuk neraka*.'	Prohibition Information
'*Kami pernah shalat Maghrib bersama Nabi ketika matahari sudah tenggelam tidak terlihat*.'	Information

TABLE II. COMBINATION OF CLASSES IN THE DATASET

No.	Advice	Prohibition	Information	Count
1	0	0	0	0
2	0	0	1	777
3	0	1	0	6
4	0	1	1	53
5	1	0	0	10
6	1	0	1	181
7	1	1	0	5
8	1	1	1	34

The class combination in the dataset is listed in Table II.

The dataset consisted of 230 Advice, 98 Prohibition, and 1045 Information data points. Based on these data points, it can be seen that the number of data points in the Advice and Prohibited classes is very small compared to that in the Information class; hence, the data is unbalanced. This can be a problem because unbalanced data can lead to less optimum classification results.

The first step to handle unbalanced data is preprocessing. This study used cleaning, case folding, tokenization, stopword removal, and stemming as the preprocessing steps to eliminate some sentences that are not used in the classification process. An Indonesian stopword list from a study conducted by Tala [20] was used and modified to match the Hadith dataset in this study. In addition, the Nazief–Andriani stemming algorithm [21] was also used. Next, the dataset was split into training data and test data using 5-fold cross-validation to make all observations in the dataset are nicely distributed in a way that the data are not biased.

Feature extraction was performed using the bag-of-words representation. In this study, a term frequency method was used to extract the feature. This method counts each word in the vocabulary list obtained from the training dataset for each data point.

Two general approaches for multi-label classification are problem transformation and algorithm adaptation. Problem transformation converts multi-labeled data into single-labeled data, while the algorithm adaptation uses algorithms specifically adapted to handle multilabel classification. Based on the research conducted by Irsan et al. [3], the problem transformation approach achieved better performance results compared to algorithm adaptation.

Binary relevance uses problem transformation approach [22] [23]. Binary relevance creates a number of k datasets (k = |L|, the total number of classes). Each dataset has the same instance as the original data; however, each dataset contains only one class. Using this method, class data representation must first be changed into one-hot encoding.

The next step involved duplicating the dataset of q, where q is the number of classes in the training data so that each dataset only has 2 classes, namely. 0 and 1.

Next, the extracted features are selected using Chi square. In this study, a Bernoulli model was used for Chi square feature selection. This model checks for the presence or absence of a word, and therefore, it only has two possible outcomes: yes or no. The Bernoulli model was used because every Hadith contains an average of 20 words, and hence, a small number of features is the type of data that the Bernoulli model can process effectively [7]. The algorithm of the Chi square Bernoulli model is presented in Algorithm 1 below.

Algorithm 1. Chi-Square Bernoulli model algorithm

```
function Bernoulli-Chi-Square-FS()
    Input: Array of attribute and its class C
    Output: Array of Chi value for each class
Initialize
arrayofchivalue (array)
arrayofclasschivalue (array)
Begin
Change class data representation into one-hot encoding
Break the class into k classes
for each c in class do
  for each a in attributes do
    for each row in a do
      if row >= 1 then
        row ← 1
      else
        row ← 0
      end if
    end for
    Calculate ChiSquare(a, c_i) and append it to arrayofchivalue
  end for
  Sort descending arrayofchivalue
  Get top-n attributes and append it to arrayofclasschivalue
end for
return arrayofclasschivalue
```

Based on Algorithm 1, each feature row is transformed into 0 and 1. In this model, words with three occurrences are the same as words with only one occurrence.

Then, feature selection is performed for each class. Feature selection is the process of selecting a subset of relevant features for training a classification model; it is used to select relevant features that will be included in the classification process, thereby efficiently and effectively improving the process [16]. Chi square feature selection is used in this study [6]; it is expressed by

$$X^2(t,c) = \frac{N*(AD-CB)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (1)$$

A Chi square statistic measures the lack of independence between term t and class c, and it can be compared to Chi square distributions with one degree of freedom to evaluate extremeness [6], where A is the number of times t and c occur; B is the number of times t occurs without c; C is the number of times c occurs without t; D is the number of times neither t nor c occurs; and N is the total number of documents.

The output of Chi square is the Chi value, which is between a feature and a class; the greater the Chi value, the greater is the relationship between the feature and the class. Each feature is calculated for each class. Once the Chi values are obtained, they are sorted in the descending order for each class, where greater the Chi value, the greater is the effect of a feature on a class [8]. Finally, the top-n features are obtained and used as inputs for the classifier.

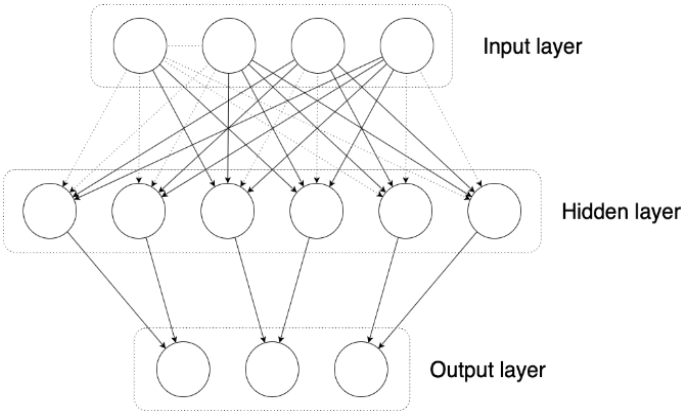


Fig. 2. Structure of the Neural Network.

The classifier is trained using the BNN. This algorithm was selected because it can process a wide variety of features to obtain a high classification performance [24], [4], [1], [10]. The classifier was trained using a modified BNN with the binary relevance approach, and therefore, the training process was conducted three times, which is equal to the number of classes. The selected features were used to train the classifier. Each class had different input data because of the different selected features. Therefore, in this the BNN was modified to tackle this problem, as shown in Fig. 2.

Fig. 2 shows that two main lines connect the input layer and the hidden layer, i.e., the bold and dotted lines. The bold line indicates that input and hidden neurons are connected, while the dotted line indicates that the input does not pass the feature selection for the class; however, it can pass the feature selection for other classes. Finally, the evaluation results of the classifier are expressed in terms of Hamming Loss. The Hamming loss is used because this method is appropriate for multilabel classification and assigns equal weight to each label [25].

IV. Results and Discussion

A. Effect of the Bernoulli Model on Chi-Square Feature Selection

The performance of the Bernoulli model was compared by varying the number of dimensions from 10% to 100%. This allows determining if the use of feature selection can help improve the performance of the classifier and to obtain the best dimensionality for optimal classification performance. The BNN input nodes are equal to the dimension of the document vector.

The results of the proposed model are compared with those of the typical BNN and Chi-square-based BNN feature selection models. The results are listed in Tables III and IV.

Based on the results listed in Tables III and IV, the proposed method achieves the best average result of 0.0938, while the CSBNN produced the best average result of 0.0991. A comparison chart of the three methods is shown in Fig. 3.

TABLE III. Hamming Loss Result of Classification using Chi-Square Bernoulli Model and Backpropagation Neural Network (BCSBNN)

Number of dimension	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
10%	0.1106	0.1064	0.0955	0.0955	0.0814	0.0979
20%	0.1308	0.1221	0.0955	0.0939	0.0861	0.1057
30%	0.1324	0.1252	0.0970	0.0908	0.0829	0.1057
40%	0.1121	0.1142	0.0970	0.1033	0.0829	0.1019
50%	0.1153	0.1111	0.0939	0.1049	0.0798	0.1010
60%	0.1075	0.0986	0.0892	0.0955	0.0782	0.0938
70%	0.1168	0.1095	0.1033	0.1002	0.0923	0.1044
80%	0.1293	0.1158	0.0939	0.1017	0.0876	0.1057
90%	0.1184	0.1064	0.1064	0.1017	0.0923	0.1051
100%	0.1137	0.1127	0.0923	0.0986	0.0814	0.0997

TABLE IV. Hamming Loss Result of Classification using Chi-Square and Backpropagation Neural Network (CSBNN)

Number of dimension	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
10%	0.1106	0.1111	0.0970	0.0986	0.0782	0.0991
20%	0.1231	0.1142	0.1049	0.1017	0.0782	0.1044
30%	0.1075	0.1299	0.1017	0.1002	0.0782	0.1035
40%	0.1199	0.1189	0.0955	0.0955	0.0829	0.1025
50%	0.1184	0.1127	0.0970	0.0845	0.0892	0.1004
60%	0.1168	0.1174	0.0955	0.0892	0.0782	0.0994
70%	0.1215	0.1142	0.0939	0.1033	0.0845	0.1035
80%	0.1199	0.1299	0.0939	0.0986	0.1158	0.1116
90%	0.1246	0.1189	0.1064	0.0970	0.0845	0.1063
100%	0.1153	0.1111	0.0939	0.1049	0.0798	0.1010

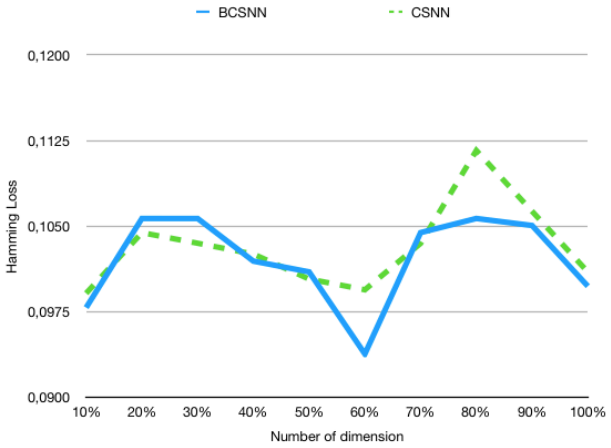


Fig. 3. Categorization Performance of BCSBNN and CSBNN according to the Number of Dimensions

As shown in Fig. 3, the performance of BCSBNN and CSBNN are not considerably different; however, on average, BCSBNN outperformed CSBNN. This is because Bernoulli distribution can select relevant features as inputs for the BNN and the Bernoulli distribution only has two possibilities (yes or no). For example, consider the word "hendak;" in the Bernoulli distribution, one word is enough to represent the word in the class to produce the probability p (hendak = 'yes' | class) and p (hendak = 'no' | class). Meanwhile, by using multinomials, the number of occurrences of each word has its respective probabilities such as p (want = 0 | class) and p (want = 1 | class). Therefore, this decreases the occurrence probability of each word.

Further, Fig. 3 shows that when using 60% of the data dimensions, the smallest Hamming Loss value is obtained. This is because the features used as inputs for the classification models match the test data. However, this can change depending on the data used. In addition, using feature selection produced better results than when the whole data (using 100% dimension) was used. This is because the feature selection technique removed irrelevant words/features from the dataset used in training and testing the classification model. However, it is necessary to determine the best parameters for choosing the number of feature dimensions to use.

B. Comparison of the Modified Backpropagation Neural Network and the Typical Backpropagation Neural Network Classification Performance

Table V lists the classification performance of the modified BNN using the binary relevance approach compared to that of the typical BNN. The performance of the networks was first compared using the Chi-Square Bernoulli Model (BCS) and then using the Chi-Square (CS) test.

TABLE V. PERFORMANCE COMPARISON BETWEEN THE MODIFIED BACKPROPAGATION NEURAL NETWORK (BINARY RELEVANCE) AND THE TYPICAL BACKPROPAGATION NEURAL NETWORK

Number of dimensions	Modified Neural Network		Original Neural Network	
	BCS	**CS**	**BCS**	**CS**
10%	**0.0979**	0.0991	0.1126	0.1129
20%	0.1057	**0.1044**	0.1135	0.1253
30%	0.1057	**0.1035**	0.1123	0.1263
40%	**0.1019**	0.1025	0.1110	0.1151
50%	0.1010	**0.1004**	0.1094	0.1119
60%	**0.0938**	0.0994	0.1094	0.1135
70%	0.1044	**0.1035**	0.1204	0.1126
80%	**0.1057**	0.1116	0.1088	0.1132
90%	**0.1051**	0.1063	0.1132	0.1163
100%	**0.0997**	0.1010	0.1094	0.1132

As shown in Table V, the modified BNN outperformed the typical BNN. This is because, in the typical BNN, the classifier must remember more patterns in the class, whereas, in the binary relevance, one classifier is focused on only remembering one pattern. For example, for the typical BNN, the classifier must remember eight different class patterns and a combination of unbalanced data, as listed in Table II. Meanwhile, the binary relevance method only focuses on each class, i.e., Advice, Prohibition, or Information.

Further, the BNN following the binary relevance approach requires less computational complexity than the typical BNN because the number of neurons connected in the former were reduced, thereby reducing the matrix computation. However, using binary relevance slightly increased time complexity because the classifier had to learn as many class patterns as possible. This can be a problem if the number of classes to be trained becomes very large.

C. Model Prediction

Samples of the classification results are listed in Table VI.

The conducted experiments and results listed in Table VI indicate that there are three types of predictions: correct prediction, partially correct prediction, and wrong prediction. Further, in the "correct prediction" row, the predictions and targets achieved the same results because the words that appear in the Hadith were relevant, and thus, only correct results were obtained.

In the "partially correct prediction" row, the system predicted only the information class, while the target classes were Advice and Information. This is because the number of Advice data points was so small that the probability of the system in retrieving the Advice class was trivial compared to that in retrieving the Information class, which has a very large number of data points. In the future, further processing of unbalanced data must be performed.

TABLE VI. SAMPLE PREDICTION

	Data	**Predicted**	**Target**
Correct prediction	*'Jika salah seorang dari kalian meludah maka janganlah ia membuangnya kearah depan atau sebelah kanannya, tetapi hendaklah ia lakukan kearah kirinya atau di bawah kaki (kirinya).'*	Advice Prohibition Information	Advice Prohibition Information
Partially correct prediction	*'Jika salah seorang dari kalian mengantuk saat salat, hendaklah tidur (dahulu) hingga ia mengetahui apa yang ia baca.'*	Information	Advice Information
Wrong prediction	*'Janganlah salah seorang dari kalian sengaja salat ketika matahari sedang terbit atau ketika saat terbenam..*	Advice Information	Prohibition

TABLE VII. SAMPLE OF ZERO PREDICTION

Data	Predicted	Target
'Luruskanlah shaf, sesungguhnya aku dapat melihat kalian dari balik punggungku.'	-	Advice Information

In the "wrong prediction" row in Table VI, which shows a sample of data that has been manually labeled before, many of the datasets used are still ambiguous when viewed in a meaningful way per word. For example, the data can be categorized into the Advice class as well. Hence, further validation of the dataset needs to be performed to achieve better performance.

A limitation of the binary relevance approach is the occurrence of zero predictions or data that cannot be categorized into any category. This happens because by following the binary relevance approach, each model is independent and so are the classes. Examples of these phenomena are summarized in Table VII.

In this research, 16 datasets could not be classified when adopting the binary relevance approach, while only 7 datasets from the overall 214 test dataset could not be classified when adopting the algorithm adaptation approach. This is attributed to unbalanced data. For example, in this study, there are only 98 data points for the Prohibition class with a total of 1066 data points, where the ratio of the Prohibition class and non-Prohibition class is 1:10, thereby making the classifier classify data as non-Prohibition class and so on for the other classes. By adopting the algorithm adaptation approach, a combination of classes connects the classes to reduce the possibility of zero predictions. Further, the use of feature selection may have an effect on the occurrence of zero predictions, as irrelevant words in documents are not selected, which causes a lack of features to sufficiently represent data. This is because the binary relevance approach entails that each model be independent and that no dependence exists among classes.

V. CONCLUSION

This research proposed a Chi-Square Bernoulli Model and a BNN model to classify Hadith into specific categories. The Bernoulli model was used as the feature selection method and was found to improve the classification performance, achieving the best average Hamming Loss result of 9.38%. This is because, in the Bernoulli distribution, one word is sufficient to represent the total number of occurrences of the word in a class, and therefore, the Bernoulli distribution can choose the relevant features as inputs for the BNN.

Furthermore, the binary relevance approach outperformed the algorithm adaptation approach. This is because when using algorithm adaptation, the classifier must remember most of the patterns in a class, whereas, in problem transformation (binary relevance), the classifier is only focused on remembering one pattern in a class.

For further research in this regard, more attention should be given to processing unbalanced data. Further, the future work should explore other methods such as the recurrent neural network, which works with data sequences or similar methods and determines their effectiveness in classifying Hadith data.

REFERENCES

[1] M. A. Saloot, N. Idris, R. Mahmud, S. Jaafar, D. Thorleuchter and A. Gani, "Hadith Data Mining and Classification: A Comparative Analysis," in Artif Intell Rev 46, pp. 113–128, 2016.

[2] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," Applied Computational Intelligence and Soft Computing, 2018.

[3] I. C. Irsan and M. L. Khodra, "Hierarchical Multilabel Classification for Indonesian News Articles," in 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016.

[4] F. Harrag, E. El-Qawasmah and A. M. S.Al-Salman, "Stemming as a Feature Reduction Technique for Arabic Text Categorization," in 10th Interna- tional Symposium on Programming and Systems, 2011.

[5] M. D. Purbolaksono, F. D. Reskyadita, Adiwijaya, A. A. Suryani and A. F. Huda, "Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature," International Journal on Advance Science, Engineering and Information Technology, vol 10, pp. 234-238, 2020.

[6] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

[7] C. D. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press, 2009.

[8] S. A. Faraby, E. R. R. Jasin, A. Kusumaningrum and Adiwijaya, "Classification of Hadith into Positive Suggestion, Negative Suggestion, and Information," Journal of Physics: Conference Serie, 2018.

[9] M. F. Afianto, Adiwijaya and S. A. Faraby, "Text Categorization on Hadith Sahih Al-Bukhari using Random Forest," Journal of Physics: Conference Series, 2018.

[10] M. Y. A. Bakar, Adiwijaya and S. A. Faraby, "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language translation) using Information Gain and Backpropagation Neural Network," in International Conference on Asian Language Processing (IALP), 2018.

[11] G. Mediamer, Adiwijaya and S. A. Faraby, "Development of Rule-based Feature Extraction in Multi-label Text Classification," J. Adv. Sci. Eng. Inf. Technol, 9(4), 2019.

[12] M. N. Al-Kabi, G. Kanaan, R. Al-Shalabi, S. I. Al-Sinjilawi and R. S. Al-Mustafa, "Al-hadith text classifier," Journal of Applied Sciences, vol. 5, no. 3, pp. 584-587, 2005.

[13] H. Liu, X. Li and a. S. Zhang, "Learning Instance Correlation Functions for Multilabel Classification for Multilabel Learning," IEEE TRANSACTIONS ON CYBERNETICS, pp. 2168-2267, 2016.

[14] H. Soleimani and D. J. Miller, "Semisupervised, Multilabel, Multi-Instance Learning for Structured Data," Neural Computation, vol. 29, p. 150, 2017.

[15] J. Huang, G. Li, Q. Huang and X. Wu, "Joint Feature Selection and Classification for Multilabel Learning," IEEE TRANSACTIONS ON CYBERNETICS, pp. 2168-2267, 2017.

[16] F. S. Nurfikri, M. S. Mubarok and Adiwijaya, "News Topic Classification using Mutual Information and Bayesian Network," in 6th International Conference on Information and Communication Technology (ICoICT), 2018.

[17] H. Aydadenta and Adiwijaya, "A Clustering Approach for Feature Selection on the Microarray Data Classification using Random Forest," Journal of Computer Science, 2016.

[18] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research 3, pp. 1289-1305, 2003.

[19] Y. Xu, G. Jones, J. Li, B. Wang and C. M. Sun, "A Study on Mutual Information-Based Feature Selection for Text Categorization," Journal of Computational Information Systems, vol. 3, no. 3, pp. 1007-1012, 2007.

[20] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," in Inst. Log. Lang. Comput. Univ. Van Amst. Neth., 2003.

[21] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi and H. E. Williams, "Stemming Indonesian: A Confix-Stripping Approach," in ACM Transactions on Asian Language Information Processing (TALIP), 2007.

[22] M.-L. Zhang, Y.-K. Li, X.-Y. Liu and X. Geng, "Binary Relevance for Multi-Label Learning: An Overview," Frontiers of Computer Science, 2018.

[23] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," IEEE Transactions on Knowledge and Data Engineering, 2014.

[24] F. Harrag and E. El-Qawasmah, "Neural Network for Arabic Text Classification," in Second International Conference on the Applications of Digital Information and Web Technologies, 2009.

[25] M. S. Sorower, "A Literature Survey on Algorithms for Multi-label Learning," in Oregon State University, Corvallis , 2010.

Transfer Learning-based One Versus Rest Classifier for Multiclass Multi-Label Ophthalmological Disease Prediction

Akanksha Bali, Vibhakar Mansotra
Department of Computer Science and IT
University of Jammu
Jammu, India

***Abstract*—The main objective of this paper is to propose transfer learning technique for multiclass multilabel opthalmological diseases prediction in fundus images by using the one versus rest strategy. The proposed transfer learning-based techniques to detect eight categories (seven diseases and one normal class) are Normal, Diabetic retinopathy, Cataract, Glaucoma, Age-related macular degeneration, Myopia, Hypertension and Other abnormalities in fundus images collected and augmented from Ocular Disease Intelligent Recognition (ODIR) dataset. To increase the data set, no differentiation between left and right eye images has been done and these images were used on VGG-16 CNN network to binary classify each disease separately and trained 8 separate models using one versus rest strategy to identify these 7 diseases plus normal eyes. In this paper, various results has been showcased such as accuracy of each organ and accuracy of the overall model compared to benchmark papers. Base line accuracy have increased from 89% to almost 91% and also proposed model has improved the performance of identifying disease drastically prediction of glaucoma has increased from 54% to 91%, normal images prediction has increased from 40% to 85.28% and other diseases prediction has increased from 44% to 88%. Out of 8 categories prediction, proposed model prediction rate has improved in 6 diseases by using proposed transfer learning technique vgg16 and eight different one versus classifier classification algorithms.**

***Keywords*—Fundus images; one versus rest strategy; transfer learning; VGG-16; augmentation**

I. Introduction

The motivation of this article comes from the fact that According to WHO at present, there are 2.2 billion people around the globe suffering with visual disability out of 2.2 billion at least a billion people could have been treated from visual impairment [1]. In the 21st century, eye blindness became normal because of high exposure towards electronic gadgets such as television, laptops etc. from early ages, though most of the eye diseases could be cured if detected in starting stages of the disease. Eyes are organs of the visual system, they capture light rays and regulate intensity using diaphragm and form an image using a lens. Eyes forward the captured image to the brain via optic nerve by converting them to electro-chemical impulses, any disturbances in the above process creates visual impairment or eye disorder. The study of eye diseases and disorders to diagnose is called Ophthalmology.

The primary sources for the cause of eye blindness are 1.exposure towards electronic gadgets 2.Lack of accessibility for medical facilities especially in developing and undeveloped countries 3. People in rural areas have a higher rate compared to their counterparts living in city 4. Aging people [2] and Indigenous people (tribes) 5. Accidents like facial fractures [3] e.t.c. The most common eye diseases occurring in day to day life are due to Diabetic Retinopathy, Glaucoma, Cataract, AMD, Hypertension, and Myopia. Some of these won't result in vision impairment but cannot be neglected from detection and treatment.

Diabetic retinopathy, a common disease for blindness in the age group of 20 to 70, initial days of this disease, diabetes patients of type -1 and 60 percent of type-2 diabetes suffered from retinopathy [4]. Glaucoma is a type of eye diseases having a common feature in cupping and atrophy of optic nerve head, visual field loss; often increase in intraocular pressure [5]. In 1990, 37 million people were estimated to be blind and 40 percent of them suffered from cataract [6] and it can be corrected with surgery but lack of facilities for treating cataract is a rising concern in developing countries and undeveloped countries.

Age related macular and degeneration (AMD) [7] is a natural thing but the numbers of these cases are increasing day by day due to the sharp rise in mortality rate because of development of medical facilities [8] and stable governments. Hypertension affects the retina significantly and the study of retaining provides valuable information to treat hypertension [9]. Myopia (short sightedness) considered as disorder, it can be corrected with glasses, contact lens and surgery (Lasik treatment), though it is less threatening but number of people suffering with this disorder has taken a step curve especially in children [10].

The state of the art solutions for retinopathy based on the classifications of these diseases are tedious tasks with the advancement of computers and computing techniques; various methods are proposed for the classification of objects. The most common types of classification are 1. Single labelled classification generates yes or no situations [11], such as person is suffering from eye disease enough to understand the

person is suffering but not to understand the reasons for suffering. 2. Multi labelled classification though it is computationally expensive, but it provides better intuitions and further depth analysis and it identifies more objects.

Machine learning algorithms have significantly done well in the field of image classifications, segmentation and enhancement techniques. Machine learning algorithms have generated better results initially but failed to learn more features with the increase of the dataset. Neural networks, a part of machine learning algorithms, have done significantly well with the data and deep learning algorithms that came into picture generated more features with increase of data. Deep learning algorithms became a common norm for any image classification, segmentation tasks.

The two common ways to represent eyes in the form of images are Fundus Photography, capturing images at the fundus. The main areas covered at fundus photography are the central and peripheral retina, optic disc and macula. The second way is Optical Coherence tomography (OCT), a technique to capture 2d, 3d and micrometer resolution images with low coherence light.

The paper findings as discussed in the upcoming section:

1) This paper focuses on Fundus images to classifyNormal, Diabetic Retinopathy, Glaucoma, Cataract, AMD, Hypertension, Myopia and other diseases using multi labelled dataset ODIR 2019 consisting of 5000 fundus images of patients of both eyes.

2) The primary reason to work on these datasets is its multi-disease data sets as most of the datasets encountered are mainly focused on one ophthalmic disease.

3) The paper carefully analyzes the machine learning, data variance and generative n probabilistic aspects of retinopathy and tries to devise a cross entropy modelling system for multi label classification.

4) It also formulates the utility of transfer learning, into the diverse and difficult field of retinopathy.

The rest of the paper, structured as: section 2, literature review briefly describes various research proposed in eye organ segmentation using both machine learning and deep learning algorithms, section 3 explains the methodology part which comprises subsection 1) data collection. 2) preprocessing and augmentation 3) Training, Section 4 explains the results and analysis which comprise subsection 1) performance metrics, section 5 describes the discussion and section 6 concludes the paper.

II. Literature Review

Deep learning and Machine Learning became solutions for computer vision problems such as image enhancement, segmentation and classification, especially in biomedical imaging. Many researchers have proposed various approaches for the classification of ophthalmological diseases.

J. Liu et al. [12], proposed an SVM based classification approach to classify myopia with an accuracy of 87% on test data from the Singapore Eye Research Institute. T.V. Phan et al. [13], proposed an SVM and random forest based method to classify AMD. V. Gulshan et al. [14], uses a deep convolutional neural network to identify diabetic retinopathy and DME in fundus images. H. Pratt et al. [15], proposed a CNN and data augmentation technique for the classification of micro-aneurysms, exudate and haemorrhages on the retina. The proposed architecture achieves sensitivity and accuracy around 95% and 75% on 5000 validation images.

J. Y. Choi et al. [16], used matConvNet for automatic detection of multiple retinal diseases on the STARE database, consisting of nine eye diseases. Optimal results were obtained by random forest transfer learning based VGG19. P. M. Burlina et al. [17], proposed a DCNN for the classification of AMD, this model compared with a pre trained DCNN by performing transfer learning. Y. Chai et al. [18], proposed a method to combine deep learning models with domain knowledge for automatic glaucoma detection on fundus images. The proposed model outperformed AlexNet, VGG16, InceptionV3 in accuracy, sensitivity and specificity. F. Grassmann et al. [19], utilized various convolutional neural networks to classify nine types of eye disease due to age, three types of AMD and one ungradable image, to classify these thirteen classes, ensembling has been done over six different neural network architectures.

M. N. Bajwa et al. [25], proposed a framework containing two stages, the first stage uses a CNN to localize and extract the optical disc from the retinal fundus image and the other one uses deep convolutional neural network for classifying disc extracted in the first stage. Due to the lack of original ground truth images, they proposed rules for generation of semi automatic ground truth images. They achieved a 2.7% improvement to the previously produced results on the ORIGA dataset. S. Keel et al. [26],proposed Inception V3 architecture for classification and severity possibility threshold on neovascular age-related macular degeneration. V. Das et al. [27], proposed CNN based classification detection techniques for DME and AMD up to two stages. The evaluation of the proposed method is performed on the OCT dataset and achieves a decent score of sensitivity, specificity and accuracy around 99.6%, 99.87% and 99.6% on test data. T. Li et al. [28], provided a new dataset of 13673 fundus images from 9598 patients for diabetic retinopathy and these images were classified into 6 types based on quality and DR level.

Y. Peng et al. [29], proposed DeepSeeNet architecture to measure the score of severity in the range 0 to 5 for the age related eye diseases study. T. Pratap et al. [30], used a pre trained CNN architecture for transfer learning to extract features for classifying levels of cataract and these features were classified using SVM. In this paper [31], M. S. Alabshihy et al. used direct technique such as problem transformation, multilabel cad system, segmentation, MSVM on dataset named as DiaretDB having two classes named as DR and hypertension and achieved an overall accuracy of 96.1%. Md. T. Islam et al. [32], proposed a classification model for eight ocular diseases using contrast limited adoptive histogram equalization as a pre-procesing step and CNN has used for feature extraction.

T. Nazir et al. [33], proposed a deep learning approach for segmentation of diabetic retinopathy, diabetic macular edema

(DME) and glaucoma using a fast region based convolutional neural network to localize and fuzzy k means to segment. A multi task loss was used as a loss for CNN. Intersection over union, mean average precision and dice coefficient as evaluation metrics and achieved mean average precision of 0.94. X. Pan et al. [34], compared DenseNet, Resnet50 and VGG16 to automatic classification and detection the four kinds of lesions of diabetic retinopathy such as non-perfusion regions, microaneurysms, leakages, and laser scars in fundus fluorescein angiography images. Sensitivity, specificity and region of curve were used as evaluation metrics. M. Aamir et al. [35], proposed a two phased CNN based architecture for classification, one for glaucoma detection and other for rating glaucoma in different scales like advanced, moderate, early, on fundus images, and adaptive thresholding was done before applying CNN. Sensitivity, specificity, accuracy and precision were used as metrics for evaluation. C. G. Gonzalo et al. [36], proposed a CNN ensembling methods to identify AMD and diabetic retinopathy in color fundus images. Inputs for CNN are contrast enhanced image and RGB image derived from original color fundus images.

R. Sarki et al. [37], proposed CNN based architecture for multi classification of diabetic eye disease in two ways, a low level multi class diabetic eye disease and another one is a high level multi class eye diabetic disease. Maximum accuracy for mild multi-classification and multi-classification are 88.3% and 85.95% using VGG16. K. Shankar et al. [38], proposed a synergic deep learning model for classifying the levels of diabetic retinopathy. The proposed method outperforms AlexNet, ResNet, GoogleNet and VggNet-19 with respect to accuracy, sensitivity and specificity. A. Ram et al. [39], used a CNN for feature extraction for classifying normal, cataract, myopia and AMD. The objective of this paper is to correlate the relationship between the number of classes and number of fully connected layers.

J. Wang et al. [40] used feature extraction based efficientnet in the first part and custom neural network in the second part for multilabel classification of fundus images. N. Gour et al. [41], used transfer learning for classification on fundus images by two approaches. In the first approach, images of both eyes were individually given as input for CNN and the results were concatenated and in the second method, images of both eyes were concatenated and given as input to CNN. Various state level architectures have used instead of CNN to generate better results and VGG16 pretrained architecture performed significantly. N. Li et al. [42], created a database of 10,000 fundus images of both eyes from 5000 patients to classify 8 diseases and multi level classification of images has improved significantly with the increase of complexity in state of art deep neural networks like AlexNet, ResNet, GoogleNet. J. He et al. [43], proposed a dense correlation network (DCN) for classifying multi labelled diseases. DCN consists of three modules for features extraction, features correlation and calculating classification score; a multi label soft margin loss was used as a loss function and produced way better results than benchmark deep neural networks.

In this paper [45], D. Muller et al. suggested ensemble heterogenous DL models for multi eye disease prediction. They also used fivefold cross validation on RFMID multilabel data containing 3200 images (1920 training data, 640 testing data and validation data each). This dataset contains twenty nine multilabel classes. The techniques used are data augmentation, bagging and stacking, transfer learning and stacked logistic regression. They achieved 0.95 AUROC for multilabel disease risk prediction by using ensemble DL models. In this paper [46], A. C. Garcia et al. used Resnet, Resnest, EfficientNet, ViT, Deit, NasNet, HRnet, CycleGAN on RFMID-2021 mainly focussed on the ERM category and achieved f1score of 86.82. In this paper [47], L. P. Cen et al. used DCNN on 249,620 images and 275,543 labels collected from different sources and achieved f1score, sensitivity, specificity, AUC of 0.923, 0.978, 0.996, 0.9984 resp. for multi-label classification dataset (Table I).

TABLE I. LITERATURE SURVEY OF RECENT STUDIES WITH SHORTCOMINGS

Reference No.	Author and year	Techniques used	Results	Limitations
[20]	M. Mateen et al., 2018	Gaussian mixture model (GMM), visual geometry group network (VGGNet), singular value decomposition (SVD) and principle component analysis (PCA)	Accuracy = 98.34	Does not consider the overfitting problem.
[21]	Q. Meng et al., 2019	Deep CNN with attention map mechanism	Accuracy = 94.5	Accurate classification accuracy can be checked on more data.
[22]	H. Chen et al., 2019	Deep hierarchical multi-label classification	AUC = 88.7	Results need to be evaluated using other performance parameters also.
[23]	L. Faes et al., 2019	Automated deep learning model	Sensitivity = 73·3–97·0 Specificity = 67–100% AUPRC = 0·87–1·00	Study should compare several state-of-art models with proposed one.
[24]	C. C. Jordi et al., 2019	VGG16 and InceptionV3	AUC = 88.71 F1-score = 88.76	Results can be improved with additional pre-processing steps.
[32]	Md. T. Islam et al., 2019	Shallow CNN architecture	F1-score = 85 Kappa score = 31 AUC = 80.5	Recent neural models can be implemented to evaluate results.
[44]	E. S. Kumar et al., 2021	Multi-Disease Classification Framework (MDCF) using stacking	AUC = 97.42 F1-score = 94.32	Different ensemble methods can be used.

A. Limitation in existing Architectures

The current state of the art methodologies has primarily 5 main limitations to highlight, as is discussed in the literature review section.

- Lack of a complete autonomous system to provide multi label classification at a medical acceptable rate. Majority of the research is done using one or multiclass eye disease. Applicability and deployment of deep learning techniques in classifying multilabel data is still in infancy stage.
- Absence of transfer running bio fueled ventures in retinopathy.
- The absence of high-volume dataset, diminishing the choice of deep learning.
- Use of traditional Machine learning techniques to measure the correlation of dimensional simulation, although feature extraction is difficult and eluding.
- Multiclass labelling pertains to quasi dimensional binary loss problem, general in multiclass labelling but deadly for retinopathy. To overcome this the paper hypotheses the use of cross-entropy modelling strategies.

III. Proposed Methodology

In this section, a pipelined architecture has been proposed based on deep learning using transfer learning techniques from Imagenet dataset to multi labelled classification of eye diseases. ODIR Dataset contains supervised data of 8 eye categories. They are Normal (N), Glaucoma (G), Diabetic Retinopathy (D), AMD (A), Hypertension (H), Cataract(C), Myopia (M) and other abnormalities (O) on fundus images as shown in Fig. 1. In this section, proposed architecture and database used has been explained.

A. Data Collection

The paper used Ocular Disease Intelligent Recognition (ODIR) dataset consisting 5000 images of ophthalmic patients of left and right eye, age and diagnostic key words of Doctor collected by Shanggong Medical Technology Co., Ltd. from different hospitals in China to classify diabetic retinopathy, glaucoma, cataract, age-related macular degeneration, hypertension, pathological myopia and other abnormalities. These multi labelled fundus images are captured by various cameras to create different resolutions. The Table II describes the count of each disease in the dataset and the augmentation details.

B. Pre-Processing and Augmentation

Images are cropped towards the centre to avoid the area which does not generate much information and various augmentation techniques were applied on the data sets labelled as Hypertension (H), Glaucoma (G), Cataract (C), AMD (A), Myopia (M) to create the balance among the datasets as these 5 diseases are largely outnumbered by other diseases. The augmented data set has increased from 7473 images to 14072 images as shown in Table II.

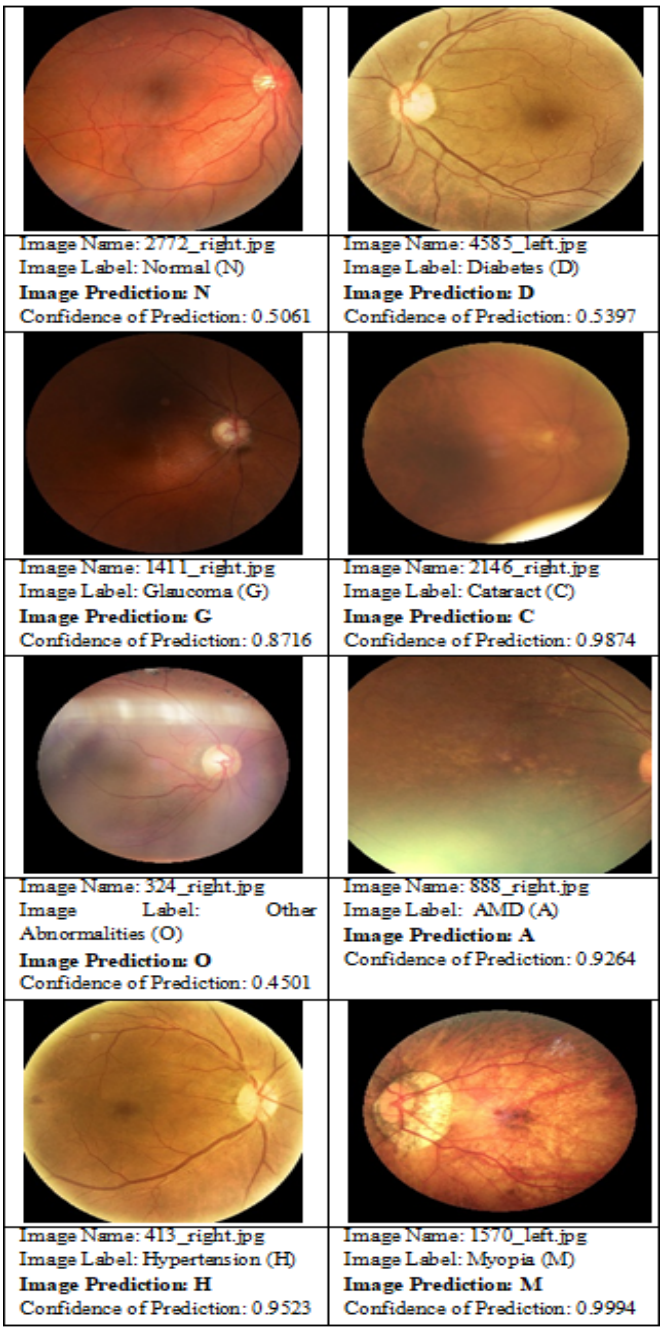


Fig. 1. Sample Images of 8 Eye Categories and their Confidence Levels of Prediction.

Each image of these 5 diseases were used to generate 8 more images using these 8 augmentation techniques, they are 1) Vertical flip, 2) Horizontal flip, 3) Both horizontal and vertical flip, 4) Clipped center of image and zoom of the original image, 5) Clipped center of image and zoom of the vertical flipped image, 6) Image rotation plus brightness enhancement, 7) Image rotation of the original image and 8) Image rotation of the vertical flipped image. Since it is a multi labelled dataset, augmentation of image containing two diseases creates 16 images in this augmentation approach, though no diabetic retinopathy images are augmented, its images are augmented due to multi labelled dataset.

TABLE II. TABULAR DESCRIPTION OF DATA SAMPLES AND AUGMENTATION DETAILS OF EACH DISEASE

Class	Total Samp-les	Training Split (70%)	Testing Split (30%)	Augmented	Augmented Sample (on training dataset)	Total Sampl-es (used for training)
N	3098	2169	929	No	0	2169
D	1801	1261	540	No	792	2053
G	326	229	97	Yes	1832	2061
C	313	218	95	Yes	1744	1962
A	277	193	84	Yes	1544	1737
H	193	135	58	Yes	1080	1215
M	268	188	80	Yes	1504	1692
O	1197	847	350	No	336	1183
Total	7473	5240	2233	Yes	8832	14072

C. Training

In proposed method, one versus rest strategy has used to classify each disease against all other 8 possibilities. The workflow of proposed methodology is shown in Fig. 2 and Fig. 3. Fig. 2 and Fig. 3 represent the flowchart and algorithm of our proposed methodology respectively. Eight different one versus rest classifier classification algorithms based on vgg16 were trained to detect N, D, G, C, A, H, M, O categories. Each eye image was taken separately (didn't consider left and right images of eyes as the same image) to double the size of the data set and implemented preprocessing and various augmentation methods to increase the data size to avoid over-fitting and to add versatility among the data. Images were cropped to the centre of fundus images and reshaped the size of image to 224×224 to avoid computational problems and these reshaped images are suitable to use VGG16 transferred weights.

1) Transfer learning: LeCun et al [48], proposed a convolutional neural network for extracting features in images, speeches and time series data. The basic layers in any convolutional neural network include convolution, pooling, batch normalization, fully connected layers as shown in Fig. 4. Various architectures are proposed by tweaking these layers by repeating more layers of one type or by changing the order of layers using different activation functions. The CNN's are great at localization and extracting features. CNNS are utilized in various fields like object detection, CNN's generated state of art results in segmentation, classification and enhancement on biomedical images and it requires a lot of data and computational power to perform matrix operations.

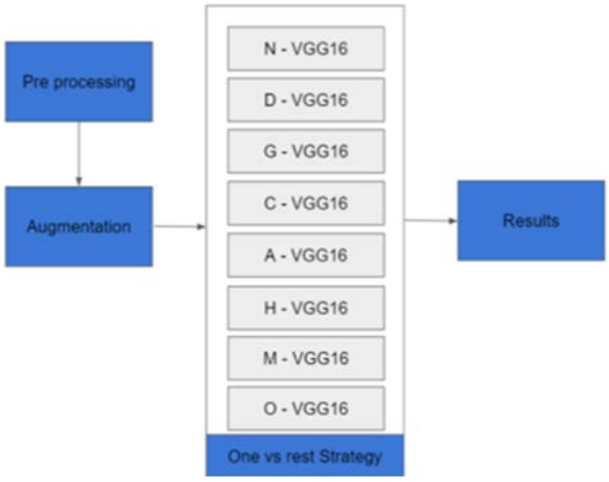


Fig. 2. Flowchart of Proposed Methodology.

For each batch on Train data do:

1. Resize image to 224 *224 Pixel.
2. Crop image to center.
3. If Image Labeled as G, C, H, A, M do:
   i. Vertical Flip.
   ii. Horizontal Flip.
   iii. Both Horizontal and Vertical Flip.
   iv. Clipped Center of image and zoom the original image.
   v. Clipped Center of image and zoom the vertical flipped image.
   vi. Image rotation plus Brightness enhancement.
   vii. Image rotation of original image.
   viii. Image rotation of vertical flipped image.
4. Else: No Augmentation.
5. For each label in data do:
   i. Train the VGG16 Model.
6. END

END

Fig. 3. Algorithm of Proposed Methodology.

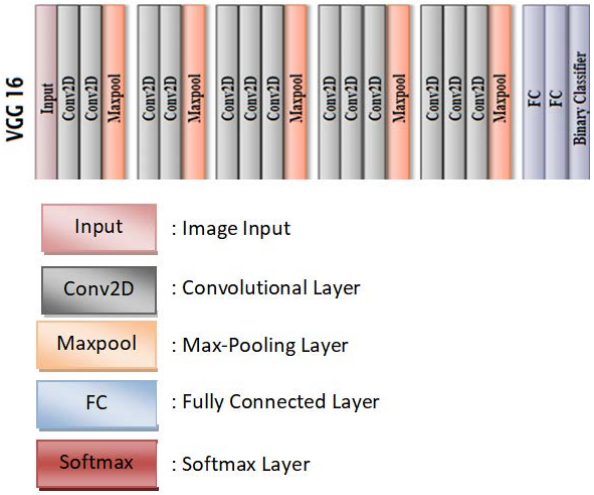


Fig. 4. Vgg16 Architecture. [48].

TABLE III. FILTERS IN CNN LAYERS

BLOCK ID	Number of layers	Number of CNN layers	Filters dimension	Number of Pooling layers
1	3	2	(3× 3 × 64)	1
2	3	2	(3× 3 × 128)	1
3	4	3	(3× 3 × 256)	1
4 & 5	4	3	(3× 3 × 512)	1

CNNs model weights are stored in open domains which are trained on large databases like ImageNet. Various CNN architectures have done well on ImageNet databases to detect and classify objects, the recent architectures are AlexNet[49], GoogleNet[50], VGGNet[51], MobileNet[52] and ResNet[53]. Collecting a large volume of multi labelled data in the medical field is a tedious task and very time consuming. Recently researchers moved to transfer learning where they use pre trained models on standard datasets and they use these pre-trained models on their datasets. There are various benefits in transfer learning like 1) it saves a lot of time for training the model and flexible enough to adjust trainable layers and non trainable layers. 2) It was trained on large dataset and has more parameters that are useful for learning, on a low sized datasets if we apply these standard models they mostly end up overfitting the data.

Proposed model uses VGG16 where its first 10 layers are not trainable and the rest are fine tuned and the last layer was adjusted to binary classification. The VGG16 contains 5 blocks, details about filters, number of convolutional layers and max pooling layers were given in Table III. In pooling layers kernel size is of 2×2 with a stride movement of 2 and in CNN layers ReLU used as activation function. Final parameters after block5 is 102764544 and these are flattened and forwarded to three sequentially fully connected neural networks. The last one uses softmax as activation for classification that outputs a vector of probability R as shown in equation (1) and the rest uses ReLU as activation function P that introduces nonlinearity to the network as shown in equation (2).

The softmax equation is defined by:

$$R = \begin{pmatrix} R1 \\ R2 \end{pmatrix} \text{ Where } R_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \quad (1)$$

The Relu equation is defined by:

$$P(z) = \max(0, z) \quad (2)$$

Detailed implementation of these architectures is shown in Fig. 4, and these architectures were trained 8 times, as shown in Fig. 2 and Fig. 3 i.e. one versus rest strategy to classify 8 diseases individually and their results are amalgamated for multi diseases classification. The model in Fig. 4 was trained for 16 epochs and utilized a batch size of 32 to train this model with a validation split of 0.2. Stochastic gradient descent [54] as shown in equation (3) was used as an optimization algorithm and hyper parameters such as learning rate, momentum, decay are adjusted as 0.0001, 0.9, 0.000006 and Nesterov Accelerated gradient as shown in equation (4) uses parameters θ calculated from momentum term γw_{t-1} gives approximation term for next parameter value through $\theta - \gamma w_{t-1}$ that improves the generalisation performance has applied with stochastic gradient to increase the speed of convergence. Binary cross entropy as shown in equation (5) was used for loss function and accuracy used as the measure of metrics. To train this model, hardware specifications of cpu with 4 crores of ram size 32 gb and 11 gb GPU of Tesla K80 Architecture have utilized.

The SGD equation for each training example $t^{(i)}$ and label $u^{(i)}$

$$\theta = \theta - \eta . \nabla_\theta \boldsymbol{J}(\theta; t^{(i)}; u^{(i)}) \quad (3)$$

The nesterov accelerated gradient equation is

$$w_t = \gamma w_{t-1} + \eta \nabla_\theta \mathbf{J}(\theta - \gamma w_{t-1})$$
$$\theta = \theta - w_t \quad (4)$$

The Binary cross entropy function is given by

$$BCE = -\frac{1}{N} \sum_{i=0}^{N} y_i . \log(\acute{y}) + (1 - y_i) . \log(1 - \acute{y}) \quad (5)$$

IV. RESULT

A comparison table was provided on various techniques used on ODIR dataset along with proposed method to classify these diseases in Table IV. These proposed models were evaluated on 2233 testing images of ODIR database as mentioned in section III (Table II) and results have been compared on the basis of accuracy of each disease with proposed method and with the base paper proposed by N. Gour et al. [41] and these results were shown in Fig. 5, contains details of training loss, validation loss of each disease and overall accuracy and individual disease accuracy. Proposed model detected myopia with more accuracy as its false positive rate and false negative rate are very less compared to other diseases as a result it showed significant improvement in accuracy, f1-score, and precision and recall whereas classification error became quite negligible. Identifying normal images became a quite tricky and its results are mediocre compared to identifying other diseases.

A. Performance Metrics

Various metrics such as Accuracy, Specificity, Precision, Sensitivity, and Classification error, F1-score, False Positive Rate, Negative Predictive Value and False Negative Rate of each disease on testing data have been shown in Table V and the barcharts for each disease and for the overall model shown in Fig. 6 and Fig. 7. The equations for performance metrics is shown in the equations 6 to 14.

1) Accuracy: It is defined as the ratio of sum of true positives and true negatives to the total number of samples.

$$Accuracy = (\text{True Positives} + \text{True Negatives}) \div \text{Total Number of Samples} \quad (6)$$

TABLE IV. COMPARISON OF EXPERIMENTS CONDUCTED ON ODIR DATASETS AND THEIR RESULTS

Reference No.	Method used	Eyes merging	Results
[32]	Proposed cnn architecture	No fusion techniques were used	F1-score: 0.85, kappa score: 0.31, AUC value :0.85
[39]	Model based on two cases 1)featured based efficient net and 2) custom based neural network for multilabel classification	No fusion techniques were used	Overall Validation Accuracy: 0.90, F1-score: 0.85 (for image size 299*299) Overall Validation Accuracy:0.92,F1-score: 0.89 (for image size 448*448)
[40]	Transfer learning applied on two cases 1) transfer learning on concatenated left and right eye images 2) transfer learning individually on left images and right images and they were merged before classification	Uses concatenation as fusion technique	Overall Validation Accuracy 0.89
[41]	Transfer learning	Both left and right eyes are merged using sum, product and concatenation	The mean of kappa, AUC and F1 score. Better results achieved for inception -v4 (0.7516) for product as fusion technique
[42]	Transfer learning with spatial correlation module	Uses concatenation as fusion technique	Uses average of kappa, AUC and F1-score. Resnet -101 produces better results around 0.827
proposed method	**Transfer learning using VGG-16**	**No fusion (considered left and right images as separate images)**	**Validation Accuracy :90.85 F1-score :0.91**

2) Specificity: Measures the accurate identification of true negative values, it is also called Selectivity and True Negativity.

$$Specificty = (\text{True Negatives}) \div (\text{False Positives} + \text{True Negatives}) \quad (7)$$

3) Precision: Measures the number of true positive values obtained over the total number of positive values, it is also called as positive Predictive value.

$$Precision = (\text{True Positives})\_ \div \_(\text{False Positives} + \text{True Positives}) \quad (8)$$

4) Sensitivity: The ratio of true positives to true positives and false negatives is called sensitivity; it is alsoreferred as True Positive Rate and Recall.

$$Sensitivity = (\text{True Positives}) \div (\text{True Positives} + \text{False Negatives}) \quad (9)$$

5) Classification error (C.E): It is defined as the ratio of sum of false positives and false negatives to the total number of samples.

$$C.E = (\text{True Positives} + \text{True Negatives}) \div \text{Total Number of Samples} \quad (10)$$

6) F1- score: It is defined as harmonic mean between precision and recall.

$$F1 - Score = 2 * (\text{Precision} \times \text{Recall}) \div (\text{Precision} + \text{Recall}) \quad (11)$$

7) False positive rate (FPR): It is measured as the ratio between false positive to false positive and true negative.

$$FPR = \text{False Positive} \div (\text{False Positive} + \text{True Negative}) \quad (12)$$

8) Negative predictive value (NPV): It is measured as the ratio between true negative to true negative and false negative.

$$NPV = \text{True Negative} \div (\text{True Negative} + \text{False Negative}) \quad (13)$$

9) False negative rate (FNR): It is measured as the ratio between false negative to false negative and true positive.

$$FNR = \text{False Negative} \div (\text{False Negative} + \text{True Positive}) \quad (14)$$

		Baseline (VGG16 - Multiclass)	Proposed Approach (VGG16 - One vs Rest, Binary model for each disease class)						
Eye Disorder	Disease Label	Baseline (Validation Accuracy)	Total Training Samples	Total Validation Samples	Training Time (per Epoch)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Normal	N	40.00%	9887	2472	282s - 29ms/step	0.2735	87.09%	0.3135	85.28%
Diabetic Retinopathy	D	89.00%	9887	2472	271s - 27ms/step	0.3302	85.16%	0.4536	82.56%
Glaucome	G	54.00%	9887	2472	270s - 27ms/step	0.0696	97.60%	0.2891	91.34%
Cataract	C	97.00%	9887	2472	272s - 27ms/step	0.0409	98.60%	0.1111	97.05%
AMD	A	88.00%	9887	2472	270s - 27ms/step	0.0505	98.29%	0.3808	90.45%
Hypertension	H	95.00%	9887	2472	272s - 27ms/step	0.1021	96.03%	0.214	93.08%
Myopia	M	90.00%	9887	2472	282s - 29ms/step	0.0384	98.68%	0.0592	98.18%
Others	O	44.00%	9887	2472	271s - 27ms/step	0.2458	91.44%	0.3911	88.88%
Overall Accuracy		**89.06%**							**90.85%**

Fig. 5. Comparison between Proposed Approach and Approach given by N. Gour et al. [41].

TABLE V. METRICS ON TEST DATA

Class	Confusion Matrix	FPR	NPV	FNR	Specificity	PPV	Sensitivity/ Recall	Classification Error	F1-Score	Accuracy
N	[[776, 386], [290, 639]]	0.3766	0.6879	0.2720	0.6234	0.6678	0.7279	0.3233	0.6966	0.6767
D	[[1545, 6], [524, 16]]	0.2727	0.0296	0.2533	0.7273	0.9961	0.7467	0.2535	0.8536	0.7465
G	[[1886, 108], [36, 61]]	0.6391	0.6289	0.0187	0.3609	0.9458	0.9813	0.0689	0.9632	0.9311
C	[[1955, 41], [11, 84]]	0.3280	0.8842	0.0056	0.6720	0.9795	0.9944	0.0249	0.9869	0.9751
A	[[1921, 86], [43, 41]]	0.6772	0.4881	0.2189	0.3228	0.9572	0.9781	0.0617	0.9676	0.9383
H	[[1887, 146], [36, 22]]	0.8690	0.3793	0.0187	0.1309	0.9282	0.9813	0.0871	0.9539	0.9129
M	[[1991, 20], [8, 72]]	0.2174	0.9000	0.0040	0.7826	0.9901	0.9959	0.0134	0.9930	0.9866
O	[[1741, 0], [350, 0]]	NaN	0.0000	0.1674	NaN	1.0000	0.8326	0.1674	0.9087	0.8326
Overall		0.41	0.49	0.09	0.45	0.93	0.90	0.12	0.91	0.87

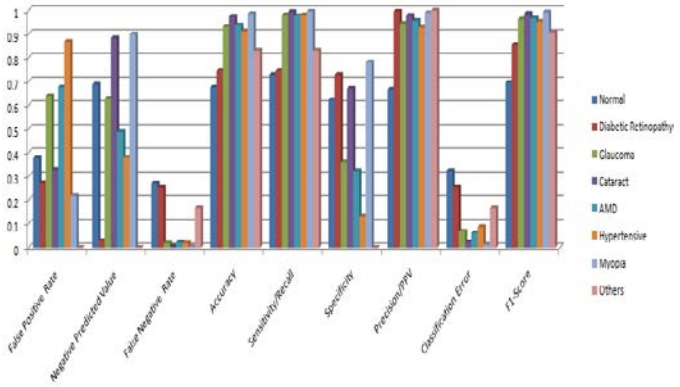


Fig. 6. Metrics in the Form of Barcharts for each and Every Disease.

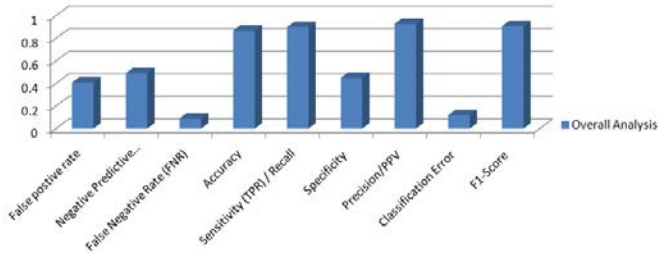


Fig. 7. Overall Analysis in the Form of Barcharts.

V. DISCUSSION

The paper demonstrates the lack of embedded machine learning techniques to solve the multi-class classification of retinopathy images. An important aspect of understanding the problem of the binary loss function is that because of the persistent 8 different classes, the exponential loss function is not able to provide substantially valid results as per the medical requirement. Because of which most of the retinopathy algorithms tend to create only theoretical steps to the earthquake in an autonomous system for multiclass classification in this field. The paper also takes into account various transfer learning accomplishments in the depth of image classification and computer vision. Furthermore, the paper can showcase the genesis idea of combining transfer learning into multiclass labelling techniques by using cross-entropy loss modelling ideas. A holistic discussion on various heuristics used is provided in the previous section and would be concluded in section 6. In the final analysis as per the literature review, we tend to understand the significance of machine learning ideas that have only been bubbled up and not implemented thoroughly. The paper ideation and implementation successfully answered where the research question is posted. While analyzing the results, it can be observed that transfer learning (Vgg16) with one verses rest classifier technique have produced exceptionally better results than baseline paper [41].

From the literature review, it is observed that no prediction is done on odir-19 testing dataset and most of the research is mainly focused on multiclass or one eye disease prediction. From the results, it is analyzed that training has produced reasonably better results (sensitivity, Accuracy, positive predictive value, F1-score) on overall testing data. Normal class shows average performance in all cases due to similarity in the features.

VI. CONCLUSION

The paper proposed a pipeline to identify multiple diseases on ODIR datasets where the aim was to increase baseline accuracy from 89% to almost 91% and also proposed model

has improved the performance of identifying disease drastically prediction of glaucoma has increased from 54% to 91%, Normal images prediction has increased from 40% to 85.28% and Other diseases prediction has increased from 44% to 88%. Out of 8 categories (seven diseases plus normal class) prediction proposed model prediction rate has improved in 6 categories, except in diabetic retinopathy and hypertension where proposed model accuracy has decreased by 6.44% and 2% respectively. The reason for achieving high accuracy in other categories is due to augmentation techniques where more balanced data is created as possible. That has clearly shown in less annotated diseases like glaucoma. The further research will be on working to create more data using other augmentation techniques like generating artificial images and working various transfer learning algorithms to improve the accuracy of each disease in multi-labeled classification problems. As for future work, other DL algorithms need to be explored for training the model and a study regarding hyperparameter optimization should be done to find the optimal model configuration. Moreover, other multi-label datasets with the latest eye diseases should be explored and predicted.

REFERENCES

[1] WHO: World report on vision. World Health Organisation (2019), https://www.who. int/publications-detail/world- report-on-vision

[2] C. C. W. Klaver, R. C. W. Wolfs, J. R. Vingerling, A. Hofman, and P. T. V. M. D. Jong, "Age-specific prevalence and causes of blindness and visual impairment in an older population: the Rotterdam Study." Archives of ophthalmology, Vol. 116, No. 5, 1998, pp. 653-658.

[3] M. H. Ansari, " Blindness after facial fractures: a 19-year retrospective study," Journal of oral and max- illofacial surgery, Vol. 63, No. 2, 2005, pp. 229-237.

[4] D. S. Fong, L. Aiello, T. W. Gardner, G.L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, F. L., 3rd, R. Klein, & American Diabetes Association, "Retinopathy in diabetes," Diabetes care, Vol. 27, 2004, pp. S84–S87. https://doi.org/10.2337/diacare.27.2007.s84

[5] B. Thylefors, and A. D. Negrel. "The global impact of glaucoma." Bulletin of the World Health Organization, Vol. 72, No. 3, 1994, pp. 323-326.

[6] D. Allen, and A. Vasavada, "Cataract and surgery for cataract," Bmj, Vol. 333, No. 7559, 2006, pp. 128-132.

[7] AREDS2 Research Group, E. Y. Chew, T. Clemons, J. P. SanGiovanni, R. Danis, A. Domalpally, W. McBee, R. Sperduto, and F. L. Ferris, "The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1)," Ophthalmology, Vol. 119, No. 11, 2012, pp. 2282–2289. https://doi.org/10.1016/j.ophtha.2012.05.027

[8] E. Nolte, R. Scholz, V. Shkolnikov, and M. McKee, "The contribution of medical care to changing life expectancy in Germany and Poland," Social science & medicine (1982), Vol. 5, No. 11, 2002, pp. 1905–1921. https://doi.org/10.1016/s0277-9536(01)00320-3

[9] L. Konstantinidis, and Y. G. Crosier, "Hypertension and the eye." Current opinion in ophthalmology, Vol. 27, No. 6,2016, pp. 514-521.

[10] I. G. Morgan, K. O. Matsui, and S. M. Saw, "Myopia." The Lancet, Vol. 379, No. 9827, 2012, pp. 1739-1748.

[11] D. T. Munroe, and M. G. Madden, "Multi-class and single-class classification approaches to vehicle model recognition from images," proc. AICS, 2005, pp. 1-11.

[12] J. Liu, D. W. K. Wong, J. H. Lim, N. M. Tan, Z. Zhang, H. Li, F. Yin, B. Lee, S. M. Saw, L. Tong, and T. Y. Wong, "Detection of pathological myopia by PAMELA with texture-based features through an SVM ap- proach." Journal of Healthcare Engineering, 2010, pp. 1-11.

[13] T. V. Phan, L. Seoud, H. Chakor, and F. Cheriet, "Automatic screening and grading of age-related macular degeneration from texture analysis of fundus images." Journal of ophthalmology, 2016.

[14] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, 2016, pp. E1-E9.

[15] H. Pratt, F. Coenen, D. M. Broadbent, , S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," in International Conference On Medical Imaging Understanding and Analysis (MIUA), Procedia Computer Science, ELSEVIER, Vol. 90, 2016, pp. 200-205.

[16] J. Y. Choi, T. K. Yoo, J. G Seo, , J. Kwak, T. T. Um, and T. H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," PLoS ONE, Vol. 12, No. 11, 2017, pp. 1-17.

[17] P. M. Burlina, N. Joshi, K. D. Pacheco, T. Y. A. Liu, and N. M. Bressler, "Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration," JAMA Ophthalmology, 2018, doi:10.1001/jamaophthalmol.2018.6156.

[18] Y. Chai, H. Liu, and J. Xu, "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models." Knowledge-Based Systems, Vol. 161, 2018, pp. 147-156.

[19] F. Grassmann, J. Mengelkamp, C. Brand, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H.F. Weber, "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography," AMERICAN ACADEMY OF OPHTHALMOLOGY, 2018, pp. 1-11.

[20] M. Mateen, J. Wen, S. Song, & Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," Symmetry, Vol. 11, No. 1, 2019.

[21] Q. Meng, Y. Hashimoto, & S. I. Satoh, " Fundus image classification and retinal disease localization with limited supervision," In Asian conference on pattern recognition, Springer, 2019, pp. 469-482.

[22] H. Chen, S. Miao, D. Xu, G. D. Hager, & A. P. Harrison, "Deep hierarchical multi-label classification of chest X-ray images. In International Conference on Medical Imaging with Deep Learning," arxiv, in Proceedings of Machine Learning Research (PMLR), 2019, pp. 109-120.

[23] L. Faes, S. K. WagnerD. J. Fu, X. Liu, E. Korot, J. R. Ledsam, & P. A. Keane, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," The Lancet Digital Health, Vol. 1, No. 5, pp. e232-e242.

[24] C.C. Jordi, N.D.R. Joan Manuel, V.R. Carles, "Ocular Disease Intelligent Recognition Through Deep Learning Architectures", Universitat Oberta de Catalunya, 2019, pp. 1–114.

[25] M. N. Bajwa, M. I. Malik, S. A. Siddiqui, A. Dengel, F. Shafait,W. Neumeier, and S. Ahmed, "Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning." BMC medical informatics and decision making, Vol. 19, No. 136, 2019, pp. 1-16.

[26] S. Keel, Z. Li, J. Scheetz, L. Robman, J. Phung, G. Makeyeva, K. Aung, C. Liu, X. Yan, W. Meng, R. Guymer, R. Chang, and M. He, "Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs," Clin. Exp. Ophthalmol., Vol. 47, No. 8, pp. 1009–1018, 2019, doi: 10.1111/ceo.13575.

[27] V. Das, S. Dandapat, P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images," in Biomedical Signal Processing and Control, Vol. 54, 2019, pp. 1-10. doi: 10.1016/j.bspc.2019.101605.

[28] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, " Diagnostic

Assessment of Deep Learning Algorithms for Diabetic Retinopathy Screening," Information Sciences, 2019, pp. 511-522. doi:10.1016/j.ins.2019.06.011.

[29] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong, E. Y. Chew, and Z. Lu, "DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," Ophthalmology, Vol. 126, No. 4, 2019, pp. 565-575.

[30] T.Pratap, P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," Biomedical Signal Processing and Control, Vol. 53, 2019.

[31] M. S. Alabshihy, A. A. Maksoud, M. Elmogy, S. Barakat, and F. A. Badria, "Diagnosis of Diverse Retinal Disorders Using a Multi-Label Computer-Aided System, "Trends in Ophthalmology Open Access Journal, Vol. 2, No. 3, 2019, pp. 140-157.

[32] Md.T. Islam, S. A. Imran,A. Arefeen, M. Hasan, and C. Shahnaz, "Source andCamera Independent Ophthalmic Disease Recognition fromFundus Image Using Neural Network," in IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON), 2019, pp. 59-63.

[33] T. Nazir, A. Irtaza, A. Javed, H. Malik, D. Hussain, and R. A. Naqvi, "Retinal Image Analysis for Diabetes-Based Eye Disease Detection Using Deep Learning." Applied Sciences, Vol. 10, No. 18, 2020.

[34] X. Pan, K. Jin, J. Cao, Z. Liu, J. Wu, K. You, Y. Lu, Y. Xu, Z. Su, J. Jiang, K. Yao, and J. Ye, "Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning." Graefe's Archive for Clinical and Experimental Ophthalmology, Vol. 258, No. 4, 2020, pp. 779-785.

[35] M. Aamir, M. Irfan, T. Ali, G. Ali, A. Shaf, A. S. S, A. Al-Beshri, T. Alasbali, and M. H. Mahnashi, "An Adoptive Threshold-Based Multi-Level Deep Convolutional Neural Network for Glaucoma Eye Disease Detection and Classification," Diagnostics, Vol. 10, No. 8, 2020, doi: 10.3390/diagnostics10080602.

[36] C. G. Gonzalo, V. S. Gutierrez, P. H. Martınez, I. Contreras, Y. T. Lechanteur,A. Domanian,B. V. Ginneken, and C. I. Sanchez, "Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration ," Acta Ophthalmologica, 2020 , pp. 368-377.

[37] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multiclass diabetic eye diseases using deep learning," Heal. Inf. Sci. Syst., Vol. 8, No. 1, 2020, pp. 1–9, doi: 10.1007/s13755-020-00125-5.

[38] K. Shankar, , A. R. W. Sait, , D. Gupta, , S.K. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model," Pattern Recognition Letters, Vol. 133, 2020, pp. 210-216.

[39] A. Ram, and C. C. Reyes-Aldasoro, "The relationship between Fully Connected Layers and number of classes for the analysis of retinal images," arxiv, 2020, [Online]. Available: http://arxiv.org/abs/2004.03624.

[40] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-Label Classification of Fundus Images with EfficientNet," IEEE Access, Vol. 8, 2020, pp. 212499–212508, doi: 10.1109/ACCESS.2020.3040275.

[41] N. Gour and P. Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," Biomed. Signal Process. Control, Vol. 66, 2021, doi: 10.1016/j.bspc.2020.102329.

[42] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, "A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Vol. 12614 LNCS, 2021, pp. 177–193, doi: 10.1007/978-3-030-71058-3_11.

[43] J. He, C. Li, J. Ye, Y. Qiao, and L. Gu, "Multi-label ocular disease classification with a dense correlation deep neural network," Biomed. Signal Process. Control, Vol. 63, 2021, doi: 10.1016/j.bspc.2020.102167.

[44] E. S. Kumar & C. S. Bindu, "MDCF: Multi-Disease Classification Framework On Fundus Image Using Ensemble Cnn Models," Journal of Jilin University, Vol. 40, No. 09, pp.35-45, 2021

[45] D. Müller, I. Soto-Rey, and F. Kramer, "Multi-Disease Detection in Retinal Imaging based on Ensembling Heterogeneous Deep Learning Models", arxiv, 2021, Available : http://arxiv.org/abs/ 2103.14660.

[46] A.C. Garcia, M.G.Dominguez, J. Heras, A. Ines, D.Royo, and M. A. Zapala, "Prediction of Epiretinal membrane from Retinal fundus images using deep learning, in CAEPIA 2021, LNAE 12882 19th ed. Springer, 2021, pp. 2-13.

[47] L. P. Cen, J. Ji, J. W. Lin, S. T. Ju, H. J. Lin, T. P. Li, Y. Wang, J. F. Yang, Y. F. Liu, S. Tan, L. Tan, D. Li, Y. Wang, D. Zheng, Y. Xiong, H. Wu, J. Jiang, Z. Wu, D. Huang, T. Shi, B. Chen, J. Yang, X. Zhang, L. Luo, C. Huang, G. Zhang, Y. Huang, T. K. Ng, H. Chen, Weiqi Chen, C. P. Pang, and M. Zhang, "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," NATURE COMMUNICATIONS, Vol. 12, 2021.

[48] Y. LeCun, and Y. Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks, Vol. 3361, No. 10, 1998, pp. 255-258.

[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[51] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv, 2015. arXiv preprint arXiv:1409.1556.

[52] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv, 2017, arXiv preprint arXiv:1704.04861.

[53] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[54] L. Bottou, "Stochastic gradient descent tricks," Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012, pp. 421-436.

Collaborative Multi-Resolution MSER and Faster RCNN (MRMSER-FRCNN) Model for Improved Object Retrieval of Poor Resolution Images

Amitha I C[1], N S Sreekanth[2]
Department of Information Technology
Kannur University, Kannur, Kerala- 670567, India

N K Narayanan[3]
Indian Institute of Information Technology, Kottayam
Valavoor (P.O.) Kottayam-686635 Kerala, India

***Abstract*—Object detection and retrieval is an active area of research. This paper proposes a collaborative approach that is based on multi-resolution maximally stable extreme regions (MRMSER) and faster region-based convolutional neural network (FRCNN) suitable for efficient object detection and retrieval of poor resolution images. The proposed method focuses on improving the retrieval accuracy of object detection and retrieval. The proposed collaborative model overcomes the problems in a faster RCNN model by making use of multi-resolution MSER. Two different datasets were used on the proposed system. A vehicle dataset contains three classes of vehicles and the Oxford building dataset with 11 different landmarks. The proposed MRMSER-FRCNN method gives a retrieval accuracy 84.48% on Oxford 5k building dataset and 92.66% on vehicle dataset. Experimental results show that the proposed collaborative approach outperform the faster RCNN model for poor-resolution conditioned query images.**

Keywords*—Faster RCNN; feature representation; multi-resolution MSER; object detection; object retrieval

I. INTRODUCTION

Object detection and recovery is one of the emerging areas of computer vision. In our daily routine, almost everything is related to object-based and its retrieval. Object retrieval can be applied as a solution to various real-time problems. Retrieving objects present in a given image scene are far more difficult than content-based image retrieval [1]. In content-based image retrieval, the image is recovered as a whole, but in object retrieval, the region of interest (ROI) only retrieved from a database. Various research studies are being conducted in this area through traditional and deep learning-based approaches. On the object retrieval task, the retrieval is mainly based on object-level features. To identify an object location in an image, it is pre-marked with a box called the object's bounding or anchor box. In the initial stage the bounding boxes are drawn for the areas of the essential object [2].

Object recovery is the process of searching for an object in an image from a large image collection or video configuration. Object recovery occurs in two important steps: first, it searches for an image in a large database, and then observes an object with an anchor box. Initially, there were content-based image recovery structures, which were later improved to recover a specific object from a scene utilizing some formal techniques. The fundamental reason for object recognition is to distinguish and find at least one efficient target from a given image or video database. It meticulously incorporates a variety of key technologies, specifically pattern recognition and digital image processing. It has wide application possibilities in different areas such as protection of brand name or logo, detection of defective parts in printed circuit boards (PCB), accident prevention in road traffics [3], alerts about hazardous products in manufacturing plants and military confined region checking [4],[5].

The object recovery process is conventionally settled by physically extricating feature representations, where the normal feature-elements are addressed by histogram-of-oriented gradients (HoG), scale invariant feature transforms (SIFT), Haar-like feature representations and other calculations that depend on grayscale [6]. In addition to the above-mentioned feature extraction procedures, specific object recovery can be performed using support vector machines (SVMs) or AdaBoost algorithms. These conventional feature extraction models are simply ready to recover low level feature components of an image data, such as colour, shape, texture, blobs and edges, and have restrictions in recognizing numerous objects under complex scenes because of their deprived generalization performances. Newer object detection models are mainly depending on deep convolutional neural network (DCNN / Deep conv-net) features. Their results are promising when compared with traditional models [7]. Some of the models based on DCNN are region-based convolutional neural network and its variants, you only look once (YOLO) and single shot multi-box detector (SSD) models. DCNN models are not just concentrating on the detailed surface features from the previous level convolution layer, but on the other hand, can get more significant level data from the next-level convolution layer [8],[9],[10],[11].

In addition to the conventional CNN process, the RCNN variants utilize a counter strategy to assume the target object regions in the feature maps, steadily adjusting the location info and optimize the object's location for categorization and retrieval. Conversely, other object discovery models will concurrently foresee the anchor boxes and categorize straightforwardly in the feature maps by applying diverse convolutional phases. The RCNN model has two activity stages that consider higher location precision, while SSD and YOLO can straightforwardly recognize the arrangement and the position data, which speeds up detection [12]. A faster RCNN model offers better object retrieval accuracy than its

predecessors. The limitation of faster RCNN is that it cannot efficiently recover objects from bad resolution images.

To overcome the limitations of detecting poor / bad resolution images in faster RCNN, the proposed system uses an existing algorithm called multi-resolution MRMSER to formulate a new collaborative MRMSER-FRCNN model that offers better retrieval accuracy, compared to individual faster RCNN or multi-resolution MSER. With this integrated approach, it can achieve better retrieval accuracy on images with poor resolution conditions. The proposed collaborative MRMSER-FRCNN model offers better retrieval accuracy, compared to individual faster RCNN or multi-resolution MSER.

The rest of this paper is arranged as follows. Section 2 gives a detailed study report on various object retrieval techniques, their feasibility and the problems in the existing models. The identification and implementation of collaborative model based on MRMSER and faster RCNN explained in Section 3. Section 4 gives a detailed discussion about the datasets used and the results obtained in the simulation experiment. Finally, Section 5 concludes the proposed model.

II. Related Work

This section provides a detailed review on the object detection and retrieval from images and also explores accessible strategies to summarize in-depth features nearby for creating conservative descriptions for image recovery. Krizhevsky A. et al. [13] achieved better categorization of images in IMAGENET by DCNN. Their study categorizes a large number of high-resolution images. They have designed a neural network consisting of MaxPooling layers and five convolution layers that are fully integrated with SoftMax functionality. An object identification strategy was proposed based on the calculation of a regional proposal by Ren S. et al. in [14]. A region based network (RPN) that expects object boundaries and object simultaneously in each pixel area proposed by Long J. et al. in [15] illustrates a complete CNN, which is remarkably capable of classifying set of images semantically. The primary goal of their study is to create a “fully convolutional network” that receives the variable-sized image inputs and produces equally-sized outcome with viable deductions and information. Here, they configured some recent taxonomic networks, for example, Alex-Net, VGGNet and GoogLeNet for taxonomic purposes. Their fully CNN completes the excellent classification of images. Recovered feature is used as a standard image representation to handle different image object classifications in [16] by Sharif Razavian A.

Hossaine D. et al. suggest a deeper belief NN strategy for object identification. Once the object identification task is done, a live-path is created. The arrangement holds the object and keeps it in a pre-defined place. This deep learning method extracts adequate feature to detect the objects utilizing a computer vision framework [17]. Experimental inferences exhibits that, their proposed model performs better than other strategies. Babenko A and Lempitsky V [18] discussed about different enhancements demonstrated by the image descriptors offered by deep CNN, which greatly enhances image classification and recovery. The convolution layer could be clarified as nearby feature sets that describe image bounds explicitly. Amitha I C and N K Narayanan discuss the state of art about the conventional approaches and their inadequacies for object retrieval in [19]. Amitha I C and N K Narayanan suggested a proficient object recovery method in images utilizing SIFT-RCNN in [20].

Li H. et al. [21] have suggested new recovery method of image objects from an image database. The suggested model utilizes the power of CNNs, they utilized the further developed variant of the faster RCNN. They have tried their technique with various freely accessible databases. Oh I. et al. [22] recommended a strategy for segmenting multi-scale image dependent on maximally stable extremal regions (MSERs). They extended the essential MSERs usefulness (blobs recognition in image) to regular image segmentations [23].

Wang R. et al. [24] have proposed an upgraded, faster RCNN situated on the MSER inference standard for SAR image transport discovery in the harbors. The test result represents excellent recognition, and it distinguishes between their prospective technique and faster RCNN approach. This faster RCNN is utilized for recovering similar objects in various scenes. Faster RCNN is presented to conquer the issues in the fast RCNN model, the previous variant of CNN. Dubey A. K et al. [25] have introduced an efficient way to deal with deformity detecting in a rail route track surfaces utilizing MSERs stamping. Through this strategy the imperfections in the railway track can be assessed without much computation. All the above activities were performed within CNN family to further develop a recovery cycle or MRMSER based calculation to work on bad resolution or improve the lower surface regions in an image. The specific collective method utilizes the remarkable components of faster RCNN and MRMSER computations to recover the queried object with better accuracy from the specified image database.

In this combined model, it inspects the use of the multi-resolution maximally stable extremal regions (MRMSER) estimation to perceive the whole space of an image, regardless of the surface details. Locating blobs in images can be performed effectively through MSER and multi-resolution MSER systems. MSER is generally utilized during edge discovery, which gives better results with the mix of faster RCNN [26],[27],[28],[29],[30]. A layer connection strategy is utilized to perceive objects in low-resolution regions. The suggested method fuses a strategy for layer connection, to distinguish object in lower surface areas.

The main focus of the proposed MRMSER-FRCNN model is to retrieve objects efficiently, even if, the query image is affected due to poor resolution conditions or lower surface areas. Here, improve the poor resolution of the query image by applying specific multi-resolution MSER, and then apply the enhanced image input to the faster RCNN phase, for further object detection and retrieval.

III. Collaborative MRMSER-FRCNN Model

Recovery of objects in images can be done by conventional methods or by deep learning based methods.

With pre-trained CNNs, a large scope image database of high-resolution images can be organized into certain categories [13],[15],[16],[31],[32]. A common problem of not being able to distinguish areas beyond a specific region, were found in both region based CNN (RCNN) [33] and fast region based CNN. As a solution for this issue, the regions proposals networks (RPNs) were presented and joined with the final layer of fast RCNN [34], and the new organization is called as faster RCNN. Faster RCNN model recovery accuracy is lower when the image has lower resolution conditions due to varied features in the image texture [21]. The maximally stable extremal regions (MSERs) method is used to increase the feature extraction capability of faster RCNNs in poor resolution conditions in [35].

Particular object recovery in images is truly challenging in lower resolution conditions. In faster RCNN, the RoI pooling layers uses only the feature-maps of the top (best) convoluted layers, which leads to incorrect movement of feature extractions at lower resolution. Because of this problem, faster RCNN cannot retain the local attributes of an object. The MRMSER algorithm is integrated to solve the problem found in the faster RCNN model for efficient object recovery of the bad resolution state of the query image. The working of this novel collaborative approach is shown in the Fig. 1.

The proposed collaborative model works on the combination of two procedures, a faster RCNN followed by MSER/multi-resolution MSER. The query image is directly applied to the MRMSER or MSER phase. During this phase the MSER algorithm produces such an output image which will help faster RCNN phase detect and retrieve image objects even if, it is from a poor resolution condition or various textured regions. MSER calculations are done through a series of tasks such as grayscale conversion, detecting the MSERs and edge recognition, filtering based on region properties, determination of bounding box region and then produce a segmented image output. The image output from our previous phase is applied as the input to faster RCNN module. The faster RCNN model used in this work consists of 7 convolutional layers marked as conv1, conv2, up to conv7. The first five convolutional layers extract the features from the input image. The last two convolutional layers, conv6 and conv7 are fixed as fully connected layers. The input image is convolved with different filter sizes with varying strides.

The stride is the number of pixels shifts over the input matrix. Apply these convolutional feature map values to region proposal network (RPN) after conv5 layer. After the generation of region proposals, perform RoI pooling and then pass it through the fully connected layer. The output layer contains the bounding box information and the class score values.

Finally, to retrieve the queried objects, perform a ranking by class score generated in the previous stage. The detailed computation of MRMSER and faster RCNN are explained in the following subsections.

A. Multi-Rsolution Maximally Stable Extremal Region (MRMSER) Calculation

MSERs maintain a well-established framework for finding blobs in images. It is utilized to measure the intelligence between image parts of two images with alternate perspectives, providing broader benchmark adjustments. MSER will be utilized for better edge-detection in the proposed object detection framework. The in-depth edge detection ability of MSER overcomes the problems due to poor resolution conditions.

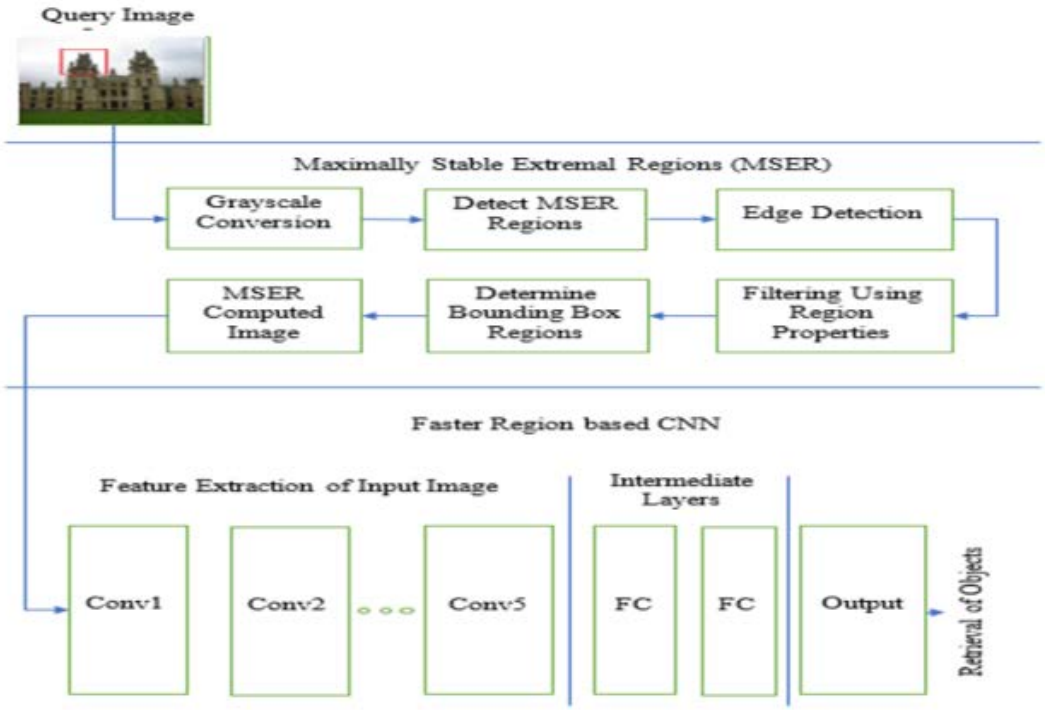


Fig. 1. Collaborative MRMSER-FRCNN Model.

MSER gives fundamental elements of an input image like all other feature detectors. MSER is reliably related fragment of two position sets of input images. This separates the image into several co-variant categories known as MSERs. MSER categories are related with regions describing uniform intensities surrounded by conflicting surroundings. The MSER procedure is described in the subsequent sections. In the initial step perform a grayscale conversion. Two options are there to perform the gray scale transformation in MSER - the average technique and the weighted technique or the luminosity technique. Each technique has its own advantages and disadvantages. Our system uses the weighted technique or the luminosity technique, which can be calculated as in Eq. (1) [24],[36].

$$I_G=0.3R+0.59G+0.11B \quad (1)$$

where I_G is the grayscale transformation and Red, Green and Blue channels are represented by R, G and B.

To guarantee the regions are maximally stable, ought to follow the requirements [35], which shown in Algorithm 1.

Algorithm 1: Algorithm for identification of areas utilizing MSER

Input:

Image I

Delta Parameters: To compute the similarities.

Step 1: For every pixel abbreviated by intensity values.

a. Spot a pixel in an image, when its turn arises.
b. Modify the pattern of the associated parts.
c. Update the region for affected related areas.

Step 2: For every single associated part.

a. Recognize areas with nearby minima concerning the speed of progress of the associated part region with an edge; characterize each such locale as MSE.

Output:

List of nested extremal areas.

Regardless, whether or not an extremely area is maximally consistent, it may be excused if:

- It is excessively large
- It is nearly nothing
- It is extremely shaky.
- It is exorbitantly similar to their parent.

Steps for the execution of MSER extractions are explained in Algorithm 2. Algorithm 3 gives the mathematical formulation of MSERs.

Algorithm 2: MSER extraction steps.

1. Perform the basic brightness of the image and change the intensity range from black to white.
2. Acquire related regions (“Extremal Regions”).
3. Identify a limit when an extremal region is “Maximally Stable”.
4. Estimate the region by an oval (optional).
5. Save those region descriptors as elements.

Algorithm 3: Mathematical Formulation of MSERs

Input:

Image I

I: D⊂Z2 →S

Extremal areas are well defined on image if:

S is completely arranged

A ⊂ D x D

p, q Є D are adjacent(pAq)

iff, $\sum_{i=1}^{d} |p_i - q_i| \leq 1$

∂Q= {q Є D\ Q:]p Є Q}

∂Q is the (Outer) Region Bounds

Extremal Region Q ⊂ D

I (p) > I (q) : Maximum

I (p) < I (q) : Minimum

Q1,….,Qi-1,….,Qi,….be the set of extremal areas/regions.

$Q_i \subset Q_{i+1}$

Q_{i*} is maximally stable

iff $q(i)=|Q_{i+\Delta} \setminus Q_{i-\Delta}| / |Q_i|$

| |: represents the cardinality

Δ Є S is a parameter of the process

To further develop the recovery exactness of the proposed procedure, a variation of MSER calculation is utilized named multi-resolution MSER (MRMSER). An MRMSER can be determined through the accompanying steps, which is shown in Algorithm 4.

Algorithm 4: Algorithm for MRMSER

Step1: Rather finding feature just from the image, generate a scale pyramid with an octave among scales.
Step 2: Identify MSERs exclusively at every resolution.
Step 3: Removes copy MSERs by erasing the best scale MSERs with comparative areas and sizes as MSERs found on the following rough scale.

Scale pyramid can be constructed through obscure and resample by a Gaussian filter.

After computing the MRMSERs of a given query image, generate the output and apply this image as an input to the faster RCNN phase.

B. Faster Region based CNN (FRCNN)

The second phase of the proposed system is based on faster RCNN. A common problem of not being able to distinguish areas beyond a specific region, were found in both region based CNN (RCNN) [33] and fast region based CNN. Faster RCNN model recovery accuracy is low when the image has lower resolution conditions.

FRCNN provides improved detection of analogous objects with the help of MRMSER. The proposed FRCNN consists of seven convolutional layers. The overall implementation task of FRCNN is performed across various convolutional layers. Initially, it receives an input image and pass it to a region proposal network (RPN), which initiates the RPN's task using the anchor boxes. Anchor boxes of different sizes were used depending on the size of the objects. After the generation of anchor boxes, compute intersection over union (IoU) on these bounding blocks / boxes. If IoU$\geq$ 0.5, accept it as the object bounding box and label it as foreground, else reject that box and recognize it as background. At the same time, the first five convolutional layers computes the feature maps and transfer it to RoI Pooling layer, which reduces the size of the feature maps created on the previous layers. A regressor refines the bounding box and classifier categorizes the object.

The real object recovery task is done in the FRCNN stage. The processed and modified image from MRMSER phase is applied as the input to FRCNN phase. The requested query image with poor-resolution condition was corrected by the MSER or MRMSER computation. The different steps associated with the FRCNN stage is described in Algorithm 5.

Algorithm 5: Algorithm for object detection using faster RCNN

1. A queried image is applied as input from MRMSER.
2. Concatenate conv3 and conv5 layers of the prospective system.
3. Finely tune the model.
4. Perform L-2 normalizations, which combines different scales and Norms in conv3 and conv5.
5. Compare the similarity of regional proposals.

a. Confidence score selections is done by setting up a threshold value.

b. According to the similarity score, rank the objects.

IV. Results and Discussion

To evaluate the efficiency of the proposed MRMSER-FRCNN method, the system is being simulated and tested on freely available Oxford buildings and vehicles database.

- Vehicle Dataset: The total number of images in this database is 150, which includes three different types of vehicles: bus, car and motorbike. Each individual vehicle class contains 50 pictures. Out of 150 images, 80% are utilized for training and the remaining 20% for testing.
- Oxford Building Dataset [37]: The total number of images in this dataset is 5062. This dataset provides 11 different landmarks for buildings. Each landmark was given five query images, so they affixed 55 query images to the entire dataset. All query images are marked with bounding boxes of required objects. Here also 80% images are utilized for training and the remaining 20% for testing.

Experimental results shows better object retrieval accuracy than faster RCNN or MSER-FRCNN. To check the retrieval accuracy of poor resolution, such images are tested and results are tabulated. The retrieval accuracy of the proposed method is compared with other similar methods also.

The object retrieval accuracy is calculated as the ratio of correctly retrieved objects and number of total objects in the reference dataset. The proposed MRMSER-FRCNN method gives a retrieval accuracy of 84.48 % in Oxford building dataset and 92.66 % in vehicle dataset. The previous method [35] gives 71.4 % in Oxford buildings dataset and 86 % in vehicle dataset. Faster RCNN shows 63.2 % in Oxford building dataset and 79 % in vehicle dataset. The sample object retrieval from Oxford building dataset is shown in the Fig. 2. The left most image surrounded by red colour box is the query image object and all other pictures surrounded by green colour box are the correctly retrieved image objects.

The detailed result analyses of MRMSER-FRCNN in Oxford building and vehicle dataset are shown in Tables I and II. The result comparison with various methods and proposed method is shown in Table III. Fig. 3 gives the bar chart comparison representation of the newly proposed MRMSER-FRCNN with other methods.

The object retrieval overall accuracy in Oxford building dataset with different methods are listed in Table IV. The proposed model retrieval accuracy is shown in bold values. Their performance comparison is shown in Fig. 4.

Fig. 2. Sample Object Retrieval from Oxford Building Dataset.

TABLE I. OBJECT RETRIEVAL WITH MRMSER-FRCNN IN OXFORD BUILDING DATASET

Land Marks in Oxford dataset	Objects retrieved correctly	Accuracy (%) Oxford Building Dataset -5k
All Souls	125	94.69
Ashmolean	146	81.1
Balliol	140	90.9
Bodleian	174	81.69
Christ Church	450	82.87
Cornmarket	48	81.35
Hertford	55	82.08
Keble	98	80.99
Magdalen	550	80.29
Pitt Rivers	88	81.48
Radcliffe Camera	229	81.20
Overall Accuracy		84.48 %

TABLE II. OBJECT RETRIEVAL WITH MRMSER-FRCNN IN VEHICLE DATASET

Vehicle Class	Vehicles retrieved correctly	Vehicle Dataset Accuracy (%)
Bike	47	94
Bus	46	92
Car	46	92
Overall Accuracy		92.66 %

TABLE III. COMPARISON WITH PROPOSED AND OTHER METHODS

Method	Datasets	
	Vehicle	*Oxford Buildings*
Faster RCNN	79.00	63.20
MSER-FRCNN	86.00	71.40
MRMSER-FRCNN (Proposed Method)	92.66	**84.48**

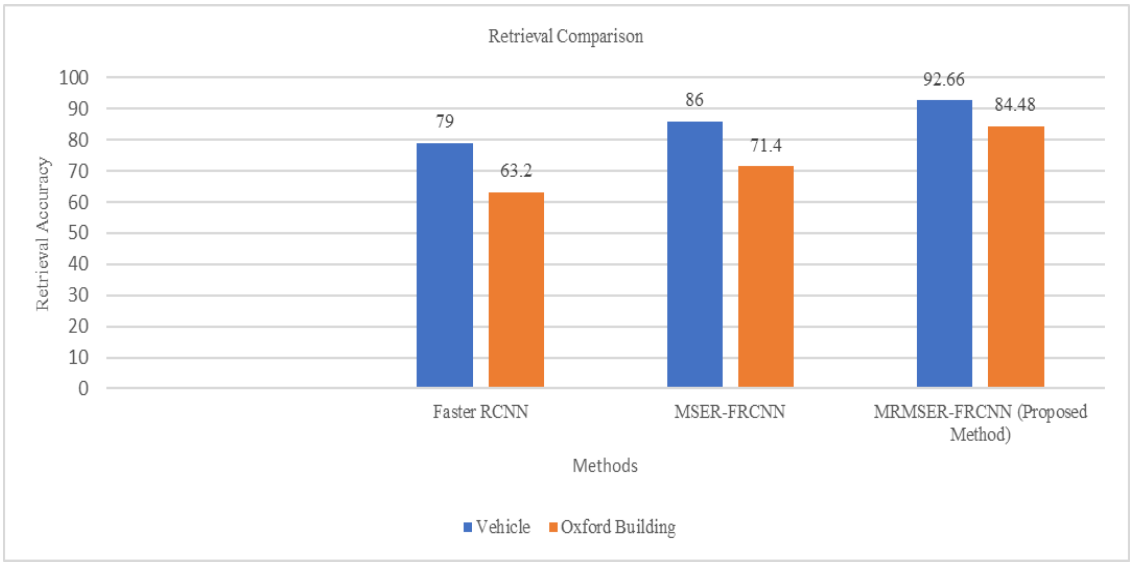


Fig. 3. Comparison with Proposed and other Methods.

TABLE IV. COMPARISON WITH PROPOSED AND OTHER METHODS IN OXFORD BUILDING DATASET

Methods	**Object retrieval Accuracy (%)**
SIFT and CNN [38]	81.60
SIFT and RCNN [20]	82.10
Faster RCNN [21]	63.20
MSER-FRCNN [35]	71.40
MRMSER-FRCNN (Proposed Method)	**84.00**

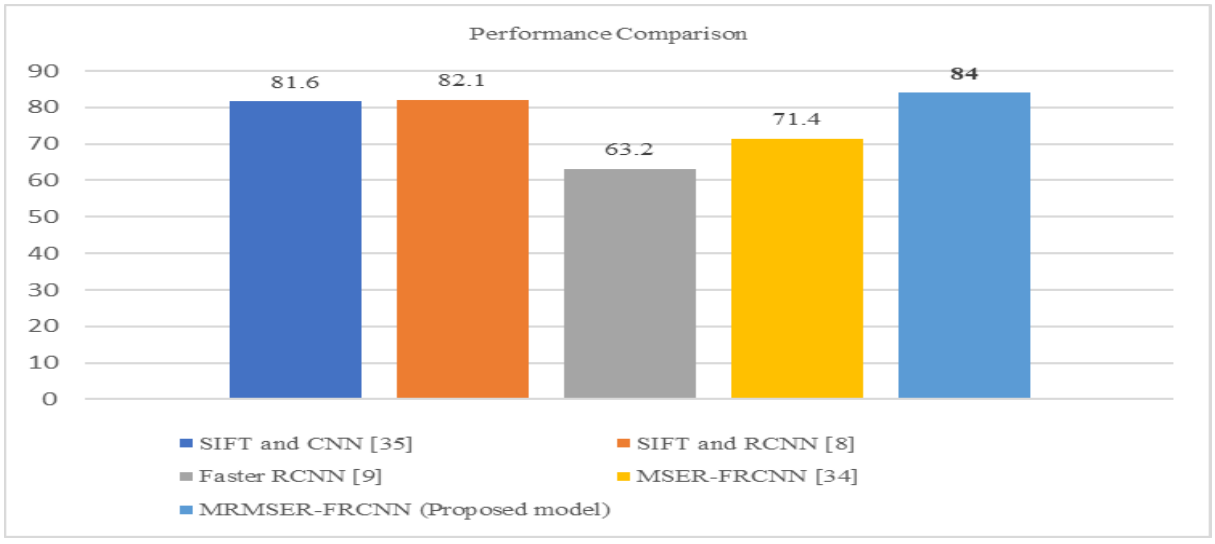


Fig. 4. Comparison with Proposed and other Methods in Oxford Building Dataset.

V. CONCLUSION

In this paper, a new collaborative approach to recovering objects from poor resolution images was proposed. The proposed MRMSER-FRCCN method can overcome the issue with the collaboration of MRMSER strategy. This will take care of smaller objects presented in the image with poor lighting conditions or poor resolution. Experimental results are obtained for collaborative method on two different publicly available datasets. The retrieval accuracies were compared with other individual as well as combined methods. The proposed MRMSER-FRCNN method gives a retrieval accuracy of 84.48% in Oxford building dataset and 92.66% in vehicle dataset. The previous works reported only a maximum

of 71.4% in Oxford building dataset and 86% in vehicle dataset. Hence it is found that the proposed MRMSER-FRCCN method outperforms conventional methods reported in literature.

REFERENCES

[1] Han, Xian-Feng, Hamid Laga, and Mohammed Bennamoun. "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era." IEEE transactions on pattern analysis and machine intelligence 43, no. 5 (2019): 1578-1604.

[2] Puranik, Vaishali, and A. Sharmila. "Integration of Basic Descriptors for Image Retrieval." In International Conference on Information Management & Machine Intelligence, pp. 629-634. Springer, Singapore, 2019.

[3] Shine, Linu, and C. Victor Jiji. "Automated detection of helmet on motorcyclists from traffic surveillance videos: a comparative analysis using hand-crafted features and CNN." Multimedia Tools and Applications (2020): 1-21.

[4] Liu, Jin, Yihe Yang, ShiqiLv, Jin Wang, and Hui Chen. "Attention-based BiGRU-CNN for Chinese question classification." Journal of Ambient Intelligence and Humanized Computing (2019): 1-12.

[5] Cao, Danyang, Menggui Zhu, and Lei Gao. "An image caption method based on object detection." Multimedia Tools and Applications 78, no. 24 (2019): 35329-35350.

[6] Hu, Rui, and John Collomosse. "A performance evaluation of gradient field hog descriptor for sketch-based image retrieval." Computer Vision and Image Understanding 117, no. 7 (2013): 790-806.

[7] He, Xinwei, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. "Triplet-center loss for multi-view 3d object retrieval." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1945-1954. 2018.

[8] Amitha, I. C., and N. K. Narayanan. "Improved Vehicle Detection and Tracking Using YOLO and CSRT." In Communication and Intelligent Systems, pp. 435-446. Springer, Singapore, 2021.

[9] Amitha, I. C., and N. K. Narayanan. "Object Detection Using YOLO Framework for Intelligent Traffic Monitoring." In Machine Vision and Augmented Intelligence—Theory and Applications, pp. 405-412. Springer, Singapore, 2021.

[10] Tasnim, Zarrin, FM Javed Mehedi Shamrat, Md Saidul Islam, Md Tareq Rahman, Biraj Saha Aronya, Jannatun Naeem Muna, and Md Masum Billah. "Classification of Breast Cancer Cell Images using Multiple Convolution Neural Network Architectures." cancer 12, no. 9 (2021).

[11] Methun, Naimur Rashid, Rumana Yasmin, Nasima Begum, Aditya Rajbongshi, and Md Ezharul Islam. "Carrot Disease Recognition using Deep Learning Approach for Sustainable Agriculture."

[12] Cao, Danyang, Zhixin Chen, and Lei Gao. "An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks." Human-centric Computing and Information Sciences 10, no. 1 (2020): 1-22.

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60, no. 6 (2017): 84-90.

[14] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster RCNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39, no. 6 (2016): 1137-1149.

[15] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.

[16] Sharif Razavian, Ali, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. "CNN features off-the-shelf: an astounding baseline for recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 806-813. 2014.

[17] Hossain, Delowar, GenciCapi, and Mitsuru Jindai. "Object recognition and robot grasping: A deep learning based approach." In The 34th Annual Conference of the Robotics Society of Japan (RSJ 2016), Yamagata, Japan. 2016.

[18] Babenko, Artem, and Victor Lempitsky. "Aggregating local deep features for image retrieval." In Proceedings of the IEEE international conference on computer vision, pp. 1269-1277. 2015.

[19] Amitha, I. C., and N. K. Narayanan. "Image object retrieval using conventional approaches: a survey." Int J Eng Technol Sci (IJETS) (2018): 1-4.

[20] Amitha, I. C., and N. K. Narayanan. "Object Retrieval in Images using SIFT and RCNN." In 2020 International Conference on Innovative Trends in Information Technology (ICITIIT), pp. 1-5. IEEE, 2020.

[21] Li, Hailiang, Yongqian Huang, and Zhijun Zhang. "An improved faster RCNN for same object retrieval." IEEE Access 5 (2017): 13665-13676.

[22] Oh, Il-Seok, Jinseon Lee, and Aditi Majumder. "Multi-scale image segmentation using MSER." In International Conference on Computer Analysis of Images and Patterns, pp. 201-208. Springer, Berlin, Heidelberg, 2013.

[23] Girshick, Ross. "Fast RCNN." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[24] Wang, Rufei, Fanyun Xu, Jifang Pei, Chenwei Wang, Yulin Huang, Jianyu Yang, and Junjie Wu. "An improved faster RCNN based on MSER decision criterion for SAR image ship detection in harbor." In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 1322-1325. IEEE, 2019.

[25] Dubey, Ashwani Kumar, and ZainulAbdinJaffery. "Maximally stable extremal region marking-based railway track surface defect sensing." IEEE Sensors Journal 16, no. 24 (2016): 9047-9052.

[26] LeCun, Yann, YoshuaBengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444.

[27] Zhou, Wengang, Houqiang Li, and Qi Tian. "Recent advance in content-based image retrieval: A literature survey." arXiv preprint arXiv:1706.06064 (2017).

[28] Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu, and Xindong Wu. "Object detection with deep learning: A review." IEEE transactions on neural networks and learning systems 30, no. 11 (2019): 3212-3232.

[29] Liu, Li, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. "Deep learning for generic object detection: A survey." International journal of computer vision 128, no. 2 (2020): 261-318.

[30] Jo, YoungJu, Hyungjoo Cho, Sang Yun Lee, Gunho Choi, Geon Kim, Hyun-seok Min, and YongKeun Park. "Quantitative phase imaging and artificial intelligence: a review." IEEE Journal of Selected Topics in Quantum Electronics 25, no. 1 (2018): 1-14.

[31] Yim, Junho, Jeongwoo Ju, Heechul Jung, and Junmo Kim. "Image classification using convolutional neural networks with multi-stage feature." In Robot Intelligence Technology and Applications 3, pp. 587-594. Springer, Cham, 2015.

[32] Jaswal, Deepika, S. Vishvanathan, and S. Kp. "Image classification using convolutional neural networks." International Journal of Scientific and Engineering Research 5, no. 6 (2014): 1661-1668.

[33] Fontdevila Bosch, Eduard. "Region-oriented convolutional networks for object retrieval." Bachelor's thesis, UniversitatPolitècnica de Catalunya, 2015.

[34] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.

[35] Amitha, I. C., and N. K. Narayanan. "Collaborative MSER and Faster R-CNN Model for Retrieval of Objects in Images." In Soft Computing for Problem Solving, pp. 673-682. Springer, Singapore, 2021.

[36] Cao, Changqing, Bo Wang, Wenrui Zhang, Xiaodong Zeng, Xu Yan, Zhejun Feng, Yutao Liu, and Zengyan Wu. "An improved faster RCNN for small object detection." IEEE Access 7 (2019): 106838-106846.

[37] Philbin, James, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object retrieval with large vocabularies and fast spatial matching." In 2007 IEEE conference on computer vision and pattern recognition, pp. 1-8. IEEE, 2007.

[38] Zhang, Guixuan, Zhi Zeng, Shuwu Zhang, Yuan Zhang, and Wanchun Wu. "Sift matching with CNN evidences for particular object retrieval." Neurocomputing 238 (2017): 399-409.

A Framework for Weak Signal Detection in Competitive Intelligence using Semantic Clustering Algorithms

Bouktaib Adil, Fennan Abdelhadi
LIST Department of Computer Science
Abdelmalek Essaadi University
Tangier, Morocco

***Abstract*—Companies nowadays are sharing a lot of data on the web in structured and unstructured format, the data holds many signals from which we can analyze and detect innovation using weak signal detection approaches. To gain a competitive advantage over competitors, the velocity and volume of data available online must be exploited and processed to extract and monitor any type of strategic challenge or surprise whether it is in form of opportunities or threats. To capture early signs of a change in the environment in a big data context where data is voluminous and unstructured, we present in this paper a framework for weak signal detection relying on the crawling of a variety of web sources and big data based implementation of text mining techniques for the automatic detection and monitoring of weak signals using an aggregation approach of semantic clustering algorithms. The novelty of this paper resides in the capability of the framework to extend to an unlimited amount of unstructured data, that needs novel approaches to analyze, and the aggregation of semantic clustering algorithms for better automation and higher accuracy of weak signal detection. A corpus of scientific articles and patents is collected in order to validate the framework and provide a use case for future interested researchers in identifying weak signals in a corpus of data of a specific technological domain.**

Keywords*—Competitive intelligence; apache spark; big data; weak signal detection; web mining; semantic clustering

I. Introduction

In the era of big data, information flows from different sources and in huge volumes. Companies and organizations are under many threats coming from different opponents and competitors. Strategic decisions must be made in order to survive the market changes and cultural, technological, or political shifts that may occur in their environment. Economists rely on the most popular models for strategies to conduct a thorough competitive intelligence activity [1][2] for example : SWOT analysis's main purpose is to analyze threats and opportunities and develop plans to react strategically to those events, this model can be supported by using weak signal detection and early warning signs techniques [3].While PETS model analyzes the data concerning the environment of the company by monitoring political, economic, technological and social factors in order to prepare strategic responses to any change so it can maintain a dominant position in the market. Many organizations invest heavily in developing systems to automate the process of competitive intelligence [4] [5] and implement their adopted strategies. One of the main goals and features of those systems is the detection of weak signals in the environment surrounding an organization. Environmental scanning is gaining the attention of many stakeholders due to the benefits and advantages [6] it brings to the well-being and the contribution to the sustainability of their companies. The aim of most environmental analysts is to detect pieces of valuable information that will give them the strategic advantage of anticipation and early response planning, this can be done through weak signal detection. A weak signal is defined as a temporal change that occurs in a domain or a topic or in the environment in general [7], and it may have an impact on the future and become a trend. Therefore, the early detection and identification of this strategic information is crucial to the evolution of an organization. Many definitions are given to this concept, and different techniques and approaches are applied to detect this kind of information automatically, which is the subject of the next sections.

Companies must be able to understand and explore their environment to extract implicit knowledge that cannot be identified by experts. But it should also be able to predict the future evolution of a specific domain. The emergence of web data and the availability of information online pushes the companies nowadays to exploit these data to extract meaningful strategic information that allows them to make optimal and strategic decisions based on a scientific accurate analysis of the data, and an intelligent approach of web mining [8] to extract high-quality data.

Competitive intelligence systems are software that groups together a set of tools and technologies that companies have to implement in order to keep track of their evolving environment [9]. Many of these solutions neglect the anticipative information model that helps predict and monitor trends that unfold threats and opportunities that must be harnessed and used to gain a competitive advantage.

Weak signals are pieces of information that will help companies to identify threats and opportunities in their environment, which in turn will allow the implementation of an anticipative strategy rather than a reactive one that responds to the events as they happen.

With the rapid stream of data available online and the vast number of documents available on the web every second,

companies must use the latest big data technologies and advanced algorithms in order to process and analyze this data efficiently to identify weak signals [10]. In this paper we propose a framework for weak signal detection in collected data from the web, using big data technologies and aggregation of semantic clustering algorithms based on Apache Spark to detect weak signals and emerging trends and monitor opportunities and threats.

This paper is structured as follows: in section 1, we present the definition of the main concepts of this work: competitive intelligence, SWOT analysis strategy, weak signals detection, competitive intelligence systems, big data analytics, semantic clustering algorithms. Section 2 will present some of the related works that try to handle and propose novel tools and methods of weak signal detection and we will highlight some of their limitations. Section 3 presents the proposed framework and our approach to detect weak signals. Section 4 presents the results of a case study in collected articles about "big data", and we show the results of our approach, then we finish by a discussion and conclusion.

II. Problematic

In order to monitor competitors and identify early warning signs that help decision makers identify companies' key intelligence needs [11], we need a framework for weak signal detection that will allow us to listen to and anticipate the changes in the market [12] by providing an unsupervised manner of analyzing data and capturing potential weak signals that evolve through time.

We define the problem and the importance of our contribution as follows: The main problem is how to process and analyze large amount of unstructured big data automatically from various sources to detect weak signals and unveil some of the strategic information hidden in a large corpus of textual documents.

We use semantic clustering algorithms with an aggregation approach to automate the detection of weak signals that share some characteristics that we defined earlier in the framework and we propose them to the final user domain expert who will then judge their usefulness in a strategic decision or action.

Most solutions do not process a variety of sources and big data, so we will try to propose a framework that is capable of analyzing data coming from multiple sources, and architecture to support the evolution of volume and velocity of data while relying on Apache Spark capabilities and semantic clustering algorithms like LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) and K-Means[13] to give accurate results and high semantically related clusters of terms that may represent a weak signal.

III. Main Concepts

A. Competitive Intelligence

Competitive intelligence is defined as a process, activity, service[14] that starts from the definition of a strategic need problem, passing by the collection of multiple data from different sources, and through the analysis of this data, analysts process the data using their set of tools and techniques to extract strategic information from the data and interpret it to transform it into a usable and a valuable knowledge to be disseminated to the stakeholders, every organization has a different model and strategy to conduct competitive intelligence, which varies depending on the size, the environment or the need of an organization, in order to enhance the decision making process.

The goal of conducting competitive intelligence is to define the position of an organization in the market and to help it be aware of the changes and competitive forces around its environment [15], by providing an organizational tool capable of generating valuable knowledge from raw data to guarantee better business performance by taking strategic actions at the right time [16].

B. SWOT Analyisis and PEST Model

Many models exist to implement competitive intelligence monitoring strategies. Economists proposed models to establish an environment scanning tools to prepare for any change in the market and give an objective perspective of the position of an organization. SWOT analysis [17] focuses on analyzing the strengths and weaknesses of a company through processing internal data, and opportunities and threats coming from external data. When talking about weak signals, we are more interested in analyzing the opportunities and threats coming from the market. The PEST model [18] stands for political, economic, social, and technological factors of an environment that could influence the existence of an organization and its evolution in the market. That external information can be collected and analyzed easily from external web data and exploited in technological intelligence to be able to detect innovation [19], which is present in both structured and unstructured form. The aim of this paper is to analyze big unstructured data using big data analytics technologies and efficient algorithms while respecting and following the main ideas and concepts of those two models, as in Fig. 1.

C. Weak Signals

According to Igor Ansoff [20], weak signals are defined as small changes and imprecise early indications that occur over a period of time on a specific topic that may have an ongoing impact on the future. Weak signals are temporal changes that hold important and strategic information that companies and organizations must detect and collect to stay ahead in the market [21]. This helps them implement an anticipative approach of handling the opportunities and threats present in the market in the form of unstructured data harvested from the web.

The identification of weak signals relies on some characteristics and key points. According to Ansoff weak signals are weakly mentioned in their first appearance, they are less frequent than the main concepts in the context where they exist, but they are new and novel and hold a sign of innovation or a surprise in the market. The interpretation of weak signals requires domain experts in order to contextualize the findings and transform data into knowledge, and classify them as opportunities or threats and disseminate them to stakeholders to make an appropriate strategic decision.

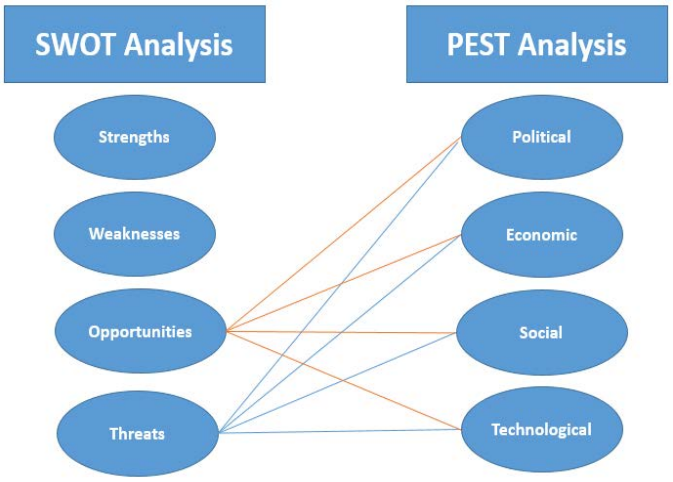


Fig. 1. SWOT Analysis vs PETS Model.

D. Apache Spark

Due to the volume of data available online, data must be collected from different sources in different formats. A homogenization step is mandatory to unify the structure of the data to be collected. Once the data is collected, we end up with huge volumes of data that cannot be processed by a normal computing approach, thus the need for big data analytics technologies that support huge volumes and fast streams of data. Few weak signal detection researchers have proposed a technological framework that addresses the issue of big data. Therefore, we propose in this paper a big data framework for weak signal detection with the implementation of semantic clustering algorithms in Apache Spark.

Apache spark [22] is one of the main big data analytics technologies, and the most well-known platforms for massive distributed computing, that are popular nowadays. This framework is gaining a lot of attention in the big data community and its use in a variety of applications [23] proved to give efficient results when dealing with large datasets. Hence we chose this framework in our attempt to develop a competitive intelligence system [24] to analyze and extract strategic information from the increasing amounts of data available in the environment of companies and organizations.

Apache spark has been used in a lot of applications [25] and it has been used to implement a variety of big data platforms and solutions. Apache Spark is a part of the Hadoop ecosystem introduced in 2009. While Hadoop processing is based on the MapReduce computing paradigm, Spark relies on the DAG (Directed Acyclic Graph) paradigm, which imposes sequential processing of RDDs, a distributed unit of data nodes in the cluster, that optimizes the consumption of resources by avoiding costly data copies used in iterative algorithms that we are going to be using in the weak signal detection framework.

IV. Related Work

Many researchers tried to apply variable methods to detect efficiently weak signals in large volumes of documents [26]. Those methods vary from supervised to unsupervised machine learning methods, automatic and semi-automatic methods, or manual methods relying on experts input, quantitative and qualitative methods, and many data sources were used to prove the approaches and detect weak signals.

One of the early approaches and attempts to discover weak signals in data, Yoon [27] used a keyword-based text mining method to identify opportunities in web news data. He used a quantitative method in which he performed a time-weighted analysis by calculating the occurrence and frequency of keywords during a period of time. But this attempt was limited to only one source of data, and it lacks an automatic crawling of data from multiple sources, and fails when it comes to dealing with large datasets. The result may not be easily interpreted when visualizing a large space of keywords.

El Hadadai.Anass et al [28] proposed a sequence data mining based method for extracting emerging trends and highlighting the evolution of domains through crossing terms with dates and other fields. With the application of correspondence analysis and multiple correspondence analysis and a visualization tool, this approach was able to extract clusters of weak signals from sorting and extracting clusters from the obtained matrix. The method was evaluated using a dataset from scientific articles and patents collected from scientific databases in order to identify technological weak signals, but this method lacks the possibility to be extended to support large datasets and its need for an expert to manipulate the tool to perform the analysis.

D. Thorleuchter et al [29], proposed a methodology based on idea mining and Latent Semantic Analysis to identify weak signals, by constructing a matrix based on the vectors and patterns discovered from the idea mining approach, by applying a dimensionality reduction on the matrix using SVD decomposition, which produces a set of semantically related clusters that may be a weak signal. The method is limited, as stated by the authors. They observed that the method lacked the possibility to discover implicitly cited weak signals and proposed an enhancement using Latent Dirichlet Allocation to get more accurate results.

Antonio.L.et.al [30] proposed a method to conduct an anticipative intelligence by analyzing text and identifying weak signals, using clustering k-medoids and a Jaccard function as a similarity function between obtained clusters in order to analyze similar clusters of weak signals, the method claims to be automatic but the dataset is collected from experts at the beginning of the process.

Julien Maitre et.al [31], the work that is closely related to what we are proposing is inspired by this paper, which presents a novel approach for weak signal detection in weakly structured data or unstructured data, by combining Latent Dirichlet Allocation and Word2Vec algorithm to perform clustering on a corpus of documents collected from the web, the article proposes also a method to identify the number of clusters k to be extracted from a corpus using LDA, which in most cases is hard to define and is crucial to the quality and robustness of the obtained results especially when it comes to weak signals, where the use of a small k may eliminate the identification of important weak signals.

In our approach, we try to group the three algorithms in order to reduce the mistakes and weakness of those approaches

by using a clustering aggregation [32] approach supported by the computational power of Apache Spark and the flexible nature of RDDs and their reusability in iterative algorithms in order to perform multiple tasks, and with using the ML pipeline feature of Apache Spark to facilitate and automate the process of weak signal identification with a minimum interaction of experts.

V. Proposed Framework

In light of the findings of the literature review conducted by C. Muhlroth et.al [26] and other reviewed approaches [33] [34] [35] [36], we found a need to propose a big data analytics framework for automatic weak signal detection. Thus we propose in this paper a framework that uses Apache Spark to implement the data analysis from data collection to weak signal identification using semantic clustering algorithms. The feature of Apache Spark that allows us to achieve this is the ML pipeline that aims at automating steps to be applied on a dataset, in order to extract implicit hidden information that may present key strategic indications to be processed and analyzed.

Fig. 2 presents the architecture of the proposed framework. It outlines the steps to be followed in the pipeline implemented using Apache Spark, starting from data collection to the identification of weak signals contained in the corpus of collected documents. In the following section, we provide a brief explanation of Apache Spark ML pipeline, and we explain the steps of the pipeline in detail.

A. Apache Spark DAG and ML Pipeline

Apache Spark provides an API to manipulate RDDs, resilient distributed datasets, which is a good structure for dealing with big unstructured data. The power of this data structure remains in the possibility to expand to huge volumes of data, thus the adoption of this technique in our work. RDDs will hold the corpus data to perform analysis using ML pipeline API that represents a set of processes to perform on a dataset to get the desired results. This makes it easier to aggregate multiple algorithms into a single pipeline. We will be using this technique in our work to implement an efficient big data analytics framework for weak signal detection, by combining the semantic clustering algorithms presented in Fig. 2.

The technology that allows Apache Spark to execute such processing is DAGs which is a new enhanced strategy to perform map-reduce tasks, as shown in Fig. 3, by organizing the planning of execution in stages and steps that form a directed acyclic graph of transformations to apply on the dataset.

All the algorithms used in this framework will be implemented using the Apache Spark MLlib library that contains a variety of tools and machine learning algorithms and clustering to be applied on the data. We combine LDA and implement LSI and K-means by using ML Pipeline to perform semantic clustering on the corpus of collected data. At the end, we communicate the findings to the stakeholders and experts to identify the clusters that hold potential weak signals.

B. Data Collection

The framework starts with data collection. We collect data from multiple scientific articles databases and patents and store them in the Hadoop file system. When we start the execution of the ML pipeline we load the data from Hadoop onto the Apache spark cluster in order to execute the outlined pipeline process depicted in Fig. 2. Scientific databases from IEEExplorer, ACM Digital, and patents from USPTO, contains many articles and documents having a lot of fields like text, date, abstract, publication date, etc. We are interested in the text and publication date of the document in order to conduct a temporal analysis of weak signals and the evolution of the topic in time to perform technological surveillance on a specific field of interest.

Many scrappers and crawlers are developed to collect data from those websites using web mining methods and Scrapy python framework [37], which gives the possibility to create scraping agents to crawl as many web pages as possible with the elimination of repeated documents. Our approach helps decision-makers and analysts to collect data automatically and conduct environmental scanning with no need for manual intervention, which could be a hard task for companies in this era of big data.

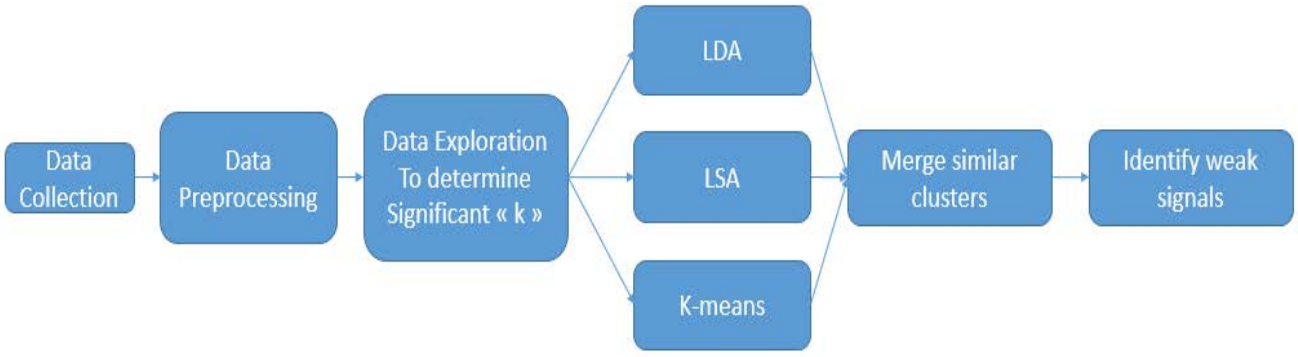


Fig. 2. Proposed Framework Architecture.

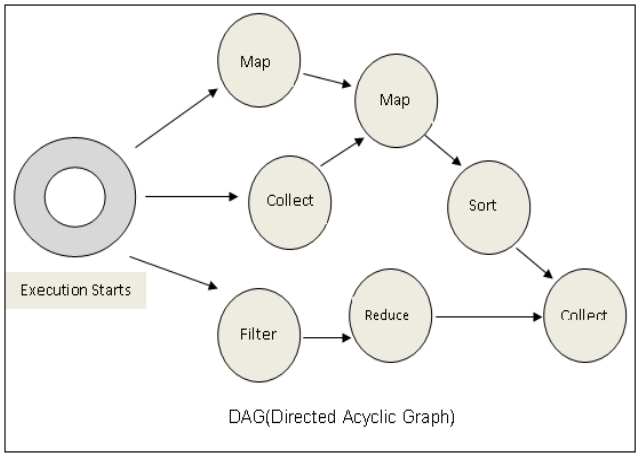


Fig. 3. DAG Excution Method in Spark.

C. Data Preprocessing

Preprocessing is an important step to clean data and format it to our needs. After choosing the text field we will use in analysis and the date field that will help us to filter emerging trends, we clean the text from ambiguous characters, then by removing StopWords, stemming and lemmatization, which will help us to get more accurate results and easily interpretable information from raw data. We create n-grams from the corpus to add them to the vocabulary of the corpus. This step is important to enable the clustering algorithms to identify multi-terms that may hold an important part of a weak signal, especially in the scientific field. Due to the nature of weak signals, which is low frequency and occurrence of words, we eliminate terms where the count is above a threshold, for example 200 occurrences, as we are not interested in highly frequently mentioned words that, in most cases, represent strong signals or trends, which are not the purpose of our analysis.

D. Data Exploration

The number of topics to be extracted cannot be determined previously as the algorithms used are unsupervised algorithms and the analyst does not have an idea about the number of clusters to be obtained. Therefore, we choose a rule of thumb and we define the number of clusters to be extracted as in eq.1 after the extraction of the vocabulary from the corpus:

$$k = \sqrt{n/2} \quad (1)$$

where n is the number of words in the vocabulary of the corpus.

The determination of an approximate k is an important step in the process of this pipeline. We can specify k based on many techniques of data exploration or using many methods from the literature [38], which is outside the scope of our research, or we can try different values of k and analyze the different clusters obtained. A small number of k though must be avoided in order to avoid the elimination of important potential weak signals that are not heavily cited.

E. LDA

Latent Dirichlet Allocation [39] is a generative probabilistic model for text classification and a topic modeling algorithm that aims at representing the documents as a set of topics, with the objective of assigning each term to a semantically related topic. When applying LDA in Fig. 4 to a corpus of documents, the algorithm tries to cluster the topics and their related terms according to their semantic relationships. It identifies k topics, k is a number specified by the analyst, many methods exist to choose the best k that gives accurate clusters.

LDA algorithm steps are defined as follows, for each document w in a corpus D:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir(α).
3. For each of the N words wn:
   (a)Choose a topic zn ~ Multinomial(θ).
   (b)Choose a word wn from p(wn |zn,β), a multinomial probability conditioned on the topic

Those steps are the standard for the LDA model, in order to cluster a distribution of semantically related words to a set of specific topics, in our case those topics may represent innovations, opportunities or threats.

So we will use this algorithm to detect underlying topics in a corpus of documents. Those topics may include weak signals that are not easily identified and are not in the scope of the knowledge of experts, especially in the case of new innovations in a domain. After removing the most frequent terms from the documents, we aim at identifying sets of words that are less frequent and semantically related and belong to the same topic. That's why a maximal number of k is essential to the extraction of latent topics that represent a small proportion of the document, which is the nature of weak signals defined by Ansoff.

F. LSA

Latent semantic analysis [40] is a text mining technique that aims to create a semantic space to identify relationships between words in a corpus of documents. Those relationships are semantically detected using a linear algebra technique called SVD decomposition. Its goal is to decompose a document-term matrix created from the corpus into a lower-dimensional space in order to detect close words and extract coherent topics and similarities between documents.

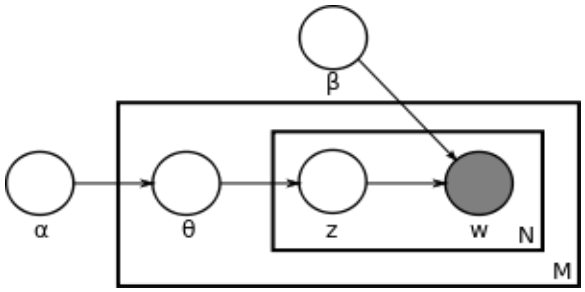


Fig. 4. Plate Notation Visualizing the LDA Model Parameters, Plate M Represent The Total Number of Documents, N Represent the Numbers of Terms in a Document, α the Per-Document Topic Distributions, β the Per-Topic Word Distribution, θ the Topic Distribution For Document m, z the Topic for the n-th Word, w is a Specific Word.

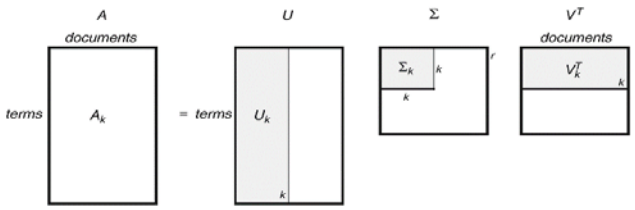


Fig. 5. LSA Algorithm Matrix Decomposition Process.

By applying this technique in weak signal detection we want to detect weak clusters that are appearing in the corpus, those highly coherent and newly emerging clusters may hold important strategic information, they may represent an opportunity for investment and collaboration, or a threat that a company has to plan a strategic response to face it and overcome its consequences.

After the creation of the matrix from the collected corpus by crossing the terms with their corresponding documents, we create an m*n matrix where m is the number of terms and n is the number of documents, then we apply SVD which will decompose the matrix into 3 new matrices as depicted in Fig. 5.

$$M = U\sum V^* \quad (2)$$

Where U is an m*k matrix that holds the word assignment to topics, Σ an k*k matrix which contains singular values that represent the importance of the topic, V* is an k*n matrix that contains the topic distribution across documents. In our case, we are interested in the first two matrices. By crossing the pairs of vectors of the two matrices, we obtain the clusters of topics and their corresponding terms. The clusters obtained in this step will be merged with the previous results to enhance the semantic understanding of the corpus.

G. K-MEANS

K-means is one the most important algorithms for clustering data, the power of this approach resides in its ability to perform unsupervised learning and clustering of data with no prior knowledge, hence the choice of this algorithm in our process to enhance the results of our approach and the quality of the obtained clusters.

As in the previous algorithms, we perform data preprocessing depicted in the previous section before applying k-means. To apply k-means we follow those steps in order to find semantically related terms, relying on the Word2Vec [41] model, and group them in cluster as depicted in Fig. 6:

- Cleaning and Preprocessing of text.
- Determination of number k.
- Feature extraction using Word2vec to represent each word semantically as a vector.
- Applying k-means.
- Getting clusters.

In the next step, we will merge the most similar clusters to unify and expand the clusters that are candidates to be weak signal clusters, containing information about potential opportunities or threats that must be noticed.

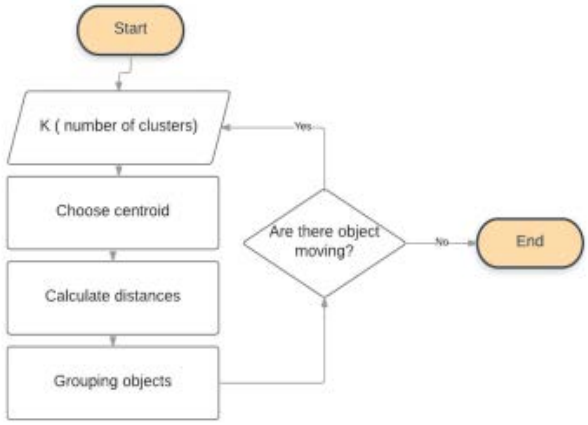


Fig. 6. K-Means Algorithm Steps.

H. Cluster Aggregation

Cluster aggregation [32] is a method that aims to apply different clustering algorithms on the dataset and find a consensus about the optimum cluster groups in order to eliminate duplicates, and eliminate the noise of each algorithm if it was applied individually, in order to improve the quality and robustness of the clustering.

After extracting the clusters from the previous steps, denoted C1, C2 and C3 from applying LDA, LSA and K-means respectively, we move to the merge step which consists of performing a similarity calculation between all pairs to identify similar clusters and merge them in order to eliminate redundancy and enhance the quality of the weak signals detection process by minimizing the disagreements between clusters according to Equation 4.

$$D(C) = \sum_{i=1}^{m} d_V(C_i, C) \quad (3)$$

where v is a set of words or multi-terms and m is the number of all clusters from the applied algorithms.

The implementation of Approximate Similarity Join of Apache Spark MLLib is used, which is based on the Jaccard similarity function eq (4). We calculate it for each pair of all clusters from all algorithms, and if it passes a threshold, we merge the clusters into one, in order to get the cluster that minimizes the number of disagreements.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

The resulted clusters from Algorithm 1 are shared with experts and stakeholders to identify potential weak signals from the corpus.

I. Weak Signal Identification

Extracted clusters will pass by the last step, which aims at calculating a score that represents the weighted term evolution inspired by Yoon et.al [27], the evolution rate “er*” of each term during a period t of the cluster “Ci” is calculated and the sum “eri” of all terms represent the score of a cluster, based on that score we can identify the clusters that may hold weak signals represented by semantically related terms from the corpus.

We order the clusters by their score, and based on that score and the interpretation of an expert in a domain, we can spot the clusters that are holding information about the weak signal, which, by interpretation, may be a threat, an opportunity of investment, or an innovation that needs further investigation or collaboration.

In the next section we present the results of applying this flow to the collected corpus, and we discuss the obtained results, advantages and limitations of our framework and we conclude with ideas for future researchers.

```
Algortihm 1: MergeClusters
Input: clusters [C1, C2, C3]
Output: megred clusters [Cm]
BEGIN
    For each pair of clusters:
    Calculate similarity
    If similarity > threshold
        DO merge_clusters ()
RETURN merged_clusters
```

VI. Result and Disccusion

In order to evaluate the proposed method, we will conduct an analysis on a dataset of scientific articles about "big data" topic, we collected a corpus of 5800 documents and scientific articles about « Big Data » containing multiple fields, from the fields we are interested in are abstract field and publication date. We will perform text mining on the text field and perform growth analysis using the publication date.

The purpose of our analysis is to perform the clustering aggregation of three algorithms, K-means, LDA, and LSA in order to combine the results of each algorithm and select from the obtained clusters the ones that are potential weak signals and may hold information about opportunities or threats.

A. Data Collection

We collect data from IEEExplorer, ACM Digital Library, SpringerLink and Sciendirect to show a case study and illustrate the processes of our approach. We use the search query "big data" and choose a publication date range from 2000 to 2020, then we collect the documents and articles published in this range of time. We are interested in the abstract, title and publication year of a document as in Fig. 7, the scraper agents extract those fields using the CSS styling of each database website in order to ease the step of homogenization of those fields in the next step.

We create a data frame from the documents containing the three fields we are interested in, and process them in the remaining steps of the framework. Fig. 8 shows an extract of the collected data.

Fig. 7. IEEEXplorer Document Example.

Document Title	Abstract	Publication Year
A survey of data ...	Computer clusters...	2020
A Novel Data-Driv...	Data-driven appro...	2018
A novel clusterin...	Big data analytic...	2019
A Framework for B...	The emergence of ...	2019
Mining conditiona...	Current Condition...	2020
Social Set Analys...	Current analytica...	2016
Big data oriented...	Due to the tremen...	2018
Big Data, Big Kno...	The idea that the...	2015
Protection of Big...	In recent years, ...	2016
Big Data Analytic...	Mobile cellular n...	2016
Big Data Analytic...	Big data analytic...	2019
Big Data-Based Im...	Big data-based ac...	2019
An Integrated Met...	The expand trend ...	2019
A Methodology of ...	The traditional b...	2018
Analysis and Visu...	With the developm...	2019
Evaluating the Qu...	The use of freely...	2015
A Novel Online an...	A sizable amount ...	2017
Data Lake Lambda ...	The advances in s...	2018
Big data analytic...	In recent years, ...	2019
A Big Data Mining...	In recent years, ...	2019

only showing top 20 rows

Fig. 8. Dataframe Collected from the Scrapping Agents of the Framework.

B. LDA Obtained Clusters

After the preprocessing and cleaning step of the articles obtained about big data, we perform the first clustering algorithm, LDA, to obtain k clusters from the corpus. The topics obtained are semantically related and clustered in one group.

A sample of clusters obtained from the corpus is presented (Table I):

TABLE I. List of Obtained Clusters from LDA

Cluster	terms	
Topic1	new concept secure communication collection data imbalanced data method consists probability distribution time big	experiments carried nir fmt proposed algorithms location data error rmse minimize total intermediate pointers
Topic2	low power performance overhead set data model predict clustering method method paper features paper	non uniform information big learning architectures cold start tasks cloud signal quality address challenge
Topic3	traffic data information extraction conducted evaluate key challenges network operators computational efficiency selection strategy	characteristics data secure communication model based factors influence change detection processing time human interaction
Topic4	important information processing techniques video analysis word embeddings enhance performance deep hashing physical systems	linear programming existing work data intensive iot networks control mechanism stream processing cover problem

C. LSA Obtained Clusters

The application of latent semantic analysis is done. After applying LSA we obtain a different set of k clusters using the matrix decomposition of singular values (SVD). For each cluster, we select a set of terms that represent this concept and that are closely related to it using the singular values in the sigma matrix.

By applying the LSA on our corpus of data we obtain the following clusters (Table II):

TABLE II. LIST OF OBTAINED CLUSTERS FROM LSA

Cluster	terms	
Topic1	defect detection data fusion sar image detection method power supply tree boosting digital twin	high performance smart manufacturing data digital key management parking lot fabric defects address problem
Topic2	trajectory data lane changing social networks data driven data processing deep neural reinforcement learning	changing model spatial temporal uav bss onset date modis derived brain health health quality
Topic3	proposed model extensive experiments multiplicative linguistic uncertain multiplicative location privacy decision making city brain	cloud computing compared state massive datasets data processing experiments conducted communication consensus group decision
Topic4	point cloud ant colony time delay electric power neutrosophic cubic point clouds dense point	security threats power data uwan security power systems algorithm based cloud generation healthcare insurance

D. K-Means Obtained Clusters

The application of k-means results in a set of k clusters after the calculation of word2vec of the text to create a feature of semantically related words. This was used as the measure of similarity between words or terms to perform semantic clustering. The following clusters were obtained from applying this algorithm:

In the next step, we will try to merge similar clusters into one cluster and build a cluster group that collects the power of all the algorithms and solves the problems and weaknesses of the other approaches (Table III).

TABLE III. LIST OF OBTAINED CLUSTERS FROM LSA

Cluster	terms	
Topic1	big personal personal data hidden transition process model industrial internet open data redundant tight	data big results indicate data processing attack path subgraph matching smart cue value data
Topic2	health big data attracted data frameworks process big computing data processing architecture analysis big dimensional big	industrial big data present processing analysis hot research data techniques lte network statistical analysis data problem
Topic3	network models network model networks cnns based convolutional recurrent neural learning based	deep convolutional compared state vector machine networks cnn outperforms state network based'
Topic4	data analysis bda applications data collected paper present smart cities applications cloud incomplete information	rare events wireless networks public key things iot based data driving range entropy loss

E. Aggregation Algorithm Obtained Clusters

By applying the approximate join similarity, we get the pairs of similar clusters, by merging similar clusters we get p clusters $p < k*3$, which gives an idea about overall clustering and solve the mistakes that could have been made by using one individual algorithm, the obtained clusters represent all the small topics and semantically related terms that may hold an opportunity or a threat (Table IV).

TABLE IV. IDENTIFIED SIMILAR CLUSTERS

Cluster id	Cluster id	Similarity distance
33	30	0.307
36	33	0.307
23	36	0.428
25	6	0.428
23	1	0.428
6	25	0.444
36	23	0.444
23	24	0.461
24	23	0.461
1	23	0.473

We merge similar clusters into one in order to eliminate redundant clusters and improve the quality of visualization.

In order to visualize the results of the approach, we create a graph from the adjacency matrix term-topic and plot the graph using Gephi to see the clusters and the relationships between them. The graph obtained contains 670 nodes and 12 006 edges. We show an extract of the graph in Fig. 10, and an identified weak signal in Fig. 9 containing semantically related words about the application of big data in health.

F. Interpretation and Discussion

From the interpretation of the results, we can spot the weak signals and hidden information that are not visible to the experts, and by combining their expertise with results obtained, we can identify clusters that are potential strategic information holders and we should cross the data back to the original document for further analysis and understanding of the context of appearance and the identification of the importance of the discovered piece of information.

In our approach we filtered weakly cited words in a specific time, year of publication, from the corpus and applied three semantic clustering algorithms in hope of finding the most accurate clusters by using an aggregation method. Those obtained clusters may contain pieces of information that is crucial to the implementation of an anticipative strategy of an organization. A weak signal is characterized by the evolution of its presence or its number of occurrences through time, which makes it a strong signal in the future, though not all weak signals are destined to be strong.

In Fig. 9 we present the graph representation of filtered words from the corpus, those words are related by their co-existence in the same document and their appurtenance to the same cluster. In Fig. 10, we singled out a cluster so we can study the semantics of this potential weak signal with the help of a domain expert.

We see in Fig. 10 that the semantic cluster of topic 32 is weakly cited and highly rated in the last period of research, which means that this low visibility cluster may be a trend in the future, though we can comment on the choice of number k, which must be chosen wisely and we must experiment with different values of k, or we have to use a different algorithm to determine the optimal value of k that will give promising and accurate results.

Extracted Potential Weak signals must be harnessed to identify threats and opportunities in the market. Our method extracts the most promising clusters of weak signal topics. Using our approach and with expert intervention, we can spot the key information that will generate value for organizations. Though the advantage of this method is not to predict which weak signal will become strong, but to enhance the quality of extracted clusters from the corpus, so we can keep and analyze only the semantic clusters holding potential weak signals through the aggregation of three algorithms: LDA, LSA, and k-means, this approach will not eventually predict which one will become strong in the future. In order to predict whether a weak signal will become strong, we require labeled data, and with the application of supervised machine learning [42], we can extract the features of weak signals that are candidates to be strong and trend in the future.

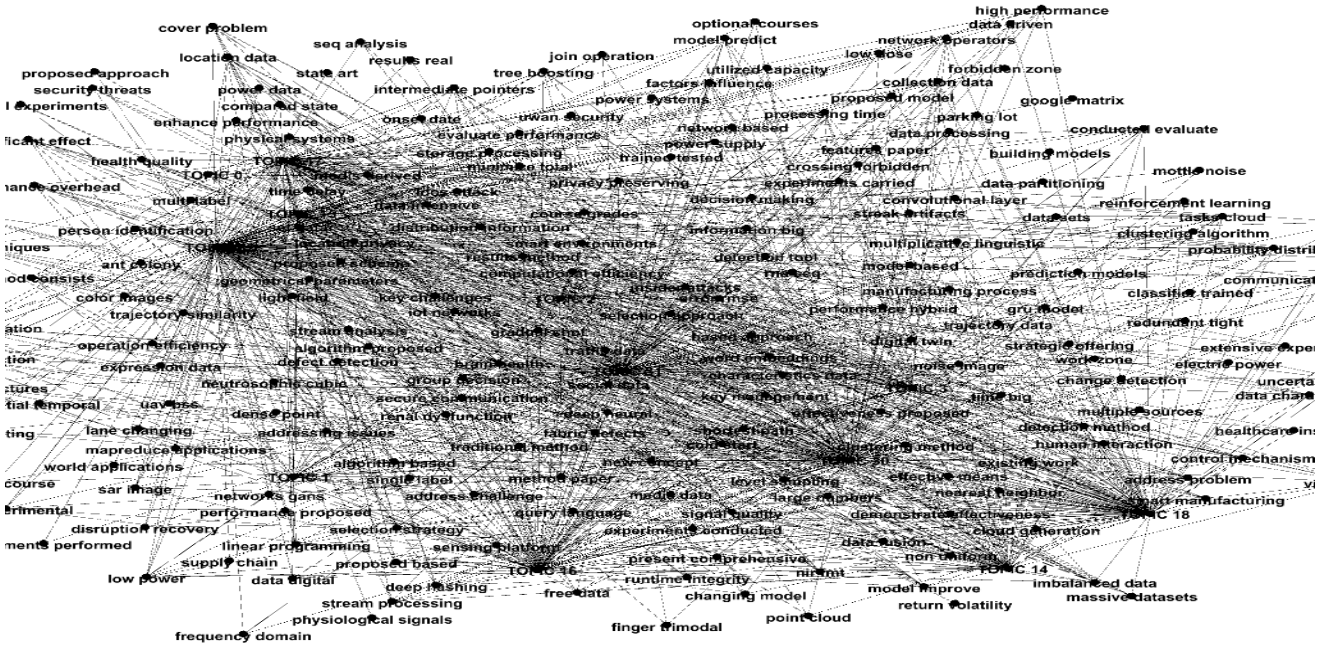


Fig. 9. Network of Collected Data from the Scrapping Agents of the Framework.

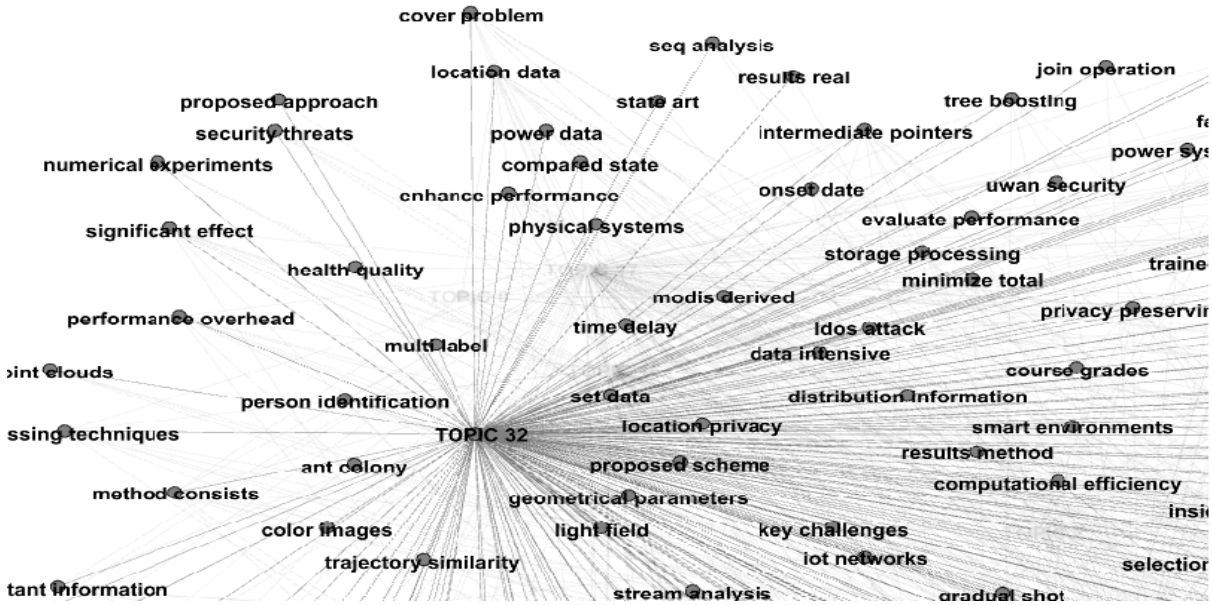


Fig. 10. Illustration of an Extracted Weak Signal Topic to be Analyzed (Topic 32).

VII. Conclusion

In this paper, we proposed a novel framework that uses semantic clustering aggregation models, made possible with the use of the computation power of Apache Spark, through the use of ML pipeline which gives the possibility of automating the process of weak signal detection in large volumes of data. The aim of aggregating clustering methods is to combine the features extracted from each method and hide its weaknesses. The use of such a tool in business will allow businesses and stakeholders to remain active and alert in the market. Semantic clustering methods have proven to be very efficient when it comes to topic modeling and extraction of variable topics semantically related from a large corpus of documents. Hence, we judge this approach to be very pragmatic in nature and it does contribute to the domain of weak signal detection.

Our framework will help stakeholders identify and prepare scenarios of intelligence needs. For each potential weak signal, there must be a strategic response ready to tackle it, which will help stakeholders implement an anticipative approach to conduct strategic competitive intelligence in a big data context, where manual extraction and analyzation of documents is impossible in this era where new data is available every millisecond.

Despite our framework contribution in the field we still think that there is more work to be done for future researchers in weak signal detection literature, for example the nature of data to be analyzed in weak signal detection research is unstructured, thus the need for more advanced clustering methods to perform unsupervised machine learning to label data as weak signals from the past data, and apply Text mining Deep Learning models [43] in order to be able to extract and identify weak signals in future data once available online, which will give a competitive advantage for organizations. In our future work we will apply Graph embedding technique [44] [45] as a technology that will allow us to reduce the dimensionality of the corpus and facilitate the semantic representation of weak signals, through the study of dynamic graph embedding to monitor the evolution of a domain terminology through time, in hope of detecting innovation, opportunity or a threat as early as possible.

In conclusion, we must mention the limitation of methods and approaches to validate the extracted weak signals in most of the literature [46]. As a future research project, we can propose a new direction of research in this field through adopting novel semantic clustering algorithms that rely on deep learning like Word2Vec and Glove word embedding [47] for a more precise semantic analysis of the corpus, and proposing novel approaches that relies on labeled data.

References

[1] Fleisher, Craig S., and Babette E. Bensoussan. Strategic and competitive analysis: methods and techniques for analyzing business competition. Upper Saddle River, NJ: Prentice Hall, 2003.

[2] Prescott, J. F., & Miller, S. H. (Eds.). (2002). Proven strategies in competitive intelligence: lessons from the trenches. John Wiley & Sons.

[3] Popescu, F., & Scarlat, C. (2015). Limits Of Swot Analysis And Their Impact On Decisions In Early Warning Systems. SEA: Practical Application of Science, 3(1).

[4] Sauter, Vicki L. "Competitive intelligence systems." Handbook on Decision Support Systems 2. Springer, Berlin, Heidelberg, 2008. 195-210. DOI: 10.1007/978-3-540-48716-6_10.

[5] Echternacht, Tiago Henrique de Souza, et al. " Competitive Intelligence: A Study Of The Involvement By Senior Corporate Management At Redepetro Companies." (2020), UFRN Electronic Journals Portal. http://rebacc.crcrj.org.br/handle/123456789/5179.

[6] Babatunde, Bayode O., and Adebola O. Adebisi. "Strategic Environmental Scanning and Organization Performance in a Competitive Business Environment." Economic Insights-Trends & Challenges 64.1 (2012).

[7] Ansoff, H.I. Managing Strategic Surprise by Response to Weak Signals. Calif. Manag. Rev. 1975, 18, 21–33. https://doi.org/10.2307/41164635.

[8] Sewlal, R. (2004). Effectiveness of the Web as a competitive intelligence tool. South African Journal of Information Management, 6(1), 1-16. DOI: 10.4102/sajim.v6i1.293.

[9] Sauter, Vicki L. "Competitive intelligence systems: Qualitative DSS for strategic decision making." ACM SIGMIS Database: the DATABASE for Advances in Information Systems 36.2 (2005): 43-57. https://doi.org/10.1145/1066149.1066154.

[10] D. Thorleuchter, T. Scheja, and D. Van den Poel, "Semantic weak signal tracing," Expert Systems With Applications, vol. 41, no. 11, pp. 5009–5016, 2014. DOI:10.1016/j.eswa.2014.02.046.

[11] Herring, J. P. (1999). Key intelligence topics: a process to identify and define intelligence needs. Competitive Intelligence Review: Published in Cooperation with the Society of Competitive Intelligence Professionals, 10(2), 4-14. https://doi.org/10.1002/(SICI)1520-6386(199932)10:2%3C4::AID-CIR3%3E3.0.CO;2-C.

[12] El Akrouchi, M., Benbrahim, H., & Kassou, I. (2020). Early warning signs detection in competitive intelligence. In Proceedings of the 25th International Business Information Management Association Conference—Innovation Vision (pp. 1014-1024).

[13] Ma, Junhong. "Improved K-Means Algorithm in Text Semantic Clustering." (2014). DOI: 10.2174/1874110X01408010530.

[14] Wolter K. (2011) Competitive Intelligence. In: Keuper F., Oecking C., Degenhardt A. (eds) Application Management. Gabler.

[15] Hall, C., & Bensousson, B. (2007). Staying ahead of the competition: How firms really manage their competitive intelligence and knowledge; evidence from a decade of rapid change. New York: World Scientific Publishing Data. https://doi.org/10.1142/6669.

[16] Charity, A. E., & Joseph, I. U. (2013). Manage competitive intelligence for strategic advantage. European Journal of Business and Management, 5(3), 1-9.

[17] Gurel E, Tat M. SWOT analysis: a theoretical review. J Int Soc Res. 2017;1051:994–1006. https://doi.org/10.17719/jisr.2017.1832.

[18] Singh, S.S. (2013). Environment & PEST Analysis : An Approach to External Business Environment, International Journal of Modern Social Sciences.

[19] Veugelers, M.; Bury, J.; Viaene, S. Linking technology intelligence to open innovation. Technol. Forecast.Soc. Chang. 2010, 77, 335–343. https://doi.org/10.1016/j.techfore.2009.09.003.

[20] Holopainen, M., & Toivonen, M. (2012). Weak signals: Ansoff today. Futures, 44(3), 198–205. https://doi.org/10.1016/j.futures.2011.10.002.

[21] Keping, W. (2009). Research on the Enterprise Crisis Early Warning System Based on Competitive Intelligence [J]. Information Studies: Theory & Application, 12.

[22] Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.;Franklin, M.J.; et al. Apache Spark: A Unified Engine for Big Data Processing. Comm. ACM 2016, 59, 56–65. DOI: 10.1145/2934664.

[23] A. Alexopoulos, G. Drakopoulos, A. Kanavos, Ph. Mylonas, G. Vonitsanos, "Two-Step Classification with SVD Preprocessing of Distributed Massive Datasets in Apache Spark", Algorithms, MDPI, March 2020. https://doi.org/10.3390/a13030071.

[24] B. Adil, F. Abdelhadi, B. Mohamed and H. Haytam, "A Spark Based Big Data Analytics Framework for Competitive Intelligence," 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 2019, pp. 1-6. DOI: 10.1109/ICSSD47982.2019.9002837.

[25] Salloum, S., Dautov, R., Chen, X. et al. Big data analytics on Apache Spark. Int J Data Sci Anal 1, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9.

[26] Mühlroth C, Grottke M (2018) A systematic literature review of mining weak signals and trends for corporate foresight. Journal of Business Economics 88(5):643–687.

[27] Yoon, J. Detecting Weak Signals for long-term business opportunities using text mining on Web news. Expert Syst. Appl. 2012, 39, 12543–12550. DOI: 10.1016/j.eswa.2012.04.059.

[28] El Haddadi, A., Dousset, B., & Berrada, I. (2012). Discovering Patterns in Order to Detect Weak Signals and Define New Strategies. In P. Kumar, P. Krishna, & S. Raju (Eds.), Pattern Discovery Using Sequence Data Mining: Applications and Studies (pp. 195-211). Hershey, PA: IGI Global. doi:10.4018/978-1-61350-056-9.ch012.

[29] Thorleuchter, D., & Van den Poel, D. (2015). Idea mining for web-based weak signal detection. FUTURES, 66, 25–34. DOI: 10.1016/j.futures.2014.12.007.

[30] MOREIRA, A. L. M. ; HAYASHI, T. W. N. ; COELHO, G. P. ; SILVA, A. E. A. . A Clustering Method for Weak Signals to Support Anticipative Intelligence. International Journal of Artificial Intelligence and Expert Systems, v. 6, p. 1-14, 2015.

[31] Julien Maitre, Michel Menard, Guillaume Chiron, Alain Bouju, "Détection de signaux faibles dans des masses de données faiblement structurées", Recherche d'information, document et web sémantique, ,vol 3, no.1. doi:10.21494/ISTE.OP.2020.0463.

[32] Gionis, A., Mannila, H., and Tsaparas, P. 2005. Clustering aggregation. In Proceedings of the 21st International Conference on Data Engineering (ICDE) (Tokyo, Japan). https://doi.org/10.1109/ICDE.2005.34.

[33] Youngjung Geum, Jeonghwan Jeon & Hyeonju Seol (2013) Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing, Technology Analysis & Strategic Management, 25:1, 1-22, DOI: 10.1080/09537325.2012.748892.

[34] Garcia-Nunes, P.I., & Silva, A.E. (2019). Using a conceptual system for weak signals classification to detect threats and opportunities from web, Futures 107(March 2019):1-16. 10.1016/j.futures.2018.11.004.

[35] Sahbi Sidhom, Philippe Lambert. "Information Design" for "Weak Signal" detection and processing in Economic Intelligence: case study on Health resources. 4th International Conference on Information Systems and Economic Intelligence - SIIE'2011, IGA Maroc, Feb 2011, Marrakech, Morocco. pp.315-321.

[36] Xianjin, Z., & Minghong, C. (2010). Study on early warning of competitive technical intelligence based on the patent map. Journal of Computers, 5(2), 274-281. DOI: 10.4304/jcp.5.2.274-281.

[37] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454. DOI: 10.1109/ICECA.2019.8822022.

[38] Patil, C., Baidari, I. Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth. Data Sci. Eng. 4, 132–140 (2019). https://doi.org/10.1007/s41019-019-0091-y.

[39] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (March 2003), 993–1022.

[40] Thomas K Landauer, Peter W. Foltz & Darrell Laham (1998) An introduction to latent semantic analysis, Discourse Processes, 25:2-3, 259-284, DOI: 10.1080/01638539809545028.

[41] Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space." CoRR abs/1301.3781 (2013): n. pag.

[42] Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317. DOI:10.1613/jair.1.12228.

[43] Chen, Liang, et al. "A deep learning based method for extracting semantic information from patent documents." Scientometrics 125.1 (2020): 289-312. https://doi.org/10.1007/s11192-020-03634-y.

[44] Wang, Yunli, René Richard, and Daniel McDonald. "Competitive Analysis with Graph Embedding on Patent Networks." 2020 IEEE 22nd Conference on Business Informatics (CBI). Vol. 1. IEEE, 2020. DOI: 10.1109/CBI49978.2020.00009.

[45] Goyal, Palash, and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." Knowledge-Based Systems 151 (2018): 78-94. DOI:10.1016/j.knosys.2018.03.022.

[46] Xiao, W. K. F. X. Z., & Xinyan, L. (2011). Enterprise Competitive Intelligence Crisis Early Warning Review [J]. Journal of Modern Information, 7. DOI: 10.3969/j.issn.1008-0821.2011.07.042.

[47] M. Mohammed, Shapol et al. "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms." Indonesian Journal of Electrical Engineering and Computer Science 22 (2021): 552-562. DOI:10.11591/IJEECS.V22.I1.PP552-562.

Virtual Reality Simulation to Help Decrease Stress and Anxiety Feeling for Children during COVID-19 Pandemic

Devi Afriyantari Puspa Putri[1], Ratri Kusumaningtyas[2], Tsania Aldi[3], Fikri Zaki Haiqal[4]
Informatics Engineering Department[1, 3, 4]
Communication Sciences Department[2]
Universitas Muhammadiyah Surakarta[1, 2, 3, 4]
Surakarta, Indonesia[1, 2, 3, 4]

***Abstract*—The occurrence of COVID-19 pandemic has changed people's life in every aspect, such as applying social distancing, the transition from offline to online activity are applied in order to decrease and stop the spread of the virus. This sudden change causes a fairly high level of anxiety and stress in society, especially for children because of activity restrictions. Various innovations, especially technology have been carried out to overcome the problems in distance restrictions that have arisen due to the COVID-19 pandemic. Virtual reality believes becoming one of the innovations that can be used to reduced anxiety levels and boredom during activity rectrictions, because it creates an artificial environment for humans to socialize. In this research, combine the Unity3D and blender software to build a virtual reality simulation with the help of virtual glassed to give a real impression of the virtual room that has been created. This VR application consists of three environments that children can use it to explore the virtual room without need being in crowded atmosphere. Based on the result of pretest and posttest questionnaire in 30 participants with the range age from seven to ten, it concludes that this VR applications can decrease the level of stress and anxiety in children by one to two levels. Besides that, this application located in acceptable area based on SUS score system.**

***Keywords*—Blender; COVID-19; Unity3D; virtual reality**

I. Introduction

In the last decade, it can be seen that the development of information technology has grown rapidly, and has contributed to advancing world civilization, including the creation of borderless world especially for the access of information. One of the hot topic in tehcnological development field is Virtual Reality (VR), which is the creation of a technological simulation resulting from a three-dimensional (3D) environment [1][2]. Along with the development of increasingly advanced technology, there are many changes that occur in the world, one of the remarkable changes is the transfer from all offline activities to online due to the coronavirus disesase 2019 (COVID-19) pandemic.

The COVID-19 pandemic began to emerge at the end of 2019 in the city of Wuhan, Hubei Province, China and began to attract a lot of world attention [3]. Recently, COVID-19 has infected more than 89 million people in 218 countries [4], due to massive transmission cause many countries implementing new regulations to stop the spread of COVID-19. One of the effective way to minimalize the spread of COVID-19 to implement a lockdown which has an impact on restrictions and the closure of public access that has the potential to cause crowds, including : school closures, restrictions on shopping places and tourist attractions. This has a fairly serious impact in all aspects of life [5].

Based on research [6],[7] there are many psychological impacts that have arisen due to the COVID-19 pandemic, including increased levels of anxiety and stress and even more severe effects if people have to undergo self-quarantine. In this era of social restrictions, technological advances are considered to be able to help reduce the negative effects caused by COVID-19, as well as become a solution in the current massive social restrictions. This argument also pointed out in [21] which highligted that digital approaches, such as AR and VR should be optimised in order to help children mental and behaviour during this pandemic era. In line with research [5] that utilizes technological advances to conduct distance learning, or the use of augmented reality (AR) applications as tutorials related to effective hand washing in order to prevent COVID-19 infection [8]. Another research on AR [20] has a conclusion that the use of AR application improve the effectivenes of training while conduct on stressful environment.

In addition, one of the technological advances that can be utilized in overcoming the COVID-19 crisis is the use of VR technology that can help overcome current health problems [9]. In the study [10] discussed the relationship between the use of VR technology that can help reduce pain in children during burn replacement. Several research [19][22] has been conducted, and gain conclusion that VR, and AR have a significant effect to reduce or prevent stress during COVID-19 Pandemic. It is happen because using the virtual tour can generate sense of belonging and affective emotion which lead people able to reduce their stress level. Furthermore, in [12] conclude that VR have a possibility to becoming one of the tools that used in stress-related treatment especially during this pandemic era. However in those research papers which has been published there are very few area which explore to implement VR to reducing strees during Covid-19 for children from seven to ten years.

Based on this background, to fill the missing link on this research area, author create a VR simulation that focuses on helping children minimize stress and boredom levels during social movement restrictions, which offers four different outdor area. This research is deemed necessary so that children can stay at home and feel comfortable doing activities at home using this application in order to prevent the transmission of COVID-19, whose movement is still very massive and fast. This research is also a form of contribution in technology field which help childrens psychological issues especially during COVID-19.

II. Related Work

One of the literatures in this research is related to making learning applications regarding proper hand washing during a pandemic [8]. In this study, the use of AR uses hand movements as a marker and divides the learning process into three stages. The results of this study were 69% of respondents felt they better understand how to wash their hands properly after seeing this AR application. Another study [11] is a study that focuses on research on the use of VR for children with special needs (ABK) when they have to do rehabilitation at home during the COVID-19 pandemic. This study found that VR technology can be used in children with special needs to maintain and improve their motor function levels during home care. Another study [13] measured the accuracy and effectiveness of using VR as a companion tool for physical exercise in the elderly during the COVID-19 pandemic. The results of this study state that by using VR during the implementation of social restrictions, the elderly can improve motor skills, and can reduce the level of obesity in their body.

Research related to literature studies [14], conducting a review of the role of AR in handling the post-COVID-19 tourism sector, concluded that the role of ICT (Information and communication technology) such as AR, or VR is an innovative application that can be utilized by the tourism sector because it minimizes contact directly and can ensure tourist satisfaction and safety during the COVID-19 pandemic or after the pandemic is over. Another research related to the use of VR was carried out by [12] regarding the development of a VR-based project called MIND-VR which focuses on developing virtual psychoeducational experiences that provide basic information about stress and anxiety disorders in health workers. The results of related studies state that the therapeutic process using VR is more effective than using video. This study also states that further studies related to the implementation of VR to deal with stress and psychological trauma due to the impact of COVID-19 are needed, especially in the use of VR at home, or in the hospitals.

The study in [19] build a VR called “the secret garden” which provide refreshing scenery and calm voice to help people overcome their stress issue related to COVID-19. It took 7 days measurement to reach the conclusion. This research has an objective to reducing stress by improve people’s positive emotions using VR application. The result of the research already published in [23] which stated that, even in prelliminary stages, it shows positive impact that VR application can reduces participant stress during COVID-19 especially after two months lockdown. Another study regarding reduce stress during COVID-19 has been implemented in [22] that utilise 360° virtual tour. The study use VR application to allow participant explore place in 360° view. This research involves 235 participants and reach conclusion that using virtual tour have a high degree of satisfaction which lead to decreasing their stress level. This research also stated that VR have a contribution to improve people psychological well-being. Beside that, VR also becoming one of the technologies that becoming recommendation to use for people who want to reduce their stress level especially during or after this pandemic era.

Based on several studies that have been discusses, it can be concluded that technology, especially VR have many positive contributions for people live especially during lockdown era. One of the benefits that appeared by using VR can be used to reducing stress level, because sudden change in people daily life. However, there are very few studies about stress levels during this pandemic era which involves children. In fact, children becoming one of the most affected subjects this time, because they cannot play outside and study freely anymore. Based on several studies that have been discussed, to complement previous studies, and complete the missing link, this research has objective to builds a VR application that focuses on reducing and preventing stress in children during the COVID-19 pandemic. In addition to building VR applications, this study also measures the effectiveness of using VR, as well as the usefullnes of this application.

III. The Proposed Method

In this research, agile methodology has been choosed to develop the application. Even though according to [24] waterfall model still become the most commonly used in software development. It has lack of flexibility when major changes are happened [25]. Therefore, this research decides to use agile methodology which offer much flexibility of design changes in the future and offers faster release compare to the traditional one [25]. In addition, based on Fig. 1 before implementation process done, the whitebox and blackbox testing are carried out to make sure that VR application already meet the requirement.

The whole process of build a VR application can be seen in Fig. 1 and 2. In Fig. 1 showed the flow of VR application development, while Fig. 2 shows the overall research flow.

In Fig. 1, the research begins by analyzing the needs needed in the software and hardware sections including Unity3D, Oculus, and VR glasses. The application design stage consists of making storyboards, interface designs as well as terrain and environment assets that are created using blender and unity3D software. The programming process for building this VR application uses the default language of the Unity 3D software, namely C# which is assisted by visual programming on Playmaker assets. After the coding process is complete, it will be combined with various assets and 3D designs that have been built, then the rendering process is carried out. This process using the target API level 30 on the android system, and it is necessary to install the VR Google SDK for unity, as well as the VR headset that will be used. Before implementing the application, a testing process is

carried out using blackbox and whitebox tests [16] to ensure the application runs well. The next process is to do a demo to participants using the help of goggle glass to run VR applications. The overall research flow in the implementation process can be seen in Fig. 2.

In this study, as shown in Fig. 2, prior to do VR app implementations, participants were pretested first to determine the level of depression experienced using a "nine-symptom checklist" questionnaire [15] which divided the severity into 5 parts, namely:

- Minimal : scor 1 – 4
- Mild : scor 5 – 9
- Moderate : scor 10 – 14
- Moderately Severe : scor 15 – 19
- Severe : Scor 20 - 27

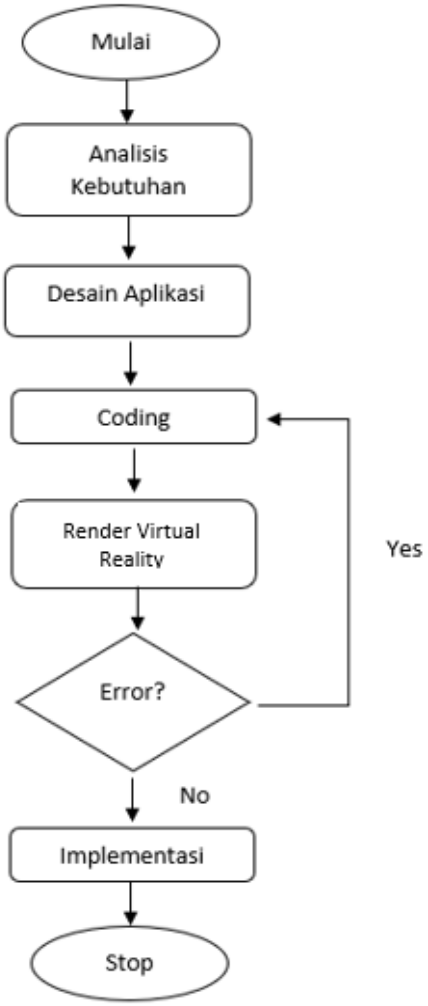


Fig. 1. VR Application Development Flow.

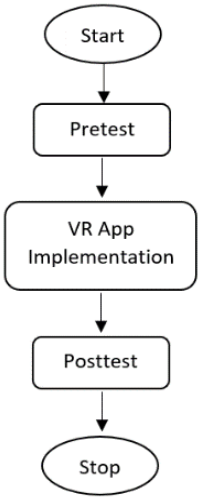


Fig. 2. VR Research Flow.

TABLE I. NINE SYMPTOM CHECKLISTS

No	Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not all	Several Days	More than half the days	Nearly every day
1	Little interest or pleasure in doing things?	0	1	2	3
2	Feeling down, depressed, or hopeless?	0	1	2	3
3	Trouble falling or staying asleep, or sleeping too much?	0	1	2	3
4	Feeling tired or having little energy?	0	1	2	3
5	Poor appetite or overeating?	0	1	2	3
6	Feeling bad about yourself - or that you are a failure or have let yourself or your family down?	0	1	2	3
7	Trouble concentrating on things, such as reading the newspaper or watching television?	0	1	2	3
8	Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual?	0	1	2	3
9	Thoughts that you would be better off dead, or of hurting yourself in some way?	0	1	2	3

The details of the "nine symptom checklists" questionnaire can be seen in Table I. This patient health questionnaire-9 (PHQ-9) is a recommendation to use as a self-measurement test to detect the rate of depression that consists from 0 to 3 level [26]. The primary reason this screening questionnaire has been chosed because it offers the ease of use the nine set of question. The ease of use the tools become the primary concern in this research because the subject that will be tested are children with has range of age from seven to ten. Beside the easiness, previous study which already used this PHQ-9 showed the accuracy screening using the nine set questions and the test also falled in acceptable area which obtain sensitifity around 0.84 and specifity about 0.77 [27]. Another reasearch that conducted in [28] showed a good sensitivy and specificity which fall below 0.83, and 0.72 respectively, using this PHQ-9 questionnaire does not necessary to divide the participants based on their gender. According to several reaseraches that already done which showed acceptable examination, it makes author believe that using this tools lead to the accurate results.

During the posttest stage, the same questionnaire table was used as already done in the pretest conditions, which aims to measure whether there is a positive impact obtained by respondents after using the VR application.

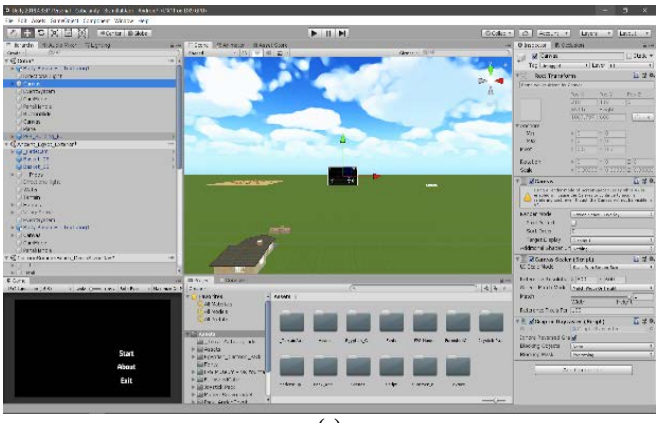
TABLE II. SUS LIST OF QUESTIONS

No	Questions	Strongly Disagree				Strongly Agree
		1	2	3	4	5
1	I think that I would like to use this system frequently.					
2	I found the system unnecessarily complex.					
3	I thought the system was easy to use.					
4	I think that I would need the support of a technical person to be able to use this system.					
5	I found the various functions in this system were well integrated.					
6	I thought there was too much inconsistency in this system					
7	I would imagine that most people would learn to use this system very quickly.					
8	I found the system very cumbersome to use					
9	I felt very confident using the system					
10	I needed to learn a lot of things before I could get going with this system					

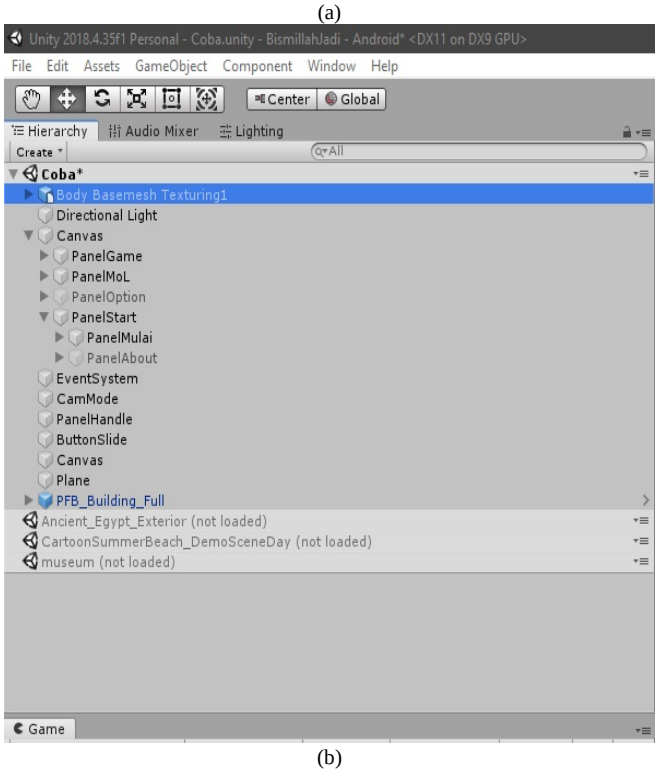
In addition, another test called System Usablitiy Scale (SUS) also carried out in order to measure the usefullnes of the application [17]. According to [29] SUS is one of the most tests that used to measure the perceive ussability and still remain relevant in the future. Based on those two researchers explain in the beginning, this SUS test is necessary to be done to meet the objectives in this research as written in Section 2. The details questioned of SUS can be seen in Table II, which have five scale, from the highest point (strongly agree) to the lowest point (strongly disagree).

IV. RESULT AND DISCUSSION

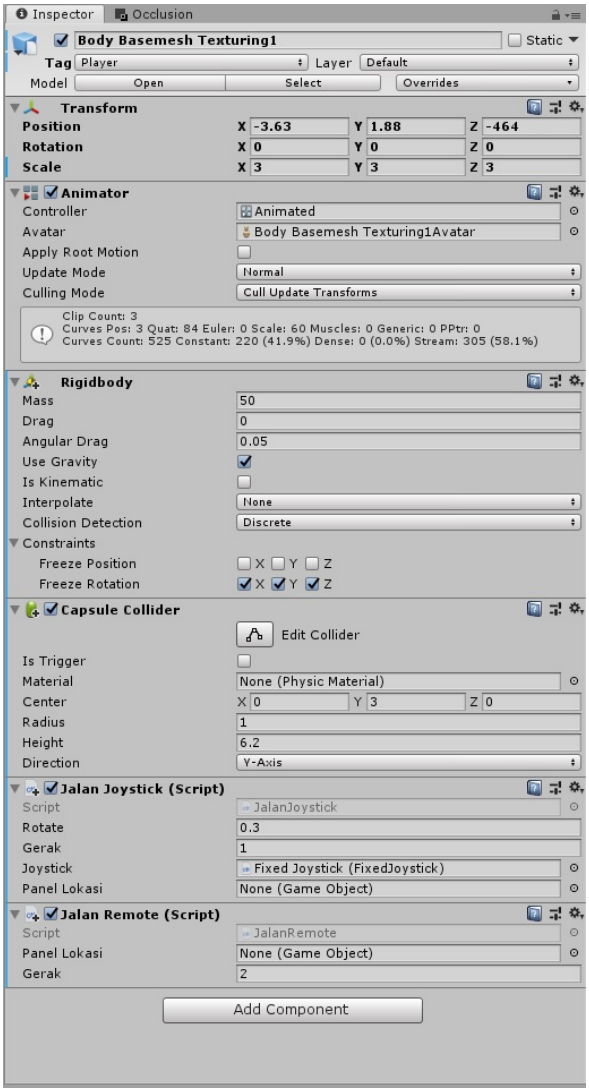
Based on the research flow in Section 3, the VR simulation application was built by adding the required assets, including Google Cardboard, Asset Image and Google VR SDK in Unity. One of the processes of making VR in unity3D can be seen in set collection of workspaces in Fig. 3a, 3b, 3c, and 3d.



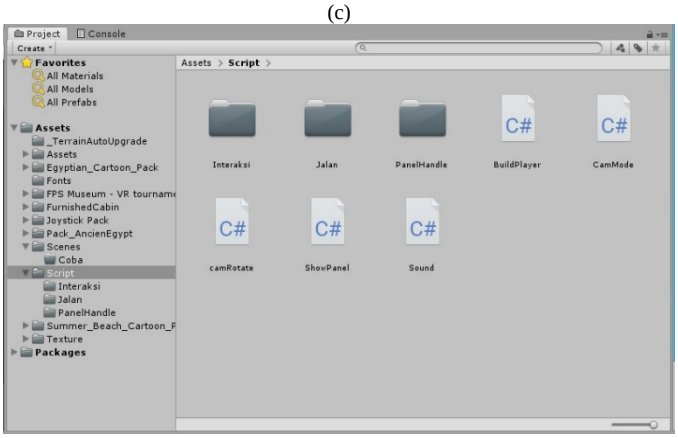
(a)



(b)



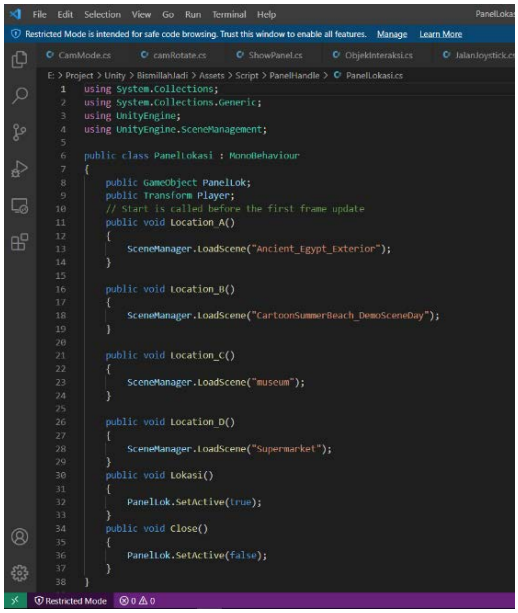
(c)



(d)

Fig. 3. (a) Whole Sample Image Process VR Development in Unity3D, (b) Hierarchy Page in Unity3D, (c) Inspector Page in Unity3D, (d) Project Struct.Ure in Unity3D.

In this VR application consists of four different scenes, namely: houses, beaches, Egypt and museums. The scene switching process is made using C# code, the snipped code of it process can be seen in Fig. 4.



```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using UnityEngine.SceneManagement;

public class PanelLokasi : MonoBehaviour
{
    public GameObject PanelLok;
    public Transform Player;
    // Start is called before the first frame update
    public void Location_A()
    {
        SceneManager.LoadScene("Ancient_Egypt_Exterior");
    }

    public void Location_B()
    {
        SceneManager.LoadScene("CartoonSummerBeach_DemoSceneDay");
    }

    public void Location_C()
    {
        SceneManager.LoadScene("museum");
    }

    public void Location_D()
    {
        SceneManager.LoadScene("Supermarket");
    }
    public void Lokasi()
    {
        PanelLok.SetActive(true);
    }
    public void Close()
    {
        PanelLok.SetActive(false);
    }
}
```

Fig. 4. Snipped Code C# for Switching Each Scene.

The final results of the VR application in this study are presented in Fig. 5, and 6 which are images based on the participants point of view when viewed using VR glasses.

Fig. 5. Sample Application VR Views 1.

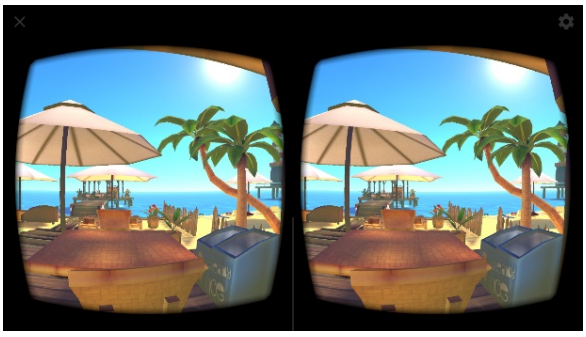


Fig. 6. Sample Application VR Views 2.

Before the application is used, as described in Section 3, blackbox and whitebox tests are carried out. The whitebox test is done by testing the program code of this application to make sure that no flaws appeared and the code written clearly. Afterwards, the blackbox testing perform to make sure all of the functionalities perform well. The overall results of blackbox testing can be seen in Table III.

TABLE III. RESULT OF BLACKBOX TEST

No	Test Class	Scenario Testing	Expected	Results
1	Start Menu	Pressing the Star Button	Displays a selection of VR mode and Mobile Mode	Valid
2	About Menu	Pressing about button	Displays about page	Valid
3	Exit Menu	Pressing the exit button	Exit the VR application	Valid
4	Mobile Mode Menu	Pointing the joystick right and left	Displays the appropriate direction with the joystick	Valid
		Pressing the Settings button	Displays the setting page	Valid
		Pressing Map button	Display various maps option	Valid
5	Map Menu	Pressing Egypt choice	Display Egypt Environment	Valid
		Pressing Beach choice	Display Beach Environment	Valid
		Pressing museum choice	Display Museum Environment	Valid
6	Egypt, Beach, and Museum Menu	Pressing the Back Button	Displays two choices menus are VR and mobile mode	Valid
		Pressing setting button	Display setting page	Valid
		Pressing Map btton	Display various map choices	Valid
		Pointing the joystick right and left	Displays the appropriate direction with the joystick	Valid

After passing the two functionality tests, the application was tested on 30 child respondents with an age range of 7-10 years which held with parental assistance using strict covid-19 protocol. In the process of testing the level of anxiety and SUS, parental assistance is necessary to obtain valid test results. The results of the comparison of pretest and posttest can be seen in Fig. 7. It can be seen that most of the children has obtain lower level of stress after used VR application.

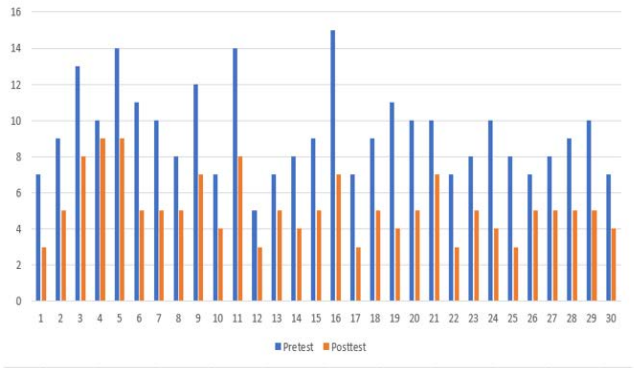


Fig. 7. Pretest and Posttest Results after Application Testing.

TABLE IV. RESULT OF SUS QUESTIONNAIRE

Respondent	Result of SUS Score										Total	Final Score
	1	2	3	4	5	6	7	8	9	10		
1	3	3	3	2	3	3	3	3	3	2	28	70
2	4	4	4	4	4	4	3	3	4	2	36	90
3	4	4	3	3	4	3	2	4	4	2	33	82.5
4	4	2	3	2	4	3	3	2	4	2	29	72.5
5	4	3	4	3	4	4	2	3	4	4	35	87.5
6	4	4	3	3	4	3	4	4	4	2	35	87.5
7	4	3	3	4	3	2	2	2	4	2	29	72.5
8	3	3	3	3	3	3	3	2	4	2	29	72.5
9	4	3	3	4	4	3	3	4	3	2	33	82.5
10	4	3	3	3	4	2	3	3	4	3	32	80
11	3	3	4	3	3	2	4	3	3	2	30	75
12	4	3	3	2	4	3	3	2	3	3	30	75
13	4	4	3	3	3	2	3	3	4	4	33	82.5
14	4	3	3	3	4	4	3	2	4	2	32	80
15	4	3	3	4	2	4	3	3	3	2	31	77.5
16	3	3	3	3	4	4	4	4	4	3	35	87.5
17	4	3	3	4	3	4	4	4	3	3	35	87.5
18	4	3	3	2	4	3	3	2	4	3	31	77.5
19	4	3	3	3	3	2	3	2	4	2	29	72.5
20	3	3	3	3	3	2	3	3	3	2	28	70
21	4	4	3	2	3	3	3	2	4	4	32	80
22	4	4	3	3	3	3	4	3	4	2	33	82.5
23	4	4	4	4	4	4	4	3	3	3	37	92.5
24	3	3	3	2	4	3	3	2	4	3	30	75
25	4	4	3	3	3	3	2	4	4	3	33	82.5
26	3	3	4	4	4	3	3	2	4	2	32	80
27	4	4	3	3	2	3	2	3	4	4	32	80
28	3	4	4	3	4	4	3	2	4	3	34	85
29	4	4	3	2	4	4	3	2	4	3	33	82.5
30	3	3	3	2	4	3	3	2	4	3	30	75
average												79.91

Based on the results of the pretest and posttest, it can be seen that before implementing VR application, most of respondents experienced depression in the mild and moderate ranges, there was only one respondent who was in moderately severe condition. However, after the respondent was given a VR application to get a new atmosphere of exploring several virtual places, it can be seen that there was a significant decrease in stress and anxiety levels, which can be seen in the survey results listed in Fig. 7. Based on Fig. 7, it can be seen that most of the respondents experienced decrease in anxiety level, about one level lower compare to pretest result. According on the results of the posttest respondents

experienced anxiety at a minimal and mild level. In order to measure the usability of the application, SUS questionaire in Table II already performed which use 30 respondents as subjects. The result of SUS testing can be seen in Table IV.

Based on the result shown in Table IV, it can be concluded that the average SUS score obtainend in this VR application around 79.91. It is located in acceptable area, between good and excellent adjective rating in SUS scoring system [18] and can be conclude that this application considered to be useful enough in order to show VR application.

V. Conclusion and Future Work

Based on the results obtained in Section 4, it can be concluded that this VR application has a contribution in reducing the level of anxiety and stress in children during the Covid-19 pandemic as evidenced by the results of the pretest and posttest presented in Fig. 7. Beside that, this VR application has a good adjective ratings SUS based on Table IV. Those two results that achieved in Section four already answer the objective which already discussed in Section two in this research, are to build VR application which allow reduce stress for children and measure the usefulness and effectiveness this application using SUS measurement.

However, many future works are needed to do in order to increase the usefullnes of this applicaion in society. One of challenges that need to tackle in this application are the addition of features to interacts in every scene so children can explore many things while enjoying the virtual environment. Beside that, the education about Covid-19 knowledge also relevant to be added in the future.

Acknowledgment

This research is fully funded by Hibah Integrasi Tridharma (HIT) Universitas Muhammadiyah Surakarta under the grant number 004/A3.III/FKI/I/2021.

References

[1] Jung, Timothy, M. Claudia tom Dieck, Hyunae Lee, and Namho Chung. "Effects of virtual reality and augmented reality on visitor experiences in museum." In *Information and communication technologies in tourism*, pp. 621-635. Springer, Cham, 2016.

[2] Putra, Ghali Adyo, Rinta Kridalukmana, and Kurniawan Teguh Martono. "Pembuatan simulasi 3D virtual reality berbasis Android sebagai alat bantu terapi acrophobia." *Jurnal Teknologi dan Sistem Komputer* 5, no. 1, pp. 29-36, 2017.

[3] Zhai, Yusen, and Xue Du. "Mental health care for international Chinese students affected by the COVID-19 outbreak." *The Lancet Psychiatry* 7, no. 4: e22, 2020.

[4] Worldometers.info. “COVID-19 Coronavirus Pandemic,” *Worldometer*, Dover Delaware, U.S.A, accessed 09 Januari, 2021.

[5] Qiu, Hanqin, Qinghui Li, and Chenxi Li. "How technology facilitates tourism education in COVID-19: case study of nankai University." *Journal of Hospitality, Leisure, Sport & Tourism Education*. p. 100288, 2020.

[6] Wang, Cuiyan, Riyu Pan, Xiaoyang Wan, Yilin Tan, Linkang Xu, Cyrus S. Ho, and Roger C. Ho. "Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China." *International journal of environmental research and public health* 17, no. 5, pp. 1729, 2020.

[7] Brooks, Samantha K., Rebecca K. Webster, Louise E. Smith, Lisa Woodland, Simon Wessely, Neil Greenberg, and Gideon James Rubin. "The psychological impact of quarantine and how to reduce it: rapid review of the evidence." *The Lancet*, 2020.

[8] Hanafi, Hafizul Fahri, Mohd Helmy Abd Wahab, Kung-Teck Wong, Abu Zarrin Selamat, Muhamad Hariz Muhamad Adnan, and Fatin Hana Naning. "Mobile augmented reality hand wash (MARHw): mobile application to guide community to ameliorate handwashing effectiveness to oppose Covid-19 disease." *International Journal of Integrated Engineering* 12, no. 5, pp. 217-223, 2020.

[9] Singh, Ravi Pratap, Mohd Javaid, Ravinder Kataria, Mohit Tyagi, Abid Haleem, and Rajiv Suman. "Significant applications of virtual reality for COVID-19 pandemic." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2020.

[10] Hua, Yun, Rong Qiu, Wen-yan Yao, Qin Zhang, and Xiao-li Chen. "The effect of virtual reality distraction on pain relief during dressing changes in children with chronic wounds on lower limbs." *Pain Management Nursing* 16, no. 5, pp. 685-691, 2015.

[11] Demers, Marika, Ophélie Martinie, Carolee Winstein, and Maxime T. Robert. "Active Video Games and Low-Cost Virtual Reality: An Ideal Therapeutic Modality for Children With Physical Disabilities During a Global Pandemic." *Frontiers in Neurology* 11, pp. 1737, 2020.

[12] Imperatori, Claudio, Antonios Dakanalis, Benedetto Farina, Federica Pallavicini, Fabrizia Colmegna, Fabrizia Mantovani, and Massimo Clerici. "Global Storm of Stress-Related Psychopathological Symptoms: a brief overview on the usefulness of virtual reality in facing the mental health impact of COVID-19." *Cyberpsychology, Behavior, and Social Networking* 23, no. 11 pp. 782-788, 2020.

[13] Gao, Zan; Lee, Jung E.; McDonough, Daniel J.; Albers, Callie. "Virtual Reality Exercise as a Coping Strategy for Health and Wellness Promotion in Older Adults during the COVID-19 Pandemic" *J. Clin. Med.* 9, no. 6: 1986, 2020.

[14] Mohanty, Priyakrushna, Azizul Hassan, and Erdogan Ekis. "Augmented reality for relaunching tourism post-COVID-19: socially distant, virtually connected." *Worldwide Hospitality and Tourism Themes*, 2020.

[15] Kroenke, K., Spitzer, R.L. and Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), pp.606-613.

[16] Nidhra, S., & Dondeti, J. (2012). Black box and white box testing techniques-a literature review. *International Journal of Embedded Systems and Applications (IJESA)*, *2*(2), 29-50.

[17] Brooke, J., 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), pp.4.

[18] Bangor, Aaron; Kortum, Philip T.; Miller, James T. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24(6),* pp.574–594.

[19] Riva, G., Bernardelli, L., Browning, M.H., Castelnuovo, G., Cavedoni, S., Chirico, A., Cipresso, P., de Paula, D.M.B., Di Lernia, D., Fernández-Álvarez, J. and Figueras-Puigderrajols, N., 2020. COVID feel good—an easy self-help virtual reality protocol to overcome the psychological burden of coronavirus. *Frontiers in Psychiatry*, *11*, p.996.

[20] Razeghi, S., Alipour, S. and Sabet, A., 2021. Enhancing stress management training and communication skills of nursing students during COVID-19 pandemic based on augmented reality. *Journal of Modern Medical Information Sciences*, *7*(1), pp.38-47.

[21] Ye, J., 2020. Pediatric mental and behavioral health in the period of quarantine and social distancing with COVID-19. *JMIR pediatrics and parenting*, *3*(2), p.e19867.

[22] Yang, T., Lai, I.K.W., Fan, Z.B. and Mo, Q.M., 2021. The impact of a 360° virtual tour on the reduction of psychological stress caused by COVID-19. *Technology in Society*, *64*, p.101514.

[23] Riva, G., Bernardelli, L., Castelnuovo, G., Di Lernia, D., Tuena, C., Clementi, A., Pedroli, E., Malighetti, C., Sforza, F., Wiederhold, B.K. and Serino, S., 2021. A Virtual Reality-Based Self-Help Intervention for Dealing with the Psychological Distress Associated with the COVID-19 Lockdown: An Effectiveness Study with a Two-Week Follow-Up. *International journal of environmental research and public health*, *18*(15), p.8188.

[24] Vijayasarathy, L. and Butler, C., Choice of Software Development Methodologies. *IEEE Software*, pp.0740-7459.

[25] Al-Saqqa, S., Sawalha, S. and AbdelNabi, H., 2020. Agile Software Development: Methodologies and Trends. *International Journal of Interactive Mobile Technologies*, *14*(11).

[26] Manea, L., Gilbody, S. and McMillan, D., 2015. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *General hospital psychiatry*, *37*(1), pp.67-75.

[27] Lotrakul, M., Sumrithe, S. and Saipanish, R., 2008. Reliability and validity of the Thai version of the PHQ-9. *BMC psychiatry*, *8*(1), pp.1-7.

[28] Beard, C., Hsu, K.J., Rifkin, L.S., Busch, A.B. and Björgvinsson, T., 2016. Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, *193*, pp.267-273.

[29] Lewis, J.R., 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), pp.577-590.

Gesture based Arabic Sign Language Recognition for Impaired People based on Convolution Neural Network

Rady El Rwelli[1]
Department of Arabic Language, College of Science and Arts in Qurayyat, Jouf University, Gurayat, Saudi Arabia

Osama R. Shahin[2], Ahmed I. Taloba[3]
Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Gurayat, Saudi Arabia

***Abstract*—The Arabic Sign Language has endorsed outstanding research achievements for identifying gestures and hand signs using the deep learning methodology. The term "forms of communication" refers to the actions used by hearing-impaired people to communicate. These actions are difficult for ordinary people to comprehend. The recognition of Arabic Sign Language (ArSL) has become a difficult study subject due to variations in Arabic Sign Language (ArSL) from one territory to another and then within states. The Convolution Neural Network has been encapsulated in the proposed system which is based on the machine learning technique. For the recognition of the Arabic Sign Language, the wearable sensor is utilized. This approach has been used a different system that could suit all Arabic gestures. This could be used by the impaired people of the local Arabic community. The research method has been used with reasonable and moderate accuracy. A deep Convolutional network is initially developed for feature extraction from the data gathered by the sensing devices. These sensors can reliably recognize the Arabic sign language's 30 hand sign letters. The hand movements in the dataset were captured using DG5-V hand gloves with wearable sensors. For categorization purposes, the CNN technique is used. The suggested system takes Arabic sign language hand gestures as input and outputs vocalized speech as output. The results were recognized by 90% of the people.**

***Keywords*—Arabic sign language; convolution neural network; hand movements; sensing device**

I. INTRODUCTION

Around 60 million people use body language around the world, and an automated tool for interpreting it might have a big effect on communication between those who use it and those who don't. Sign language is a means of wordless communication that includes the use of body parts. In sign speaking and listening, face features, as well as eye, hand, and lip gestures, are used to transmit information. People who are deaf or hard of hearing rely heavily on sign language as a form of communication in their daily lives [1]. Nevertheless, the lack of consistency in shape, size, and posture of the hands or fingers in an image made computer interpretation of hand signals exceedingly difficult. SLR can be approached in two aspects: image-based and sensor-based. The main advantage of expression frameworks is that users do not need to use complicated gadgets. In any scenario, extensive operations are required during the pre-processing stage. The importance of language in growth cannot be overstated. It facilitates the internalization of social norms and the development of communication control in addition to serving as a channel for interpersonal communication. Even though they can hear the language spoken across them, deaf children do not learn a language to express themselves in the same way that hearing impairment children do.

SLR research has recently been divided into two categories: vision and contact-based approaches. This between sensing users and devices is part of the interaction technique. It usually employs an interferometric glove that collects finger motion, bending, movement, and angle information of the produced sign via EMG signals, inertial estimate, or electromagnetism. As the platform's input, the vision-based technique uses data obtained from video streams or photos captured with the camera. It's also divided into two categories: presence and 3D model-based techniques [2]. The majority of 3D model-based strategies begin to gather the position of the hand and joint angle of the hand in 3D spatial into a 2D image. Whereas demeanor identification relies on features extracted from the image's PowerPoint display, recognition is completed by matching the characteristics [3]. Although many hearing-impaired people have mastered sign language, few "regular" people understand or can use it. This has an impact on impaired people's communication and creates a sense of separation among them and the "regular" society. This chasm can be bridged by deploying a technology that constantly converts sign language to textual and vice versa. Numerous paradigm advancements in many scientific and technological fields have now aided academics in proposing and implementing systems that recognize sign languages.

Disabilities people interact through hand signals, which is a gesture-based communication strategy rather than written or spoken language. Arabic is the official language of 25 different countries. In certain nations, Arabic is spoken by only a small percentage of the population [4]. According to some accounts, the total number of countries is between 22 and 26. Although the Arabic language is deontological, the Arabic gesture is not. Arabic is spoken by Jordanians, Libyans, Moroccans, Egyptians, Palestinians, and Iraqis, to name a few. Each country, though, has its unique dialect. To put it another way, there seem to be two forms of Arabic: standard and colloquial. As they all employ the same alphabets, the Arabic sign language (ArSL) is also the same. This function is quite beneficial to research studies. The Arab

deaf communities are a close-knit group. Interaction between the deaf and hearing communities is low, focusing primarily on communities with deaf people, relations of the deaf, and occasionally play companions and professionals.

The recognition of Arabic Sign Language includes a continuous identification program based on the K-nearest neighbor classifier and a feature extraction method for the Arabic sign language. However, the fundamental flaw with Tubaiz's method would be that patients must wear interferometric hand gloves to gather information on specific gestures, which can be extremely distressing again for users [5]. For the construction of an Arabic sign language recognition, an interferometric glove was created. Using hidden Markov model (HMM) and temporal characteristics, continual identification of Arabic sign language is possible [6]. A study was performed on the translation of Arabic sign language to text for usage on portable devices. While the above papers cover a wide range of sign languages, Arabic Sign Language was also the subject of research in a few instances. Using a Hidden Markov Model (HMM) quantifier, the researchers achieve 93% accuracy for a sample of 300 words. In comparison to HMM, they use KNN and Bayesian classifications [7], which produce equivalent results. This presents a network matching method for continual detection of Arabic Sign Language sentences. Decision trees and the breakdown of motions into stationary poses are used in the model. Using a polynomial runtime technique, they attain at least 63% accuracy when interpreting multi-word phrases.

This paper deals with the Gesture Based Arabic Sign Language Recognition for Impaired People and this uses the Convolution Neural Network process as the research system. There is different section that deals with the process of Arab Sign Language and the Convolution Neural Network. Section 1 organizes the Introduction of the gesture sign Arab language and the Machine Learning system. Different methods and research involved in Arab language recognition were expressed in detail in Section 2; the proposed methodology is presented in Section 3. The result and Discussion are investigated in Section 4 and finally, the paper gets concluded.

II. Literature Review

The most comfortable and creative way for the hard of hearing to communicate is through hand signals. With improvements in multimedia tools and networks, academics have long been drawn to innovation. Sign language communications systems as a way to increase network technology for the hearing and speech impaired, offering increased social possibilities and integration. This study introduces a framework for leveraging the Microsoft Kinect device to communicate in Arabic sign language. The proposed method is based on the gesture recognition architecture for Arabic signs proposed by [8] for language communication systems. The suggested Language for Arab sign technique has a sign identification rate of 96 %, according to experimental data. In addition, the typical mission completion time for an Arabic sign was roughly 2.2 seconds. As a result, the suggested technology can be used to develop a real-time Arabic sign languages communications network. Finally, survey respondents stated that the projected procedure is consumer and simple to use and that it may be utilized to recognize and show Arabic signs at a minimal cost.

Researchers attempt to be using ICT to improve the Deaf community at large life quality by building solutions that can help them improve communication with the rest of the world and among themselves. Designers describe work on the construction of an Avatar-based translation for Deaf individuals from Arabic Speech to Arabic Sign Language in this paper. The study begins with an overview of the Deaf community's situation in the Arabic-speaking population, as well as a brief assessment related to particular research. [9] a translation system based on the avatar of Arabic speech and Arabic sign language for deaf people is recognized. The study begins with an overview of the Deaf community's situation in the Arabic-speaking population, as well as a short assessment of various related research. The next section describes the research system's design considerations. The technology will be built around a dataset of captured 3D Arabic sign language movements. Data gloves will be used to capture the gesture recognition motion.

The use of an automatic speech recognition method for Arab sign language (ArSL) has significant societal and humanitarian implications. With the growing deaf culture, such technology will aid in integrating such individuals and allowing them to live a regular life. Arab sign language, like other languages, has many subtleties and distinct qualities that necessitate the use of a useful weapon to treat it. The author in [10] propose a novel system based on deep learning that will automatically recognize words and numbers in Arab sign language when fed with a genuine dataset. Research conducted comparison research to demonstrate the effectiveness and robustness of suggested method vs established approaches based on k-nearest neighbors (KNN) and Support Vector Machines (SVM). Hearing is essential for normal language and speech development, and hearing impairment occurs whenever the acuity to normally heard noises is reduced [11]. Many studies show that one out of once each three to four educated citizens with any degree of hearing loss faces educational, social, and impede learning. The goal of this study was to assess the hearing impairment in hearing-impaired child's psychological traits (communication, social, emotional, and cognitive), and then connect this pattern to a linguistic scale.

The author in [12] proposed the Arab sign Language recognition and for those with hearing impairments, sign language entails the movement of the arms and hands as a channel of understanding. The identification of certain characteristics and the classification of specific input data are the two key steps in an automatic sign identification system. Many methods for categorizing and identifying sign languages have been proposed in the past to improve reliability. However, recent advances in the field of machine learning have prompted us to pursue more research into the identification of hand signs and gestures using deep neural networks. The Arabic gesture has seen significant research on hand gestures and gesture recognition. This research proposes a vision-based system that uses CNN to recognize Arabic hand sign-based letters and translate them into Arabic speech. With a deep learning model, the suggested system

automatically recognizes hand sign symbols and shouts out another output in Arabic. This system recognizes Arabic hand sign-based letters with 90% accuracy, indicating that it is a very reliable technology. Using more powerful hand gesture recognition technologies like Motion Sensors or Xbox Kinect can enhance accuracy even more. The result will be given to the text into the voice engine, which will create the sounds of the Arabic hand sign-based characters.

Hearing-impaired people can be found throughout the world, so developing good local level sign language recognition (SLR) systems is critical. This did a thorough assessment of computerized sign language identification based on supervised learning strategies and processes published between 2014 and 2021 and found that existing systems require theoretical categorization to accurately interpret all available data. As a result, focus on aspects that are included in practically all basic sign detection methods. The author in [13] present a comprehensive framework for investigators that analyzes their advantages and weaknesses. This study also demonstrates the importance of types of sensors in this sector; it appears that acknowledgment based on the integration of datasets, such as vision-based and webcam channels, outclasses unimodal analysis. Furthermore, recent advances in research facilities have enabled them to advance from the official press of sign language protagonists and turns of phrases to the capacity to change ongoing gesture conversations with minimal latency. Many of the models available are adequate for a range of tasks, but neither of them currently has the requisite generalization potential.

III. Methodology

A. Arabic Sign Language Architecture

The Architecture for Arabic Sign Language Communication System is a proposed solution that serves as a low-cost multiple languages translation. A sign language-based communications network, all while retaining high precision and economical usability [7]. The following Fig. 1 Architecture of Arabic Sign Language is shown. Hardware, software, and the network are the three elements that make up the system architecture. A gesture authentication method and a video representation are included in the hardware device [14]. The Gesture recognition digital storage repository, the Sign recognition center, and the Sign media center are all part of a software component. The network device is responsible for sending and receiving Arabic sign language data across a network connection.

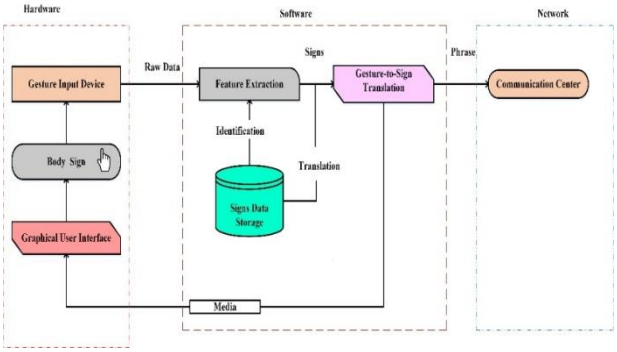


Fig. 1. Architecture of Arabic Sign Language.

1) Hardware: The transmitter and receiver channels via which the user controls the system are provided by the hardware components of the system. The Gesture input unit is a Microsoft Kinect from the first generations, which collects data and delivers it to the Sign recognition center [15]. The output data from the Sign media center is shown via the Display devices (audio system and digital display). In its current state, the stability supports visual data.

2) Software: The program's software is in charge of extracting features and movement translation, as well as offering a simple Graphical User Interface (GUI). Based on the existing and established lexicon, the Sign Identification Center turns the raw input into a collection of predefined signs [16]. The Sign language datastore is being used to obtain information about signs. The Sign media center translates the signs obtained from the Sign recognition center into the desired medium and speech, which it then sends to the Display technologies or engaging in dialogue.

The gesture recognition data store includes both gesture dictionaries and translating dictionaries from one vernacular to another. The size of something like the information storage is limited due to the project's concept stage.

Source: ArSLAT: Arab Sign Language Alphabets Translator [17].

3) Network: The Signs media player sends media to the Communication center, which then transmits this through the system to its intended destination [18]. This section is presently unimplemented because it is unrelated to usability testing. The above Fig. 2 shows the Arab Sign Language Alphabets and this image could be used for the further processing system.

Fig. 2. Arab Hand Sign.

B. Preprocessing of Data

The first stage in creating a working deep learning model is data preprocessing. This is used to convert raw data into a format that is both usable and efficient. The flowchart of data preprocessing is shown in Fig. 3.

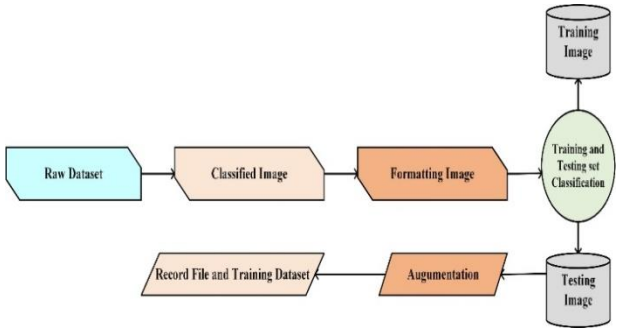


Fig. 3. Data Preprocessing.

1) Raw data: The image captured using the camera is termed the raw data in this section the raw image suits the hand sign image of the Arabic language and this is implemented in the proposed method [19]. The following environment is considered for image representation:

- Different angles.
- Lighting condition changing.
- Focus and good quality.
- Object size and distance adjustment.

The purpose of making raw photographs is to create a dataset that can be used for training and testing. the Arabic Alphabet from the dataset of the suggested program.

2) Classified image: The presented method categorizes the Arabic Alphabet's pictures. To understand the system, one subfolder is utilized to store photographs from one classification. In the developed framework, all subfolders that describe categories are maintained together in one primary special folder "dataset."

3) Number of epochs: The number of epochs represents how many times the complete dataset is processed into the neural network during training. There is no perfect number for it though, and it is determined by the facts.

4) Formatting image: The graphics of hand signs are usually uneven and have a varied background. To obtain the hand component, it is important to remove the unneeded elements from the photos. Images are referred to as digital data that has been rasterized for usage on a display device or printing in some of those types [20]. The extract was subjected is the process of converting visual data into a set of pixels.

5) Classification of training and testing dataset: The image taken for the formatting could be classified based on the training or testing image. A controlled learning method for classification examines the training data set to find, or learn, the best relationships between two variables that will produce a strong forecasting model. The goal is to create a trained (fitted) model that does a good job of generalizing to new, unknown data.

6) Augmentation: Real-time data is always incomplete and unusual due to several modifications (rotating, moving, and so on). Image augmentation is a technique used to improve the achievement of deep neural networks. It purposefully tries to manipulate images with methods such as shear, shifts, flips, and rotation. Using this image enhance raw images [21], the suggested system's images are rotated dynamically from 0 to 360 degrees. A small number of photographs were also ripped at random with a 0.2-degree range, and a small number of images were inverted horizontally.

C. Frame Work

The design of the Arabic sign language recognition utilizing CNN is shown in Fig. 4, Convolution Layer. CNN is a machine learning (ML) system that uses perceptron algorithms in the implementation of its operations for data gathering. These systems are categorized as artificial neural networks (ANN). The discipline of machine learning is where CNN is most useful. It mostly aids in the recognition and classification of images [22]. CNN is made up of two parts: feature extraction and classification. Each element has its own set of features that must be investigated. These elements will be explained in detail in the following sections. A convolutional neural network (CNN, or ConvNet) is a type of neural network that is used to analyze visual information [23]. One of the main reasons that researchers have realized the efficacy of deep learning is the vitality of convolution layers nets in image processing. They are in charge of significant advancements in computer vision (CV), which has significant application in self-driving cars, mechatronics, unmanned aerial vehicles, safety, medical advances, and treatment options for the visual impairments.

Convolutional neural networks employ an architecture that lends itself especially well to image classification. These systems allow neural nets to learn quickly. This enables us to approach enhanced deep multi-layer systems for image classification. CNN continues to learn from data using the Backpropagation algorithm and its derived products. Modern implementations make use of specialized GPUs to improve results even further.

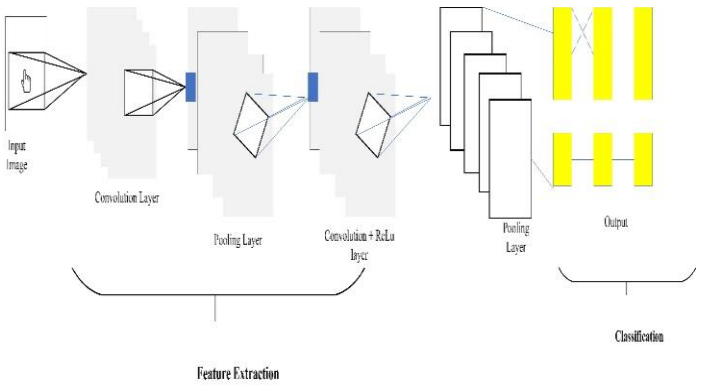


Fig. 4. Convolution Layer.

1) Input blocks: Squeeze Net requires an input block that would be at least 224 X 224 (With 3 channels for RGB). CNN is made up of numerous components. The Convolution operation, on the other hand, is CNN's most important component. The statistical combining of two roles to create a third function is referred to as a convolution layer. To create a feature map, the combination of the inputs using filtering or kernel is necessary. Convolution is performed by dragging each filtering over a specific input [24]. A matrix combination is performed at each location, and the outcome is added to a different feature map. Every image is turned into a 3D matrix with a defined width, height, and depth. Because the image (RGB) has color channels, the thickness is shown as a measurement.

Various convolutions can be conducted on raw data with various filters [25], resulting in various feature extraction. The output of the convolution layer is created by combining the multiple feature images. The output is then passed via an input layer, resulting in complex output. The length of a given step that the Fourier filter performs each time is referred to as the stride. The size of a step is typically 1; this indicates that the convolution filter is moving image pixel. When increasing the size of a step, the filters will slide across the input with a larger frequency, resulting in less overlap between the cells. Researchers should do something to stop extracted features from decreasing because it is always less than the input size. These are going to apply to cushion here.

$$Output_{size} = \frac{input_{size} - Filter_{size} + 2 * padding_{size}}{Stride_{\ size}}$$

2) Max pooling layer: In between Convolution layers, a pooling layer is naturally added. However, its primary goal is to reduce dimensionality and reduce calculation time by using fewer parameters. It also prevents overtraining and cuts down on training time. Pooling can take numerous forms, the most frequent of which is max pooling. It employs the maximum value in all windows, resulting in a smaller feature map with the same amount of information. To estimate the size of the pooling layer's output produced, the panel sizes must be specified ahead of time; the following equations can be used.

$$Output_{size} = \frac{input_{size} - Filter_{size}}{Stride_{\ size}} + 1$$

The pooling layer provides some high accuracy in all cases, indicating that a certain component will be recognizable regardless of where it appears on the panel.

The categorization component of CNN is the second most critical component. The objects are classified up of a few tiers that are all interconnected (FC). An FC layer's neurons have a strong relationship between the two to every one of the preceding layer's activations. The FC layer aids in the mapping of representations between inputs and outputs. The layer's functions are carried out using the same concepts as a conventional Neural Network [4]. One Dimensional data, on the other hand, can only be accepted by an FC layer. The flatten function of Python is utilized to create the new method for converting three-dimensional data to just one data.

3) Dropout regularization techniques: Overfitting is a significant and serious issue in deep neural networks. Dropout is a method of dealing with this difficult challenge. It is accomplished by randomly discarding some neural units in the neural network with an artificially designed ratio throughout training. The degree of co-adaptation between neuronal units has been greatly reduced. Using dropout on a neural network is equivalent to extracting a thinned network from the original entire network. During the training process, a series of thinning networks are collected using dropout at a specific dropout ratio. During the testing phase, it is not possible to directly help determine by combining the forecasts from exponentially thinning models. This should employ a whole untinned network with fewer weights to forecast outcomes by implicitly aggregating the results of all those thinned systems. Dropout greatly lowers overfitting and outperforms other regularization methods. The convolutional neural network using dropout would be discussed.

4) Activation function: The activation function is a node placed at the end of or between Neural Networks. This has various sorts of activation functions, but this discussion will concentrate on Rectified Linear Units (ReLU). The ReLU function is the most often used objective function in neural networks. ReLU has a significant benefit over other training algorithms in that it does not stimulate all neurons at the same time. The picture for the ReLU algorithm above shows that it turns all negative input to zero but does not stimulate the neuron. Because just a few neurons are stimulated at a time, it is incredibly computationally efficient. It does not reach saturation in the positive area. In reality, the ReLU activation function converges six times quicker than the $tanh$ and sigmoid activation functions.

5) Features extraction: The Convolutional Neural Network is made up of several basic parts. The convolution layer is a critical component of the CNN network. This layer denotes the mathematical description of functions that result in a third function. To generate a feature map, convolution must be performed within the input using a kernel or a filter. The convolution implementation consists of sliding each filter with sufficient input. At each location, matrix multiplying is conducted, and the outcome is placed on a feature map. Each image is converted to a 3D matrix with depth, height, and width. Because the image is made up of color channels, the depth has been deemed a dimension. Multiple convolutions are performed on the input dataset using appropriate criteria, resulting in distinct feature maps. The result of the convolution layer is obtained by combining the multiple feature maps. The kernel is a two-dimensional (2D) array of elements that would be used as weights generally. As demonstrated in Fig. 5, the convolution procedure is conducted by dragging the kernel across the picture. The result of the convolution layer is a feature mapping. Every section

that is subjected to the dragging and convolution processes is referred to as an interesting region (IR). The accompanying equation is used to carry out the convolutional procedure.

$$Z_{ij} = (I * K)_{ij} = \sum_m \sum_n I_{i-m,j-n} K_{mn} - 1$$

Where K is the kernel and I is the input image. Every layer's output in Convolution Neural Network can be stated as follows:

$$y_i^l = f(z_i^l)$$

Then, $z_i^l = \sum_{j=1}^{l-1} w_i^j x_j^{l-1} - 2$

Where y represents the outputs of the layer, z represents the activation function, i represents a layer l neuron, w represents the weight, and x represents the input information.

$$w = \{w_{ij}^l : l = 1,3, \dots L-1; i = 0,1, \dots I; j = 0,1, \dots J\} - 3$$

$$x = \{x_j^l : l = 1,3, \dots L-1;\ j = 0,1, \dots J\} - 4$$

The pooling layer is the CCN's second layer. The primary goal of this layer is to make the feature mapping from the convolution layer easier to understand. It emphasizes the characteristic by using the maximum, summation, or averaging operations. The fully connected layer is the third layer. This layer's primary function is to transform a two-dimensional (2D) feature map into a one-dimensional (1D) one. This style is appropriate for deciding on feature categorization based on pre-defined features.

This employed an Arabic sign languages dictionary in this research. In the training phase of hand symbol identification, employed certain motions from a lexicon as a ground truth. The lexicon is distributed in the form of graphic groups of motions. Every set of images symbolizes a different type of social setting. This chooses over 40 motions performed with one hand and over 10 motions performed with both hands. This database is used to train Convolutional Neural networks. The technology was put to the test with real-life motions performed by coworkers.

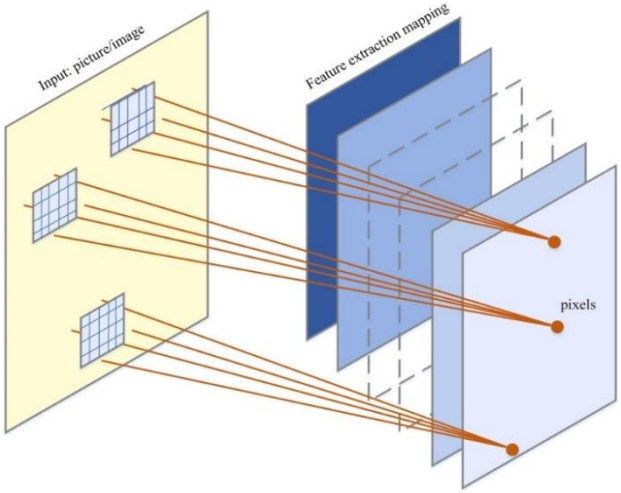


Fig. 5. Feature Extraction.

To recognize fingers and hands, computer vision architecture was created. Within its area of view, it separates and monitors them. The architecture collects movement monitoring data in the form of a series of pixels. The measured positions, sign orientations, and other information about every object recognized in the current frames are stored in the monitored data frames. A unique pixel is used to illustrate the identified fingers and hands. The flow chart for CNN proposed model is shown in Fig. 6. On still photos, summarize a technique using the methods below.

- Gesture frames are captured.
- Image denoising is a technique for improving the appearance of images.
- Employing Convolution Neural Network (CNN) to separate the face of a signer.
- Using Convolutional Neural Network to segment hand and finger gestures.
- Utilizing Convolution Neural Network to detect motions.
- By evaluating motions to elements in a pre-built database, motions can be recognized.
- Getting the translations for the motion that was recorded.

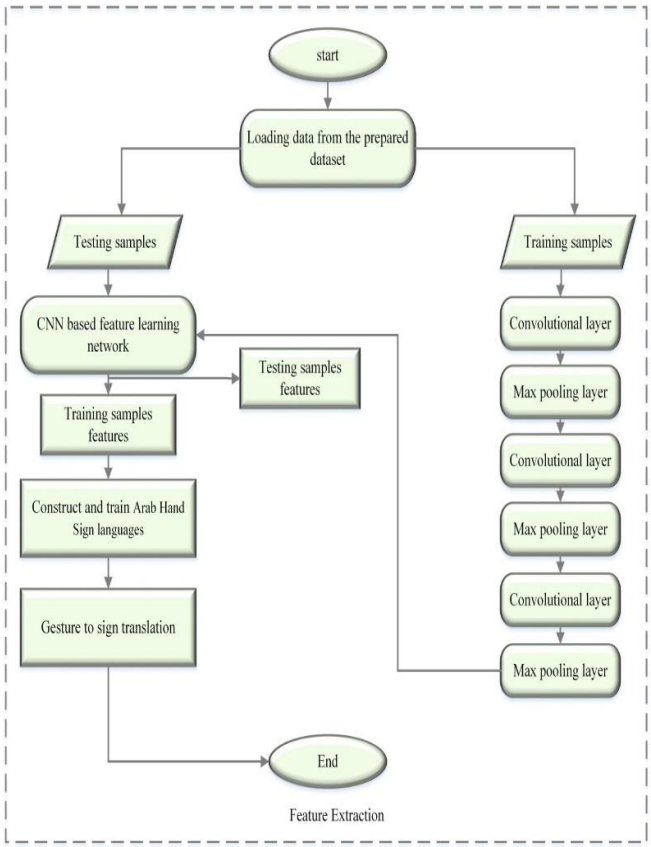


Fig. 6. Flow Chart for CNN Proposed Model.

IV. Result and Discussion

Two convolution layers are used to test the suggested system. After that, each fully connected layer is followed by two max-pooling layers. The first layer of the convolution operation has a different pattern; there are 30 kernels in the first layer, while the second level has 64 kernels; nonetheless, the kernel size in both layers is 3×3 layer. Each pair of convolution and max-pooling was examined using two alternative dropout regularization values of 25% and 50%, respectively. As a result, this option allows for the elimination of one input out of every four inputs (25%) and two inputs out of every four inputs (50%) from each combination of convolutional and pooling layers.

The various sizes of training sets, as can be seen in Fig. 7, when training the network using 80 % of the images from the dataset, the accuracy reaches its highest point of 90.03%. Table I shows the Percentage Training Set.

Researchers compared the suggested system's outcomes with KNN (k-nearest neighbor) with Euclidean distance and SVM (support vector machines) with various kernels processors typically shown in this field to demonstrate its effectiveness.

Other factors that influence identification, including such facial movements, have been explored in prior studies. Various input detectors, like the jump action controllers, are also employed, as well as integrating multiple input detectors to handle the various characteristics described above. In addition, a novel learning approach was applied in this study, which yielded encouraging outcomes.

TABLE I. Percentage Training Set

Training Image	Detection Rate
60%	85.65%
50%	83.28%
80%	90.03%
70%	88.45%

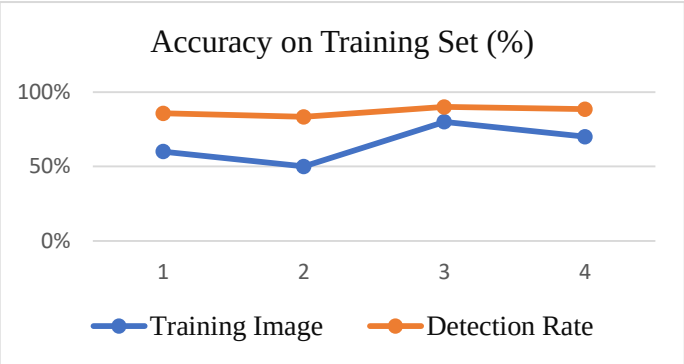


Fig. 7. Accuracy on Training Set.

TABLE II. The CNN Configuration's Categorization Results

Input depth	CNN layer	Accuracy	
Configuration 1	Kernel	$5 \times 5 \times 8$	91.5%
	Subsampling	2×2	
Configuration 2	Kernel	$5 \times 5 \times 8$	90.6%
	Subsampling	2×2	
Configuration 3	Kernel	$5 \times 5 \times 8$	94%
	Subsampling	2×2	
Configuration 4	Kernel	$5 \times 5 \times 8$	92%
	Subsampling	2×2	
Configuration 5	Kernel	$5 \times 5 \times 8$	89%
	Subsampling	2×2	
Configuration 6	Kernel	$5 \times 5 \times 8$	89.8%
	Subsampling	2×2	
Configuration 7	Kernel	$5 \times 5 \times 8$	86%
	Subsampling	2×2	

The scheme then exhibits an optimistic accuracy rate with reduced loss rates in the following phase (testing phase). The accuracy rate was reduced even further when augmented graphics were used while maintaining nearly the same precision. Each digital image in the testing stage was processed before being used in this model. The proposed system generates a vector of 10 values, with 1/10 of these values being 1 and all other values being 0 to represent the predicted class value of the given data. The system is then linked with its signature step, in which a hand sign is converted to Arabic speech, Table II.

The Fig. 8 Detection Rate Comparison is shown based on Table III Comparison of Detection Rate. In this, the comparison is done between the proposed system and the other method such as SVM with RBF kernel and linear kernel and KNN. Among those methods, a higher rate of accuracy is found in the proposed method. So, by implying this method the impaired people could easily recognize the sign.

TABLE III. Comparison of Detection Rate

Classifier	Detection Rate
Support Vector Machine with linear Kernel	86%
Support Vector Machine with RBF kernel	85%
KNN	68%
Proposed system	90.03%

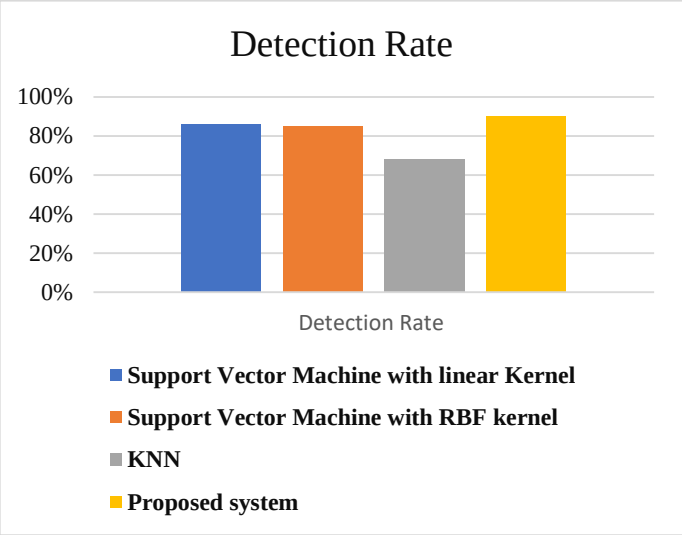


Fig. 8. Detection Rate Comparison.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research at Jouf University for supporting this work by Grant Code: (DSR-2021-04-0317).

Funding Statement: This work was funded by the Deanship of Scientific Research at Jouf University under grant No (DSR-2021-04-0317).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

V. CONCLUSION

The Identification of sign languages and Arabic Sign Language (ArSL), as well as several types of classifications and their outputs, were studied using various symbols and signs and ArSL. This suggested survey is carried out to take the best classifier for hand gesture recognition systems that depend on several sign languages. Some of the developed models were shown to be quite efficient, however only on limited applications. The studies originate from all across the world and include a wide range of sign language variances, which is critical for assuring worldwide coverage. The entire operation and performance of sign language recognition are represented using neural networks, machine learning, and deep learning classifiers, among others. In terms of accuracy, the Deep learning-based classifier CNN produced the results in research. Thus, the gesture Based Arabic Sign Language Recognition for Impaired People is based on Convolution Neural Network System. In addition, the size of the data collection could be enhanced further in future study projects. The suggested system's result is Arabic-language speech obtained through the detection of Arabic sign language. Furthermore, the solution presented here would be excellent for impaired people.

REFERENCES

[1] M. A. Almasre and H. Al-Nuaim, "A comparison of Arabic sign language dynamic gesture recognition models," Heliyon, vol. 6, no. 3, p. e03554, Mar. 2020, doi: 10.1016/j.heliyon.2020.e03554.

[2] A. S. Elons, M. Abull-Ela, and M. F. Tolba, "A proposed PCNN features quality optimization technique for pose-invariant 3D Arabic sign language recognition," Applied Soft Computing, vol. 13, no. 4, pp. 1646–1660, 2013.

[3] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat, "SIFT-Based Arabic Sign Language Recognition System," in Afro-European Conference for Industrial Advancement, vol. 334, A. Abraham, P. Krömer, and V. Snasel, Eds. Cham: Springer International Publishing, 2015, pp. 359–370. doi: 10.1007/978-3-319-13572-4_30.

[4] A. Shahin and S. Almotairi, "Automated Arabic Sign Language Recognition System Based on Deep Transfer Learning," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 10, pp. 144–152, 2019.

[5] M. A. Bencherif et al., "Arabic Sign Language Recognition System Using 2D Hands and Body Skeleton Data," IEEE Access, vol. 9, pp. 59612–59627, 2021.

[6] M. Mustafa, "A study on Arabic sign language recognition for differently abled using advanced machine learning classifiers," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 3, pp. 4101–4115, 2021.

[7] B. Hisham and A. Hamouda, "Supervised learning classifiers for Arabic gestures recognition using Kinect V2," SN Appl. Sci., vol. 1, no. 7, p. 768, Jul. 2019, doi: 10.1007/s42452-019-0771-2.

[8] T. Aujeszky and M. Eid, "A gesture recogintion architecture for Arabic sign language communication system," Multimedia Tools and Applications, vol. 75, no. 14, pp. 8493–8511, 2016.

[9] S. M. Halawani, "An Avatar Based Translation System from Arabic Speech to Arabic Sign Language for Deaf People," p. 8.

[10] S. Hayani, M. Benaddy, O. El Meslouhi, and M. Kardouchi, "Arab sign language recognition with convolutional neural networks," in 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), 2019, pp. 1–4.

[11] H. Desoky, O. Raaft, and S. N. Azab, "Psycho-communicative interruptions in hearing-impaired Egyptian Arabic-speaking children," Beni-Suef Univ J Basic Appl Sci, vol. 10, no. 1, p. 35, Dec. 2021, doi: 10.1186/s43088-021-00124-9.

[12] M. Kamruzzaman, "Arabic sign language recognition and generating Arabic speech using convolutional neural network," Wireless Communications and Mobile Computing, vol. 2020, 2020.

[13] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," IEEE Access, vol. 9, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[14] Y. Saleh and G. Issa, "Arabic sign language recognition through deep neural networks fine-tuning," 2020.

[15] W. Abdul et al., "Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM," Computers & Electrical Engineering, vol. 95, p. 107395, 2021.

[16] A. M. Ahmed, R. Abo Alez, G. Tharwat, M. Taha, B. Belgacem, and A. M. Al Moustafa, "Arabic sign language intelligent translator," The Imaging Science Journal, vol. 68, no. 1, pp. 11–23, 2020.

[17] N. El-Bendary, H. M. Zawbaa, M. S. Daoud, A. E. Hassanien, and K. Nakamatsu, "ArSLAT: Arabic Sign Language Alphabets Translator," in 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krackow, Poland, Oct. 2010, pp. 590–595. doi: 10.1109/CISIM.2010.5643519.

[18] N. Aouiti, M. Jemni, and S. Semreen, "Arab gloss and implementation for Arabic Sign Language," in 2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA), 2017, pp. 1–6.

[19] A. Boukdir, M. Benaddy, A. Ellahyani, O. E. Meslouhi, and M. Kardouchi, "Isolated Video-Based Arabic Sign Language Recognition Using Convolutional and Recursive Neural Networks," Arabian Journal for Science and Engineering, pp. 1–13, 2021.

[20] N. A. Sherbiny and M. A. Tessler, "Arab oil: impact on the Arab countries and global implications.[16 papers]," 1976.

[21] J. Napier and L. Leeson, "Sign language in action," in Sign language in action, Springer, 2016, pp. 50–84.

[22] M. Kardouchi, “Arab Sign language Recognition with Convolutional Neural Networks”.

[23] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat, “Sift-based arabic sign language recognition system,” in Afro-european conference for industrial advancement, 2015, pp. 359–370.

[24] R. Ahuja, D. Jain, D. Sachdeva, A. Garg, and C. Rajput, “Convolutional neural network based american sign language static hand gesture recognition,” International Journal of Ambient Computing and Intelligence (IJACI), vol. 10, no. 3, pp. 60–73, 2019.

[25] K. Al-Fityani, “Deaf people, modernity, and a contentious effort to unify Arab sign languages,” PhD Thesis, UC San Diego, 2010.

Micro Expression Recognition: Multi-scale Approach to Automatic Emotion Recognition by using Spatial Pyramid Pooling Module

Lim Jun Sian, Marzuraikah Mohd Stofa, Koo Sie Min, Mohd Asyraf Zulkifley
Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, Bangi, Malaysia

***Abstract*—Facial expression is one of the obvious cues that humans used to express their emotions. It is a necessary aspect of social communication between humans in their daily lives. However, humans do hide their real emotions in certain circumstances. Therefore, facial micro-expression has been observed and analyzed to reveal the true human emotions. However, micro-expression is a complicated type of signal that manifests only briefly. Hence, machine learning techniques have been used to perform micro-expression recognition. This paper introduces a compact deep learning architecture to classify and recognize human emotions of three categories, which are positive, negative, and surprise. This study utilizes the deep learning approach so that optimal features of interest can be extracted even with a limited number of training samples. To further improve the recognition performance, a multi-scale module through the spatial pyramid pooling network is embedded into the compact network to capture facial expressions of various sizes. The base model is derived from the VGG-M model, which is then validated by using combined datasets of CASMEII, SMIC, and SAMM. Moreover, various configurations of the spatial pyramid pooling layer were analyzed to find out the most optimal network setting for the micro-expression recognition task. The experimental results show that the addition of a multi-scale module has managed to increase the recognition performance. The best network configuration from the experiment is composed of five parallel network branches that are placed after the second layer of the base model with pooling kernel sizes of two, three, four, five, and six.**

Keywords*—Micro expression recognition; facial expression; spatial pyramid pooling module; multi-scale approach; deep learning

I. Introduction

According to the research from [1], [2], faces are the main human "tools" to express information in terms of emotion. Facial expression is an important means that enable humans to undergo social interaction with each other. This is because 55% of human feelings are manifested by their facial expression. For example, an observer can deduce that someone is feeling disgusting if his/her upper lip is rising upward.

Facial expression can be broken down into two categories, which are macro-expression and micro-expression. A macro-expression is an intentional facial expression, while a micro-expression is an unintentional facial expression. Benjamin et al. [3] investigated that the major differences between them are the intensity and time taken to manifest the expression. Deng et al. [4] reported both expressions are widely used as an input to various applications and the most obvious application is to estimate the hidden emotions.

On the other hand, Micro-expression (ME) is an unintentional, quick facial movement that is primarily used to express the emotions of happiness, sadness, and surprise [5]. A ME happened in a short time, usually happened in the range of 0.04s until 0.2s. Hence, it is a hard task for a human to use their bare eyes to detect the occurrence of ME. Even if a human is undergoing training to detect an ME, their average performance is only slightly better than other people who do not undergo the training process. Hence, Zhao and Li [6] showed that machine learning is proposed to aid humans in analyzing the ME to understand human's true emotions.

Machine learning (ML) can be broadly classified into traditional machine learning and deep learning. Researchers in pattern recognition tasks have frequently applied both techniques to the applications of facial expression recognition [7], human activity recognition [8], recycling system [9], and image recognition [10]. Traditional machine learning relies on a set of handcrafted features, which is then passed to a decision-making module algorithm such as decision tree, neural network, and Support Vector Machine (SVM) [11], [12]. However, it is a time-consuming task for a computer vision engineer to judge which features are the best to describe the emotions.

The deep learning methodology is different compared to the traditional machine learning approach, whereby the features of interest are obtained through iterative optimal training such as through the convolution process [10], [13]. Usually, after the feature maps have passed through a convolution process, they will undergo a pooling process. Fig. 1 shows the generalized framework of traditional machine learning and deep learning algorithms for human emotion recognition tasks.

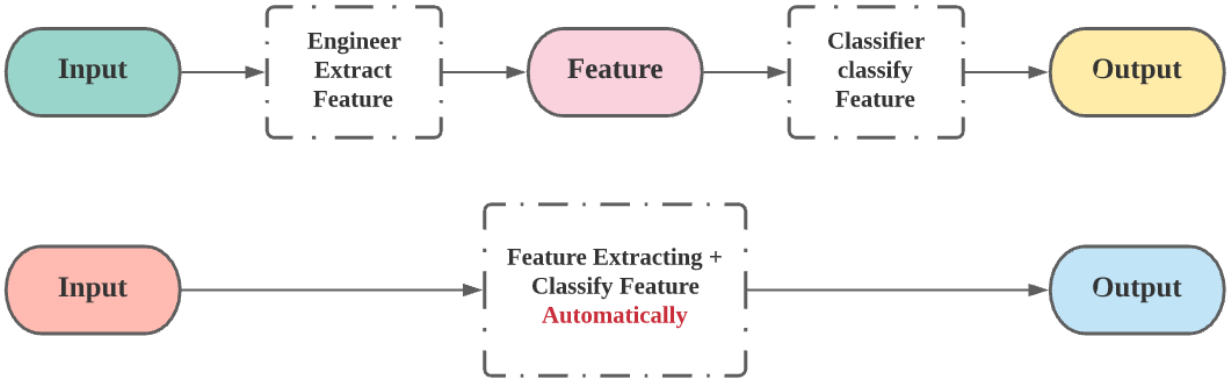


Fig. 1. Generalized Framework of Traditional Machine Learning (above) and Deep Learning (below).

Spatial Pyramid Pooling (SPP) originates from spatial pyramid matching that utilizes spatial statistical properties to represent global information for recognition purposes. Ke et al. [14] discussed the main benefit of the SPP module is it can produce constantly desired outputs without considering the input size requirement to the deep learning model. The training process of a deep learning model with multiple sizes of the image can also prevent over-fitting problems [15]. In this work, several configurations of SPP have been explored that include different numbers of layers and kernel sizes, as well as placement of the module. Besides that, the base model that has been used in this work is also compact in nature, whereby it is commonly used in tracking application, which requires fast computational model [16]. In Oh et al. [17] have surveyed different algorithms from different researchers, and their respective accuracy is summarized in Table I.

By referring to Table I, the previous works' accuracy using CASME II dataset is within the range of 40% to 60%, while for the SMIC dataset, the accuracy range is between 50% and 70%. In general, these accuracies are not satisfactory enough for real-life application. By referring to Table I, the previous works' accuracy using CASME II dataset is within the range of 40% to 60%, while for the SMIC dataset, the accuracy range is between 50% and 70%. In general, these accuracies are not satisfactory enough for real-life application. Micro-expression is a crucial set of facial cues that are extensively employed in all parts of human society. However, simple facial macro expression cues are not enough to effectively relay the real emotions. In order to resolve the issues, this study presents several variants of SPP to improve the recognition accuracy of the automated micro-expression recognition applications.

TABLE I. ACCURACY OF MICRO-EXPRESSION RECOGNITION FROM DIFFERENT PAPERS

Papers	**Features**	**Classifier**	**Accuracy (%)**	
			CASME II	***SMIC***
Huang et al. [18]	SpatioTemporal Completed Local Quantiza-tion Patterns (STCLQP)	SVM	58.39	64.02
He et al. [19]	Multi-task mid-level feature learning (MMFL)	SVM	59.81	63.15
Huang et al. [20]	Discriminative Spatiotemporal Local Radon-based Binary Pattern (STLPB-IP)	SVM	64.37	60.98
Li et al. [21]	Histograms of Image Gradient Orientation (HIGO)	SVM	67.21	68.29
Liong et al. [22]	Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP)	SVM	46.00	54.00
Happy et al.[23]	Fuzzy Histogram of Optical Flow Orientations (HFOFO)	SVM	56.64	51.83
Le Ngo et al. [24]	LBP-TOP	SVM	49.00	58.00
Xu et al. [25]	Facial Dynamics Map	SVM	45.93	54.88
Ping et al. [26]	LBP-TOP	Group Sparse Spation-Temporal Reature Learning (GSLSR)	67.89	70.12
Zong et al. [27]	Hierarchical STLBP-IP	Kernelized Group Sparse Learning (KGSL)	63.83	60.78

Continuing from this introduction section will be Section II that presents a comprehensive overview of the basic architecture, hyperparameter function, and related layers used in modeling the CNN model. In Section III, the new improved CNN model is introduced by embedding spatial pyramid pooling into the basic model. Then, the experimental results were discussed in Section IV. Finally, the last section concludes the paper with some suggestions for future works.

II. Related Work

A. Basic Architecture of CNN

The structure of a neural network is very much similar to a human being's brain neuron. When the neuron is excited, it will deliver a chemical substance to its neighboring neuron, which will alter the state potential. If the next neuron's potential is higher than the threshold, the state will be activated and vice versa [28]. A compact deep learning structure of a convolutional neural network (CNN) mainly comprises three convolutional layers, pooling layers, and full-connected layers. The convolution process happened in convolutional layers to extract features from the input image through a sliding window operation. Then, the resultant feature maps will be passed to the pooling layers to reduce the map dimension. Fully-connected layers will be used to segregate the data into different classes [29], [30]. Fig. 2 shows the basic architecture of a CNN.

B. Convolutional Layer

A convolution operator performs a linear operation to a spatial map through a sliding window process. The process starts by applying a small number array (kernel) on the input data (tensor) to compute the product of each element of kernel and tensor for all tensor data. After that, each of the computed outputs will be summed up to form a new value in the respective position of the tensor (feature map). The whole steps will be repeated by applying multiple kernels into the tensor [31]. Based on Ma et al. [32], one kernel can extract one pattern characteristic of the input image. Fig. 3 shows the summary of a convolution process.

C. Padding

According to Rikiya et al. [31], the overlapping between the center element of the kernel with the outermost element of the input tensor should be avoided. Hence, a padding operator was introduced to enlarge the feature map dimension. There are two popular types of padding operations which are zero padding (or called the same padding) and valid padding. The process of zero padding is to make the convolution image larger by adding the zeros to its borders. The size of the output from the zero padding will be the same as the size before undergoing the convolution process [33], [34]. Fig. 4 shows the summary of a zero-padding process.

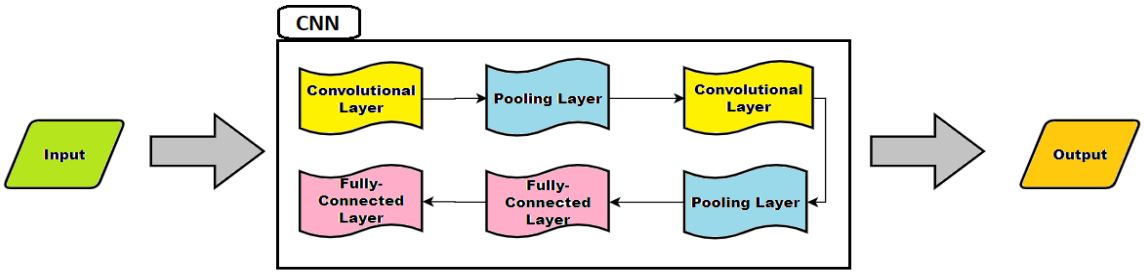


Fig. 2. Basic Architecture of CNN.

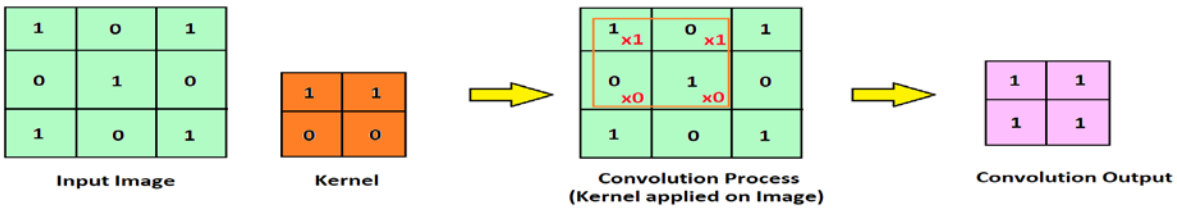


Fig. 3. Summary of Convolution Process.

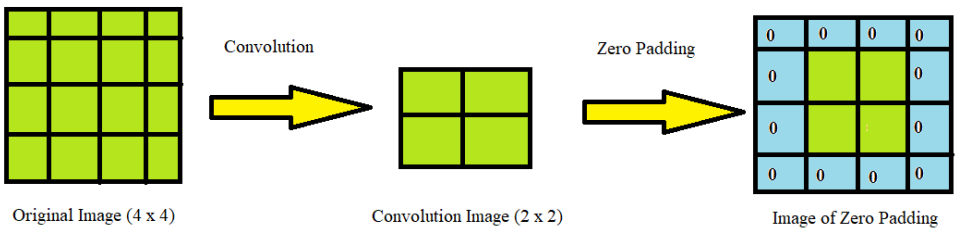


Fig. 4. Summary of Zero Padding Process.

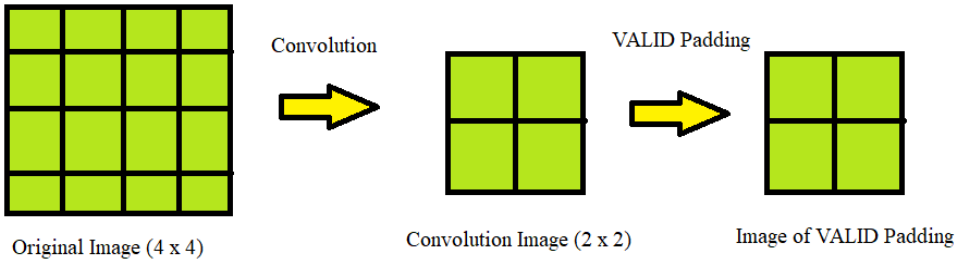


Fig. 5. Summary of VALID Padding Process.

If a Keras-Tensorflow library is used, a valid padding option means that no padding will be applied. The size of the image going through a valid padding option will be the same as the size of the convolution image as shown in Fig. 5.

D. Pooling Layer

There are two advantages of applying a pooling layer to the deep network. Firstly, it helps to decrease the size of the feature map and hence reduces the complexity of the network. Secondly, Guo et al. [35] shows it also helps to extract the important feature optimally. Based on Victor and Isabel [36], there are three types of pooling operators, which are maximum pooling, average pooling, and attentive pooling. For the maximum pooling operator, the maximum element in each overlapping area between the kernel and the feature map will be chosen as the resultant output. Shallu and Rajesh [37] identify the only disadvantage of maximum pooling operation is if most of the values on the feature map are high values, the significant features may be discarded. Fig. 6 shows the operational flow of the maximum pooling process with a 4 x 4 feature map, 2 x 2 kernel size with a step size of two pixels.

According to Sharma et al. [37], the average pooling operation process differs from the maximum pooling process. The new value in the respective position of the feature map is obtained by calculating the average pixels of each overlapping area between the kernel and the feature map. Fig. 7 shows the summary of an average pooling process with a 4 x 4 feature map, 2 x 2 kernel size with a step size of two pixels.

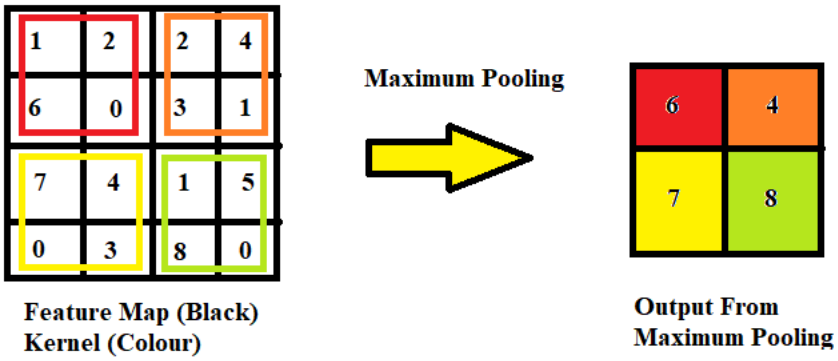


Fig. 6. Operational Flow of a Maximum Pooling Process (4 x 4 Feature Map, 2 x 2 Kernel Size with Step Size of Two Pixels).

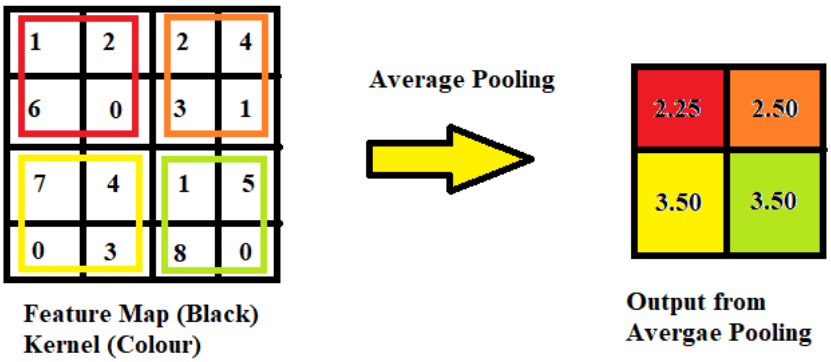


Fig. 7. Operational Flow of an Average Pooling Process (4 x 4 Feature Map, 2 x 2 Kernel Size with Step Size of Two Pixels).

E. Activation Function

The activation function, which is also known as the transfer function, is used to determine the neurons that will be excited and passed as the high state to the next neurons. According to Chigozie et al. [38] and Feng et al. [39], an ordinary neural network without an activation function, the output of each layer of the network will consist of a linear combination of its last layer, which is shown in (1). Let y = output, x = input, n = number of nth layers, b = bias.

$$y = w_1x_1 + w_2x_2 + ... + w_nx_n + b \quad (1)$$

According to the formula above, the range of the output will start from the negative infinity until positive infinity. This shows that the neurons in the network are not limited to a certain finite range. Conversely, with the presence of an activation function, the linear output will be converted to a non-linear result and the output range will fall within a finite value. According to Chigozie et al. [38] and Feng et al. [39], the non-linear result is shown in (2).

$$y = a(w_1x_1 + w_2x_2 + ... + w_nx_n + b) \quad (2)$$

In this study, two types of activation functions, which are ReLu and Softmax function will be utilized. Wang et al. [40] stated that ReLu is an activation function that is based on a piece-wise function. ReLu function is known to be good in handling gradient-vanishing problems. This is the main advantage of a ReLu function compared to other activation functions such as Sigmoid and Tanh functions [41]. Another advantage of a ReLu function is it can be computed at a faster speed compared to the other functions. A positive gradient with a value equal to one will be produced for positive input, while a negative gradient is produced when the input is negative. Fig. 8 shows the graphical representation of the ReLu function.

According to Martin et al. [42], the Softmax function is a useful function that converts the weight vector to the probability distribution. Such a function will make sure that the output is in the range between zero until one and the sum of the outputs will be unity [43]. The softmax function is commonly used in models with multiple classes. Chigozie et al. [38] showed that the probability of each class will be provided and the class with the highest probability is considered as the target class. The only disadvantage of the Softmax function is the output value of zero cannot be produced and hence, a sparse probability distribution cannot be produced through this function. This is because any small output value in the sparse probability distribution will be treated as a negligible value which is zero.

F. Fully Connected Layer

A fully connected layer means each neuron is connected to every neuron to its next layer. The major function of the fully connected layer is to classify the input image into a variety of classes. The Softmax function is used in its output layer [44].

G. Local Response Normalisation

Local Response Normalisation (LNR) is a non-trainable layer based on the lateral inhibition process. It decreases the neighboring pixels activation state, which deems to be too huge in order to form a big contrast in a feature map. This normalization process involves a decreasing operator by squaring and normalizing the pixel values of the feature map in a local neighborhood [45]. From Alex et al. [46], the (3) representation for LNR operation is shown as below, where $b^i_{x,y}$ = output neuron, $a^i_{x,y}$ = input neuron, N = total kernel in the layer, x, y = position of it kernel, others = constant value.

$$b^i_{x,y} = a^i_{x,y} / \left(k + a \sum_{j=\max(0,i-\frac{n}{2})}^{\min(N-1,i+\frac{n}{2})} (a^j_{x,y})^2\right)^B \quad (3)$$

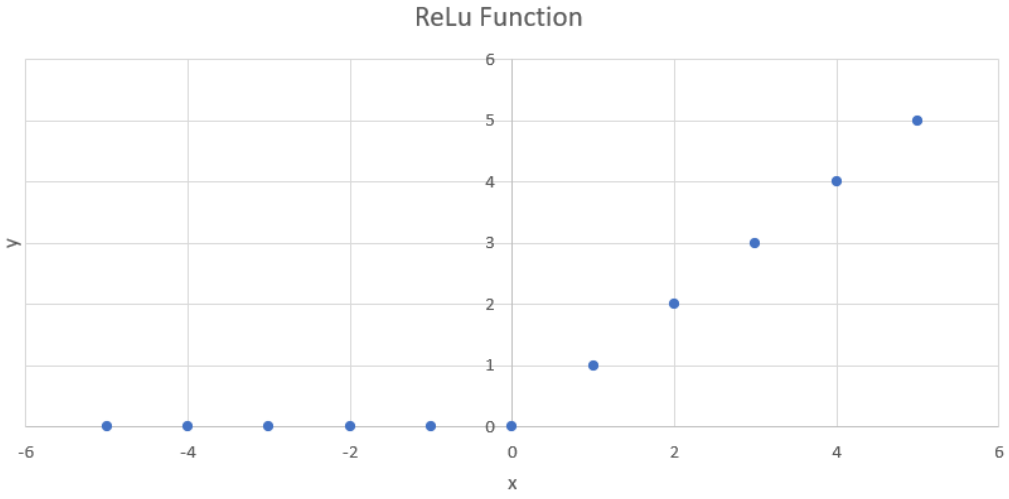


Fig. 8. Graphical Representation of ReLu Function.

III. METHODOLOGY

A. CNN Model

In this study, a compact CNN model [47] derived from the VGG-M model, which was introduced by Chatfield et al. [48]. A compact model is utilized because of the small number of available training data will lead to the problem of over-fitting according to the study by Nicholas et al. [49]. By having a compact property, a deep learning model can be optimized well even with a small number of data. In this study, the combined datasets have only 441 micro-expression videos, which is considered as a small training set. This is the primary reason deeper and newer models, such as Resnet [50] and DenseNet-SPP [51] are not used as the base model.

VGG-M has a good balance between computational speed and accuracy. Hence, it has excellent average performance and has been widely used in tasks involving the fields of vision such as multi-biometric recognition. The original VGG-M consists of nine layers network with five convolutional layers, three fully-connected layers and one flatten layer. The five convolutional layers have the kernel number of 96 for the first convolutional layer, 256 for the second convolutional layer, and 512 for the third, fourth, and fifth convolutional layer. Jiang et al. [52] proposed the fully-connected layers have a kernel size of 128 for the first and second layers, while three nodes for the third fully-connected layer to reduce the complexity of the training process. Each convolutional and full-connected layer is coupled with the ReLu activation function except for the last fully-connected layer, which is coupled with the Softmax activation function. In [53], LNR is applied after the first and second convolutional layers only. A maximum pooling layer is applied after each LNR layer and also after the fifth convolutional layer to make the model more robust and have a better generalization capability [54]. Table II shows the summary of the modified VGG-M architecture used in this study.

According to Table II, the primacy change that can be observed is the reduction in stride size for the first and second maximum pooling (Pool1 and Pool2) from two to one. This is because a larger feature map size is needed to insert the Spatial Pyramid Pooling (SPP) layer. Note that the size of the input image used in this study is 75 x 75. If the original stride size of the first and second maximum pooling layer is used, the feature map size after going through the second layer will become 3 x 3 only, which is not enough to embed the multiple average pooling processes in the SPP layers. Conversely, if the stride size is changed to one, the feature map size will be 13 x 13, which is enough to implement the multi-scale average pooling in the SPP layer.

TABLE II. SUMMARY OF THE MODIFIED VGG-M ARCHITECTURE

Table Head	Type of Layer	Kernel Number	Kernel Size	Stride	Padding	Activation Function
Conv1	Convolution	96	7 x 7	2 x 2	Valid	ReLu
Norm1	LRN	-	-	-	-	-
Pool1	Maximum Pooling	-	3 x 3	1 x 1	-	-
Conv2	Convolution	256	5 x 5	2 x 2	Valid	ReLu
Norm2	LRN	-	-	-	-	-
Pool2	Maximum Pooling	-	3 x 3	1 x 1	-	-
Conv3	Convolution	512	3 x 3	1 x 1	Same	ReLu
Conv4	Convolution	512	3 x 3	1 x 1	Same	ReLu
Conv5	Maximum Pooling	512	3 x 3	1 x 1	Same	ReLu
Pool5	Flatten Layer		3 x 3	2 x 2	-	-
Flat1	Full-connected Layer		-	-	-	-
FC1	Full-connected Layer	128	-	-	-	ReLu
FC2	Full-connected Layer	128	-	-	-	ReLu
FC3	Convolution	3	-	-	-	Softmax

B. Spatial Pyramid Pooling

For general SPP, the input will undergo three down-sampling operations separately which are average pooling, batch normalization, and rectified linear unit activation function (ReLu). The kernel sizes of average pooling are different between the parallel layers. After that, the size of the down-sampling output will be adjusted (resized process) according to the skip-connection layer size. Then, these outputs are combined using a concatenation operator [55] as shown in Fig. 9.

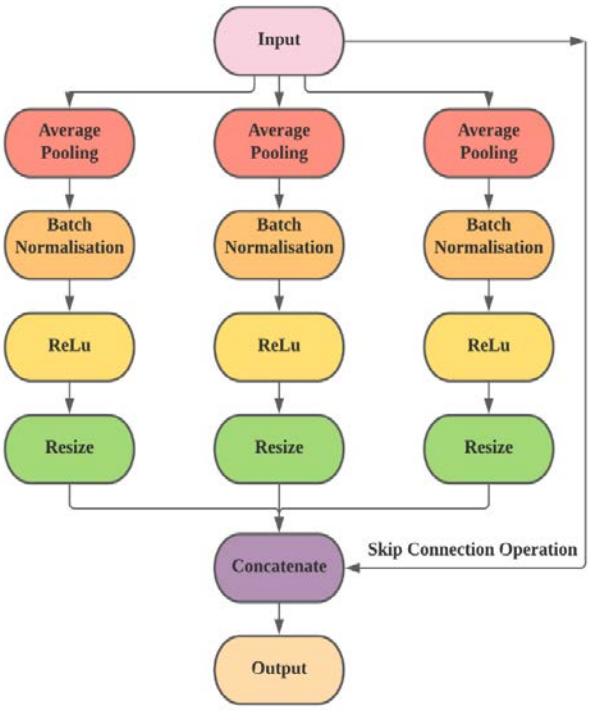


Fig. 9. General SPP Layer.

IV. Result and Discussion

A. Basic Settings

Google Colaboratory (Colab) was used as the platform to conduct the training and training process of the CNN model for micro-expression recognition for detecting a human's true emotion. Colab is a free cloud service developed by Google Research that enables the researcher to write and run Python code through the internet browser. It is a platform that is well suited for machine learning-based research. Another advantage of using Colab is it provides free GPU for the user that can be used for deep learning training. In this study, the programming language used is Python with Tensorflow library 2.4.1 through "Python three Google compute Engine Backend (GPU)" with Ram size of 12.72 GB and disk size of 68.40 GB. The virtual GPU used is Tesla P100-PCIE-16GB.

B. Hyperparameter

According to Lisha et al. [56] and Kandel et al. [57], hyperparameters can be considered as an input to a CNN algorithm which determines the performance of the deep learning algorithm towards the new and unseen data. Several hyperparameters that will be optimized in this work are learning rate, batch size, epochs, and type of optimizer. The type of optimizer determines how the weights are renewed by decreasing the loss or error [58]. In this study, Adamax optimizer is used as the sole optimization algorithm. This is because according to the research from [59], [60], Adamax produces a stable calculation method to renew the weights that ensure the stability of the CNN model.

Batch size is the number of data used for training of CNN model before the weights are updated. Smaller batch size can lead to slower convergence, while a larger batch size enables the CNN model to reach optimum minima. After performing several tests, a batch size of 64 is used throughout the experiment which achieves better stability and convergence compared to other batch sizes. Learning rate determines the rate of updating the weights. We are using 0.0001 as our learning rate for the CNN model. Jaya et al. [61] stated that the learning rate is not very high because a high value will cause the CNN architecture to become very unstable.

Epoch is defined as the number of iterations for a CNN model being trained by the whole datasets. Colab has a limitation that requires the user to interact with the system without idling by more than 90 minutes, after which it will stop the session automatically. Therefore, in [62] states that the number of epoch and learning rate need to be selected carefully so that the number of epoch can be minimized. This study has set the maximum number of epoch to be 120 iterations. Table III shows the summary of the other hyper-parameters that have been set in this study.

C. Dataset

Three ME datasets have been selected for this study, which are CASMEII [63], SMIC [64], and SAMM [65]. The first dataset, CASME II consists of 247 ME video clips from 26 subjects. The resolution of all videos is initially set to 640 x 480 pixels while the cropped image resolution is 340 x 280 pixels. Only five types of emotions are being considered, which are happiness, surprise, disgust, regression, and others [66]. The second dataset, SAMM involving 159 ME video clips from 29 subjects. The initial resolution of the videos is 2040 x 1080 pixels, while the cropped video resolution is 400 x 400 pixels. Rather than five emotion categories, SAMM consists of eight emotion classes, which are angry, contempt, disgust, fear, happiness, sadness, surprise, and others [67],[68]. The last dataset, SMIC comprises of 164 ME video clips from 16 subjects. The size of every image is 640 x 480 pixels. This dataset has three categories of emotion only that include positive, negative, and surprise [64], [66].

TABLE III. Hyperparameter Setting

Hyperparameter	**Value**
Learning Rate	0.0001
Batch Size	64
Epochs	120
Optimizer	ADAMAX

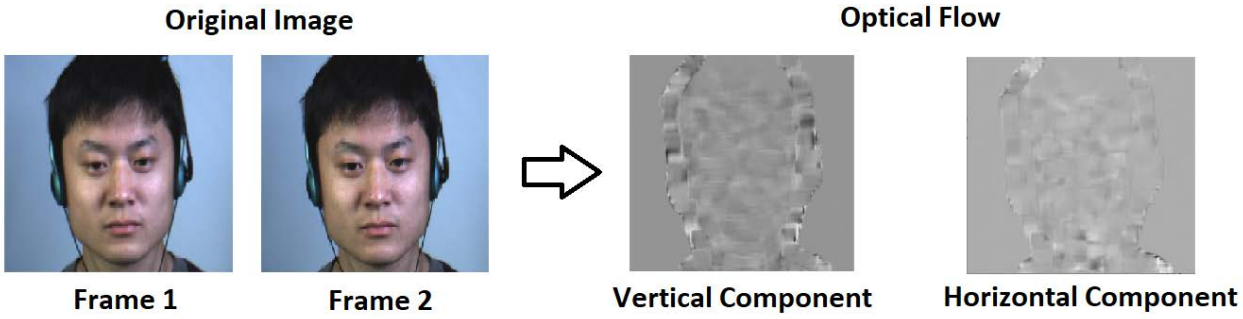


Fig. 10. Samples of Optical Flow between Two Consecutive Frames.

In this study, all images from all three datasets will be resized to 300 x 300 pixels, which will result in the same resolution for the respective vertical and horizontal components of the optical flow image. After that, the optical images will be down-scale again to the CNN input requirement, which is 75 x 75 pixels. According to Song and Zengfu [69], optical flow is the apparent motion between the video frame or image frame. It has a high-level feature in analyzing the visual motion information compared to the original image sequences, which allows it to have a better and more efficient data representation for ME [70]. Fig. 10 shows some examples of the calculated optical flow images, whereby the black and white color shows the presence of motion between the frames, while the grey color indicates that there is no motion for that respective pixels

The proposed system performance was verified by using 570 videos from 71 subjects that comprise of CASME II, SAMM, and SMIC datasets. There is no validation dataset used during the training phase. For the training and testing process, the dataset is divided according to leave-one-subject-out for testing, while the rest subjects will be used as training. This means that the model is trained using 70 subjects and 1 subject is used for testing. The output of this study will be labeled and classified into three classes: positive, negative, and surprise emotions. Positive emotion involves happy micro-expression which is considered as a "good" human emotion, while negative emotion is considered as a "bad" human emotion that can be further broken down into disgust, sadness, and fear. The third class of emotion, surprise is the emotion that a human expresses when he senses any difference between the expectation and the reality. Fig. 11, Fig. 12, and Fig. 13 portray an example of each type of emotion from different datasets.

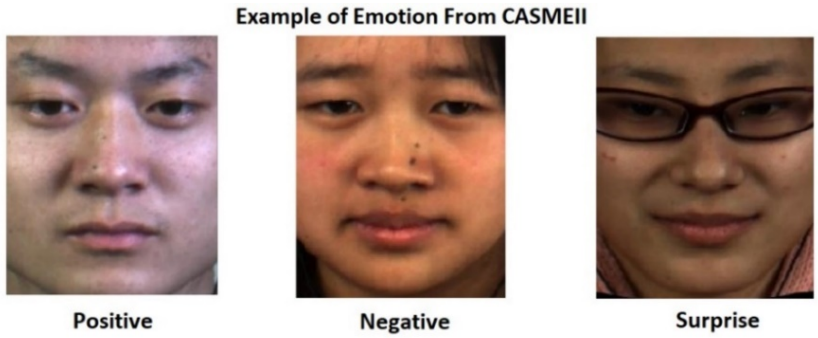


Fig. 11. Example of each Emotion from CASMEII.

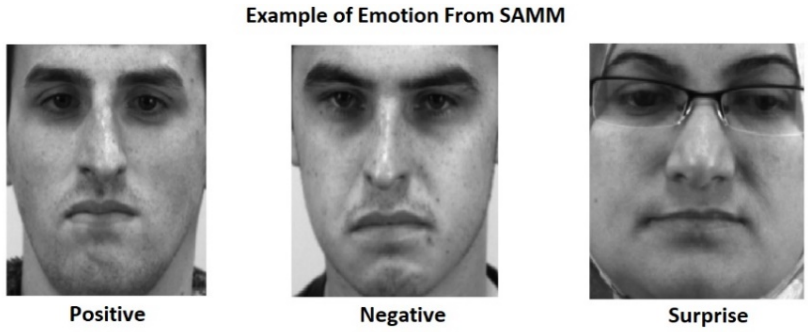


Fig. 12. Example of each Emotion from SAMM.

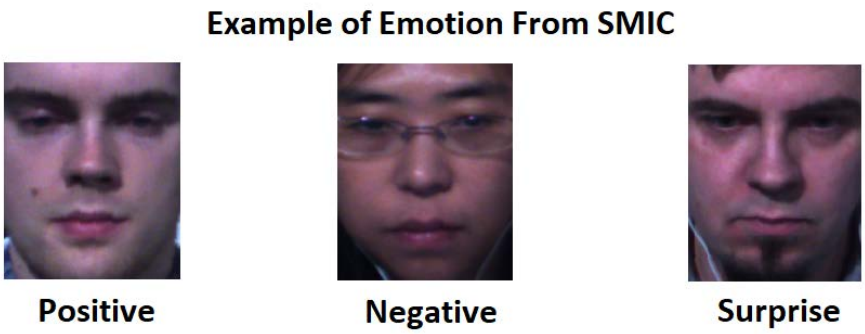


Fig. 13. Example of each Emotion from SMIC.

D. Evaluation Metric

The evaluation metric used in this study is the accuracy of micro-expression recognition for detecting a human's true emotion. According to Duygu [71], accuracy is defined as the ratio of true classification to total classification (true and false classification), which is formulated as in (4), whereby TP is defined as true positive, TN is true negative, FP is false positive, and FN is false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

E. Experimental Setting

Several training processes were performed based on the modified VGG-M architecture for micro-expression recognition to detect human's authentic emotions. For every experiment, the values of other hyperparameters are set to constant to make sure that the experiments are conducted in fair conditions. After that, the Spatial Pyramid Pooling (SPP) layer of various configurations (in terms of the number of layers, kernel size of average pooling, and position where SPP layers are added) will be embedded into the modified VGG-M architecture. The performances for each configuration will be compared and evaluated by computing their recognition accuracy by using combined datasets of CASMEII, SMIC, and SAMM. Fig. 14 shows the summary of the major procedures that were performed to extract the performance accuracy.

There will be eight variants of SPP architectures that will be tested as detailed out below:

- First Variant = two parallel layers, Kernel size: two, four, Position: After first Layer of VGG-M.
- Second Variant = two parallel layers, Kernel size: two, four, Position: After second Layer of VGG-M.
- Third Variant = three parallel layers, Kernel size: two, four, six, Position: After first Layer of VGG-M.
- Fourth Variant = three parallel layers, Kernel size: two, four, six, Position: After second Layer of VGG-M.
- Fifth Variant = four parallel layers, Kernel size: two, four, six, Position: After first Layer of VGG-M.
- Sixth Variant = four parallel layers, Kernel size: two, four, six, eight, Position: After second Layer of VGG-M.
- Seventh Variant = five parallel layers, Kernel size: two, four six, eight, ten, Position: After first Layer of VGG-M.
- Eighth Variant = five parallel layers, Kernel size: two, four, six, eight, ten, Position: After second Layer of VGG-M.

The SPP layers with the different number of SPP layers and kernel size of average pooling will be added in two ways, one of the ways is added after the first layer and another way is inserted after the second layer of the VGG-M module as shown in Fig. 15.

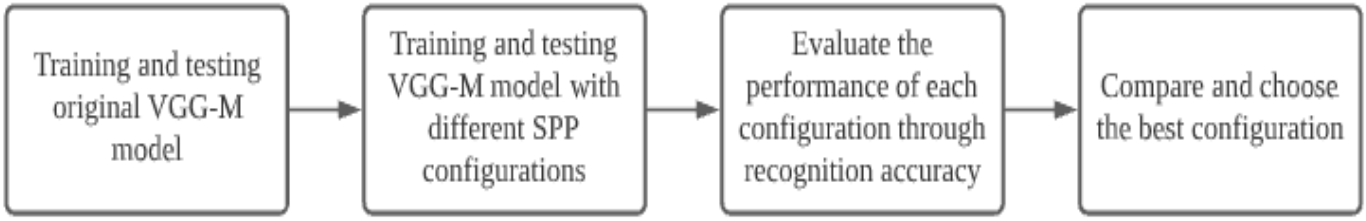


Fig. 14. Summary of the Experimental Procedures.

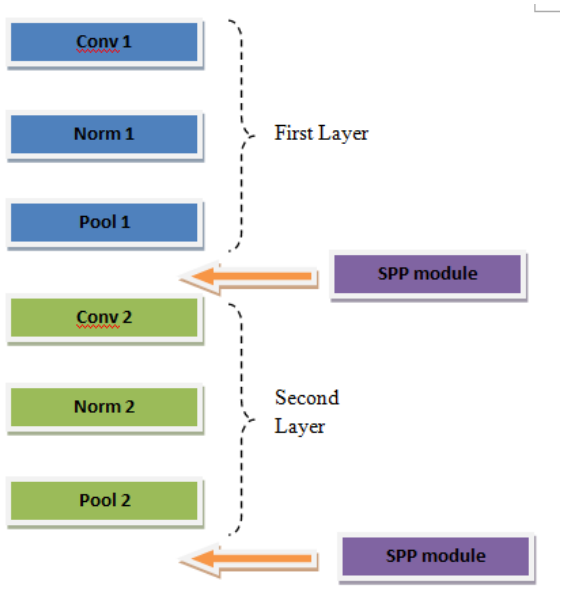


Fig. 15. Placement of the SPP Module.

F. Dataset Analysis

Table IV shows the results of modified VGG-M with several settings of parallel branches and placement of the SPP modules. Besides that, the results of combining both hyper-parameters are also reported in the same table. Fig. 16 shows the graph for training process of the original VGG-M and Fig. 17 shows the graph of the training results of VGG-M with SPP inserted after the second CNN layer.

TABLE IV. PERFORMANCE RESULTS OF VARIOUS SPP CONFIGURATIONS

Types of datasets	**Accuracy (%)**								
	Original (without SPP)	***After first layer***				***After second layer***			
		2 SPP	3 SPP	4 SPP	5 SPP	2 SPP	3 SPP	4 SPP	5 SPP
Combined	75.96	74.75	75.21	75.96	75.06	76.42	75.36	76.87	76.27
CASME II	86.21	84.37	86.67	83.45	86.67	88.05	83.91	86.67	86.67
SAMM	71.21	70.2	68.18	69.19	69.7	70.02	71.21	69.7	69.7
SMIC	70.73	69.92	70.73	74.8	69.11	71.14	71.14	73.98	72.36

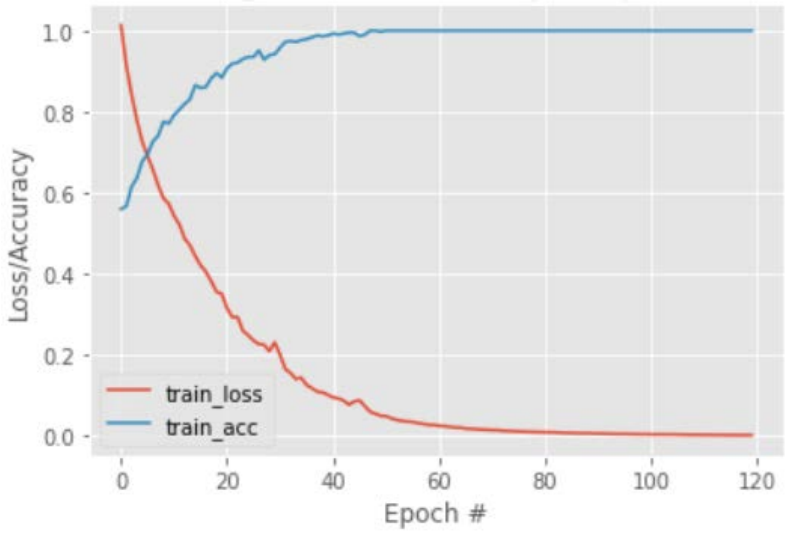


Fig. 16. The Original VGG-M Training Process.

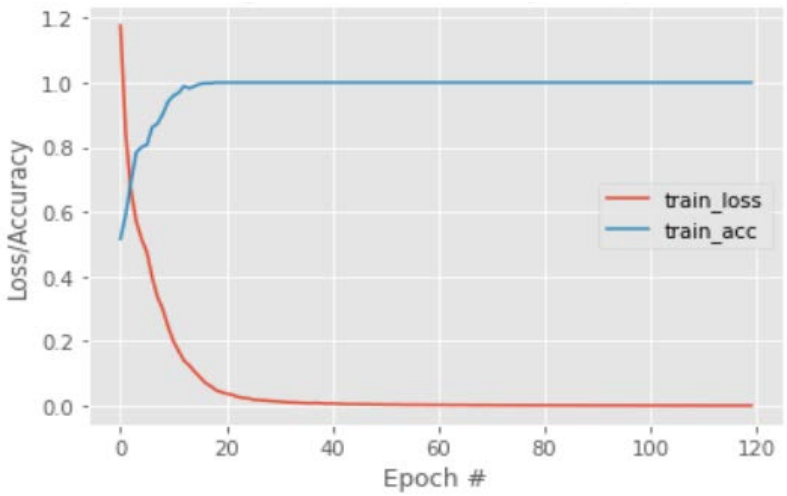


Fig. 17. VGG-M Training Process using SPP after the Second VGG-M Layer.

By optimizing the network configurations (different SPP parallel branches and module placement), the recognition accuracy of the proposed deep model has improved the recognition performance by four percent to seven percent when compared to the original network for the SAMM dataset. Meanwhile, for the SMIC dataset, the recognition accuracy of the model using the combined dataset has improved from 1% to 5%.

G. SPP Configuration and Placement

From Table IV, the results show that the recognition accuracy of the model using the combined dataset, CASME II and SMIC after the addition of the SPP module has improved the performance, except for the case of the SAMM dataset, in which the performance has become worst. Firstly, based on the combined dataset results, the best recognition accuracy is obtained with four SPP parallel layers, added after the second layer with an accuracy of 76.87%, which is an improvement of 0.91% compared to the base model.

For the CASME II dataset analysis, the best recognition accuracy is obtained by using two SPP parallel layers inserted after the second layer with 88.05% accuracies, which is an improvement of 1.84% compared to the base model. Among many datasets and configurations, this is the highest recognition accuracy obtained. While, for the case of the SAMM dataset, the addition of more SPP parallel layers into the base model does not increase the recognition accuracy, whereby the best number of SPP layers is three that is added after the second layer with a recognition accuracy of 71.21%.

On the other hand, for the SMIC dataset, four SPP parallel layers that are embedded after the first layer produced the highest recognition accuracy of 74.80% with an improvement of 4.07% compared to the base model. This is the greatest improvement in terms of recognition accuracy among many configurations that have been tested. Hence, the best setup for each dataset and the combined datasets are shown in Table V.

However, among all the configurations, SPP with four parallel branches produces the best general performance where it produces the best recognition accuracies for the combined datasets and SMIC dataset. Besides that, the best overall placement of the SPP module is if it is added after the second layer of the base model. On average, it produces results with higher recognition accuracy. In addition, most of the highest recognition accuracy is obtained after adding the SPP right after the second layer to the base model.

H. Improvement of the SPP Configuration

To further improve the base model performance, a few new variants of the SPP is introduced as follow:

- Ninth Variant = five parallel layers, Kernel size: two, three, four, five, six, Position: After first Layer of VGG-M
- Tenth Variant = five parallel layers, Kernel size: two, three, four, five, six, Position: After second Layer of VGG-M

TABLE V. SUMMARY OF THE BEST SETUP IN TERMS OF SPP CONFIGURATION AND PLACEMENT

Types of datasets	**Best configuration of SPP**	
	SPP number	***Position***
combined	Four	After second layer
CASME II	Two	After second layer
SMIC	Three	After second layer
SAMM	Four	After first layer

TABLE VI. PERFORMANCE BETWEEN DIFFERENT SETS OF KERNEL SIZE

Types of datasets	Accuracy (%)			
	After first layer		*After second layer*	
	first set Kernel	second set Kernel	first set Kernel	second set Kernel
Combined	75.06	74.91	76.27	76.42
CASME II	86.67	84.37	86.67	86.21
SAMM	69.70	71.21	69.70	70.20
SMIC	69.11	69.51	72.36	72.76

TABLE VII. SUMMARY OF THE BEST CONFIGURATION IN TERMS OF THE AVERAGE POOLING KERNEL SIZE FOR DIFFERENT DATASET SETUP

Table Head	**Table Column Head**
Combined	two, three, four, five and six
CASME II	two, four, six, eight and ten
SMIC	two, three, four, five and six
SAMM	two, three, four, five and six

The performance difference between the new variants of SPP with the earlier variants is due to the kernel sizes of average pooling, where the sizes are smaller for the latter variants. In short, the ninth and the tenth variants have bigger kernel sizes compared to the seventh and the eighth variants. Table VI shows the recognition accuracy of the modified model with different kernel sizes.

According to Table V, when adding the SPP layer after the first layer to the base model, the first set of kernel sizes produce better recognition accuracy compared to the second set for the combined datasets and CASME II dataset. On the other hand, it is the opposite trend for the SAMM and SMIC datasets. Besides that, if the SPP is added after the second layer, the second set of kernel sizes produces better accuracy compared to the first set for the combined, SAMM and SMIC datasets, except for the CASMEII dataset.

By comparing the overall results, the second set of kernel sizes result in higher average recognition accuracy for the test done on the combined, SAMM and SMIC dataset (76.42%, 71.21%, and 72.76% respectively). However, the first set of kernel sizes return the best recognition accuracy for the test done on the CASME II dataset (86.67 % accuracy). Hence, it can be concluded that the second set of average pooling kernel sizes is the better alternative compared to the first set as shown in Table VII.

V. CONCLUSION

This study has managed to improve the recognition accuracy of the CNN-based deep learning model by embedding SPP to the base model. Various configurations have been tested to find the optimal setup. For the combined dataset, the best SPP configuration is obtained by adding four parallel branches of the SPP after the second layer of the base model. This configuration has produced a recognition accuracy of 76.87%, which is an improvement of 0.91% over the base model. For the CASME II dataset, two parallel branches of the SPP layers added after the second layer of the base model have produced 88.05% recognition accuracy, which is an improvement of 1.84%. Meanwhile, the best SPP configuration for the SMIC dataset is four parallel branches added after the first layer with an improvement of 4.07%. There is not much performance improvement that can be observed with the addition of the SPP module when the test is done on the SAMM dataset. Generally, the overall best SPP configuration is by embedding four parallel branches of SPP with average pooling kernel sizes of two, three, four, five, and six, added after the second layer of the base model. For future works, more datasets can be combined to produce a more robust micro-expression-based automated emotion recognition system. Other than that, the dataset will be resampled using data augmentation methods to balance the class distribution between the emotion class. Besides that, synthetic data augmentation through the generative adversarial method can be employed to further increase the training samples.

ACKNOWLEDGMENT

The authors would like to acknowledge funding from Universiti Kebangsaan Malaysia (Geran Universiti Penyelidikan: GUP-2019-008) and Ministry of Higher Education Malaysia (Fundamental Research Grant Scheme: FRGS/1/2019/ICT02/UKM/02/1).

REFERENCES

[1] C. Correia-Caeiro, K. Guo, and D. S. Mills, “Perception of dynamic facial expressions of emotion between dogs and humans,” Animal Cognition, vol. 23, no.3, May 2020.

[2] Y. Huang, F. Chen, S. Lv, and X. Wang, “Facial Expression Recognition: A Survey,” Symmetry, vol. 11, no.10, p. 1189, Sep. 2019.

[3] B. Allaert, I. M. Bilasco, and C. Djeraba, “Micro and macro facial expression recognition using advanced Local Motion Patterns,” IEEE Transactions on Affective Computing, pp. 1–12, 2019.

[4] D. Deng, Z. Chen, Y. Zhou, and B. Shi, “MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no.03, Apr. 2020.

[5] D. Borza, R. Danescu, R. Itu, and A. Darabant, “High-Speed Video System for Micro-Expression Detection and Recognition,” Sensors, vol. 17, no.12, Dec. 2017.

[6] G. Zhao and X. Li, "Automatic Micro-Expression Analysis: Open Challenges," Frontiers in Psychology, vol. 10, Aug. 2019.

[7] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial Expression Recognition Based on Random Forest and Convolutional Neural Network," Information, vol. 10, no.12, Nov. 2019.

[8] S. O. Slim, A. Atia, M. M.A., and M.-S. M.Mostafa, "Survey on Human Activity Recognition based on Acceleration Data," International Journal of Advanced Computer Science and Applications, vol. 10, no.3, 2019.

[9] M. A. Zulkifley, M. M. Mustafa, A. Hussain, A. Mustapha, and S. Ramli, "Robust Identification of Polyethylene Terephthalate (PET) Plastics through Bayesian Decision," PLoS ONE, vol. 9, p. e114518, Dec. 2014.

[10] Y. Lai, "A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition," Journal of Physics: Conference Series, vol. 1314, p. 012148, Oct. 2019.

[11] Niall O' Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan and Joseph Walsh, "Deep Learning vs. Traditional Computer Vision," 2020.

[12] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," Journal of Network and Computer Applications, vol. 153, Mar. 2020.

[13] J. Wang, H. Wang, X. Zhu, and P. Zhou, "A Deep Learning Approach in the DCT Domain to Detect the Source of HDR Images," Electronics, vol. 9, Dec. 2020.

[14] Z. Ke, C. Le, and Y. Yao, "A multivariate grey incidence model for different scale data based on spatial pyramid pooling," Journal of Systems Engineering and Electronics, vol. 31, pp. 770–779, Aug. 2020.

[15] C. Dewi, R.-C. Chen, and S.-K. Tai, "Evaluation of Robust Spatial Pyramid Pooling Based on Convolutional Neural Network for Traffic Sign Recognition System," Electronics, vol. 9, May 2020.

[16] M. A. Zulkifley, N. A. Mohamed, and N. H. Zulkifley, "Squat Angle Assessment Through Tracking Body Movements," IEEE Access, vol. 7, pp. 48635–48644, 2019.

[17] Y.-H. Oh, J. See, A. C. le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges," Frontiers in Psychology, vol. 9, Jul. 2018.

[18] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns," Neurocomputing, vol. 175, Jan. 2016.

[19] J. He, J.-F. Hu, X. Lu, and W.-S. Zheng, "Multi-task mid-level feature learning for micro-expression recognition," Pattern Recognition, vol. 66, Jun. 2017.

[20] X. Huang and G. Zhao, "Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern," Oct. 2017.

[21] Li X, Xiaopeng H, Moilanen A., Huang X., Pfister T. and Zhao G., "Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods," IEEE Transactions on Affective Computing, vol. 9, Oct. 2018.

[22] S.-T. Liong, J. See, R. C.-W. Phan, K. Wong, and S.-W. Tan, "Hybrid Facial Regions Extraction for Micro-expression Recognition System," Journal of Signal Processing Systems, vol. 90, Apr. 2018.

[23] S. L. Happy and A. Routray, "Fuzzy Histogram of Optical Flow Orientations for Micro-Expression Recognition," IEEE Transactions on Affective Computing, vol. 10, Jul. 2019.

[24] A. C. le Ngo, J. See, and R. C.-W. Phan, "Sparsity in Dynamics of Spontaneous Subtle Emotions: Analysis and Application," IEEE Transactions on Affective Computing, vol. 8, Jul. 2017.

[25] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression Identification and Categorization Using a Facial Dynamics Map," IEEE Transactions on Affective Computing, vol. 8, Apr. 2017.

[26] Ping, L., Zheng, W., Ziyan, W., Qiang, L., Yuan, Z., and Minghai, X., et al., "Micro-Expression Recognition by Regression Model and Group Sparse Spatio-Temporal Feature Learning," IEICE Transactions on Information and Systems, vol. E99.D, 2016.

[27] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning From Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition," IEEE Transactions on Multimedia, vol. 20, no.11, pp. 3160–3172, Nov. 2018.

[28] X. Zhou, "Understanding the Convolutional Neural Networks with Gradient Descent and Backpropagation," Journal of Physics: Conference Series, vol. 1004, Apr. 2018.

[29] S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, "Deep Convolutional Neural Network Model for Automated Diagnosis of Schizophrenia Using EEG Signals," Applied Sciences, vol. 9, Jul. 2019.

[30] Y. Xie, W. Dai, Z. Hu, Y. Liu, C. Li, and X. Pu, "A Novel Convolutional Neural Network Architecture for SAR Target Recognition," Journal of Sensors, vol. 2019, pp. 1–9, May 2019.

[31] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, Aug. 2018.

[32] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction," Sensors, vol. 17, Apr. 2017.

[33] Yang Chen, Shengwu Qin, Shuangshuang Qiao, Qiang Dou, Wenchao Che, Gang Su, Jingyu Yao and Uzodigwe Emmanuel Nnanwuba., "Spatial Predictions of Debris Flow Susceptibility Mapping Using Convolutional Neural Networks in Jilin Province, China," Water, vol. 12, Jul. 2020.

[34] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," Journal of Big Data, vol. 6, Dec. 2019.

[35] S. Guo, T. Yang, W. Gao, C. Zhang, and Y. Zhang, "An Intelligent Fault Diagnosis Method for Bearings with Variable Rotating Speed Based on Pythagorean Spatial Pyramid Pooling CNN," Sensors, vol. 18, Nov. 2018.

[36] V. Suárez-Paniagua and I. Segura-Bedmar, "Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction," BMC Bioinformatics, vol. 19, Jun. 2018.

[37] S. Sharma and R. Mehra, "Implications of Pooling Strategies in Convolutional Neural Networks: A Deep Insight," Foundations of Computing and Decision Sciences, vol. 44, Sep. 2019.

[38] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," pp. 124–133, Nov. 2018.

[39] J. Feng and S. Lu, "Performance Analysis of Various Activation Functions in Artificial Neural Networks," Journal of Physics: Conference Series, vol. 1237, Jun. 2019.

[40] Y. Wang, Y. Li, Y. Song, and X. Rong, "The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition," Applied Sciences, vol. 10, Mar. 2020.

[41] Y. Yu, K. Adu, N. Tashi, P. Anokye, X. Wang, and M. A. Ayidzoe, "RMAF: Relu-Memristor-Like Activation Function for Deep Learning," IEEE Access, vol. 8, pp. 72727–72741, 2020.

[42] A. F. T. Martins and R. F. Astudillo, "From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification," Feb. 2016.

[43] A. Aldhahab, S. Ibrahim, and W. Mikhael, "Stacked Sparse Autoencoder and Softmax Classifier Framework to Classify MRI of Brain Tumor Images," International Journal of Intelligent Engineering and Systems, vol. 13, Jun. 2020.

[44] S. Das and J. Mukherjee, "Automatic License Plate Recognition Technique using Convolutional Neural Network," International Journal of Computer Applications, vol. 169, Jul. 2017.

[45] S. Samir, E. Emary, K. El-Sayed, and H. Onsi, "Optimization of a Pre-Trained AlexNet Model for Detecting and Localizing Image Forgeries," Information, vol. 11, May 2020.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, May 2017.

[47] S. R. Abdani, M. A. Zulkifley, and N. Hani Zulkifley, "A Lightweight Deep Learning Model for COVID-19 Detection," in 2020 IEEE

Symposium on Industrial Electronics & Applications (ISIEA), pp. 1–5, Jul. 2020

[48] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," pp. 121–129, May 2014.

[49] Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda and Jean M Vettel., "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," Journal of Neural Engineering, vol. 15, Dec. 2018.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Jun. 2016.

[51] S. R. Abdani and M. A. Zulkifley, "DenseNet with Spatial Pyramid Pooling for Industrial Oil Palm Plantation Detection," in 2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE), pp. 134–138, Dec. 2019.

[52] H.-J. Jiang, Y.-A. Huang, and Z.-H. You, "Predicting Drug-Disease Associations via Using Gaussian Interaction Profile and Kernel-Based Autoencoder," BioMed Research International, vol. 2019, Aug. 2019.

[53] I. Omara, X. Wu, H. Zhang, Y. Du, and W. Zuo, "Learning pairwise SVM on hierarchical deep features for ear recognition," IET Biometrics, vol. 7, Nov. 2018.

[54] Z. Li, F. Li, L. Zhu, and J. Yue, "Vegetable Recognition and Classification Based on Improved VGG Deep Learning Network Model," International Journal of Computational Intelligence Systems, vol. 13, p. 559, 2020.

[55] S. R. Abdani, M. A. Zulkifley, and M. Mamat, "U-Net with Spatial Pyramid Pooling Module for Segmenting Oil Palm Plantations," Sep. 2020.

[56] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," Mar. 2016.

[57] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," ICT Express, vol. 6, Dec. 2020.

[58] N. M. Aszemi and P. D. D. Dominic, "Hyperparameter Optimization in Convolutional Neural Network using Genetic Algorithms," International Journal of Advanced Computer Science and Applications, vol. 10, no.6, 2019.

[59] D. Yi, J. Ahn, and S. Ji, "An Effective Optimization Method for Machine Learning Based on ADAM," Applied Sciences, vol. 10, no.3, Feb. 2020.

[60] D.-S. Kwon, C. Jin, M. Kim, and W. Koo, "Mooring-Failure Monitoring of Submerged Floating Tunnel Using Deep Neural Network," Applied Sciences, vol. 10, no.18, Sep. 2020.

[61] J. T. Hardinata, H. Okprana, A. P. Windarto, and W. Saputra, "Analisis Laju Pembelajaran dalam Mengklasifikasi Data Wine Menggunakan Algoritma Backpropagation," J-SAKTI (Jurnal Sains Komputer dan Informatika), vol. 3, no.2, Sep. 2019.

[62] S. S. Liew, M. Khalil-Hani, S. Ahmad Radzi, and R. Bakhteri, "Gender classification: a convolutional neural network approach," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 24, 2016.

[63] Yan W.J, Li X.; Wang S.J; Zhao G, Liu Y.J; Chen Y.H. and Fu X., "CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation," PLoS ONE, vol. 9, no.1, Jan. 2014.

[64] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A Spontaneous Micro-expression Database: Inducement, collection and baseline," Apr. 2013.

[65] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A Spontaneous Micro-Facial Movement Dataset," IEEE Transactions on Affective Computing, vol. 9, no.1, Jan. 2018

[66] C. Guo, J. Liang, G. Zhan, Z. Liu, M. Pietikainen, and L. Liu, "Extended Local Binary Patterns for Efficient and Robust Spontaneous Facial Micro-Expression Recognition," IEEE Access, vol. 7, 2019.

[67] D. Y. Choi and B. C. Song, "Facial Micro-Expression Recognition Using Two-Dimensional Landmark Feature Maps," IEEE Access, vol. 8, 2020.

[68] J. Li, C. Soladie, and R. Seguier, "A Survey on Databases for Facial Micro-Expression Analysis," 2019.

[69] S. Wang and Z. Wang, "Optical Flow Estimation with Occlusion Detection," Algorithms, vol. 12, no.5, May 2019.

[70] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition," Frontiers in Psychology, vol. 8, Oct. 2017.

[71] D. Kaya, "Optimization of SVM Parameters with Hybrid CS-PSO Algorithms for Parkinson's Disease in LabVIEW Environment," Parkinson's Disease, vol. 2019, May 2019.

Automated Telugu Printed and Handwritten Character Recognition in Single Image using Aquila Optimizer based Deep Learning Model

Vijaya Krishna Sonthi[1]
Research Scholar
Department of CSE, FEAT
Annamalai University
Chidambaram, India

Dr. S. Nagarajan[2]
Associate Professor
Department of CSE, FEAT
Annamalai University
Chidambaram, India

Dr. N. Krishnaraj[3]
Associate Professor
School of Computing, SRM Institute
of Science & Technology
Kattankulathur, India

***Abstract*—Machine printed or handwritten character recognition becomes a major research topic in several real time applications. The recent advancements of deep learning and image processing techniques can be employed for printed and handwritten character recognition. Telugu character Recognition (TCR) remains a difficult task in optical character recognition (OCR), which transforms the printed and handwritten characters into respective text formats. In this aspect, this study introduces an effective deep learning based TCR model for printed and handwritten characters (DLTCR-PHWC). The proposed DLTCR-PHWC technique aims to detect and recognize the printed as well as handwritten characters that exist in the same image. Primarily, image pre-processing is performed using the adaptive fuzzy filtering technique. Next, line and character segmentation processes are performed to derive useful regions. In addition, the fusion of EfficientNet and CapsuleNet models is used for feature extraction. Finally, the Aquila optimizer (AO) with bi-directional long short-term memory (BiLSTM) model is utilized for recognition process. A detailed experimentation of the proposed DLTCR-PHWC technique is investigated using Telugu character dataset and the simulation outcome portrayed the supremacy of the proposed DLTCR-PHWC technique over the recent state of art approaches.**

Keywords*—Optical character recognition; Telugu; deep learning; Aquila optimizer; BiLSTM; handwritten characters; printed characters

I. INTRODUCTION

Recognition of hand printed or machine printed documents plays a significant role in applications such as text-speech converter, automated language-language translator, and intelligent scanning machines [1]. The aim of document image analyses is to identify the graphics and text modules in the paper documents and for extracting the proposed data, as human beings do. The two modules of document image analyses are textual and graphical processing. The former textual processing handles the text element of the document images and the graphical processing handles symbol and non-textual line elements which form line diagrams, delimit straight lines among company logos and text sections, and so on [2]. The image processing method is utilized for recognizing the handwritten Telugu character. Telugu is widely speaking in Telangana and Andhra Pradesh in India.

Optical character recognition (OCR) is a procedure which transforms the handwritten or printed files to their equivalent text format [3]. The OCR can be separated into two classes: online and offline character identification. Offline character identification is additionally separated into two components i.e., machine and handwritten printed character identification. In hand-written character recognition, there is a large number of problems in comparison with machine printed documents [4]. It is a challenging and fascinating field of pattern detection using many real-time applications. Many commercial OCR systems are accessible for printed Arabic characters however they have a lot of technological issues, particularly in the segmentation phase the result isn't enviable [5]. During the past years, OCR has become increasingly important since the need for translating the scanned image into computer identifiable forms like text documents has improved application. The OCR challenging problems such as distortion, lighting variations, variance in font size, and blurring of the printed character images, have improved the demands for OCR in the study work. The significant disadvantage witnessed in OCR is that the infra-class variation is larger because of the huge amount of images accessibility for the processing [6]. The OCR system is currently under development for most of the common languages and Telugu is no exemption for it. The OCR process has become really hard for Telugu and possess its individual challenges to the developer of Telugu character recognition (TCR) system. In recent times, significant studies have been made toward the development of an effective TCR scheme [7].

This study focuses on the design of effective deep learning based TCR technique for printed and handwritten characters (DLTCR-PHWC) in the same image. The DLTCR-PHWC involves adaptive fuzzy filtering (AFF) technique for image pre-processing. Besides, line and character segmentation processes are performed to derive useful regions. Moreover, the fusion of EfficientNet and CapsuleNet models is used for feature extraction. Furthermore, the Aquila optimizer (AO). With bidirectional long short-term memory (BiLSTM) model is applied for recognition process. A comprehensive experimental analysis takes place using Telugu character dataset a detailed comparative analysis is carried out in terms of different evaluation parameters.

The rest of the paper in Section 2 discusses literature review, Section 3 briefs proposed model and Section 4 discusses results and discussion and Section 5 ends up with conclusion.

II. Literature Review

Prameela et al. [8] proposed an OCR method for Telugu documents that consists of 2 steps, i.e. classification, preprocessing, and feature extraction. In preprocessing phase, they used median filtering on the input character and employed skeletonization and normalization methods over character for extracting boundary edge pixel points. In feature extraction phase, first, all the characters are separated into 3×3 grids and the equivalent centroid for every 9 regions is calculated. Cheekati and Rajeti [9] deal in emerging a reliable, fast Telugu hand-written ResNet for offline and online character identification and improve classification accuracy. The method is estimated by IIITS-Telugu Handwriting Dataset; HP Lab databases (Telugu) India and attained a remarkable result. The presented residual net (ResNet-50) attains 2.37% error on ResNet-18 & 34 test set.

Lakshmi and Babu [10] perform a new Telugu script identification and retrieval method named HCH model. Hash coding would be utilized as a feature extractor and the hamming distance would be used as a replacement for traditional Euclidean distance for measuring the similarity among database and query images. Experimental analyses exhibit that the presented model has outstanding performances to the traditional methods proposed in the survey. Burra et al. [11] proposed 2 methods for improving the glyph/symbol segmentations in a Telugu OCR scheme. The main features having an effect on the entire performance of a Telugu OCR scheme can be able to divide/segment scanned document images into identifiable units. In Telugu, this unit is generally interconnected component and is known as glyphs. Once a document is removed, interconnected components-based algorithm for segmentation fails. They provide malformed glyphs (a) partial and results of break in the character because of uneven dispersal of ink on the noise /page, and (b) are an integration of more than two glyphs due to smudging in noise/print. The previous one is labeled broken and the last one is merged character. Hebbi et al. [12] introduce a cross language environment for recognizing the words and characters of lower resources script that is script which doesn't have typical database and the dataset isn't accessible for public access. Indic script comes from a popular origin and few scripts have a standard 3 regional structures. Identification of this script could be performed using another script having same structures. In order to identify this character, the models are trained using source language Kannada with zone wise testing and training is made by Kannada and the targeting language Telugu.

Rani et al. [13] proposed a method for feature classification and extraction of Telugu hand-written script based personalized template matching method through caching method for achieving better results. The method of caching is executed by main databases with a cache database maintain the frequently employed character template for a set of each character template. The XML databases are utilized in the class for different character templates and the class representation is given by new class structures depending on XML tags. In Sarika and Sirisala [14], OCR methods like preprocessing, digitization, feature extraction, recognition, and segmentation were discussed. HWCR with distinct ML methods like Bayesian decision theory, SVM, and Bayesian classifier was discussed. The current technique of HWCR is examined for native language and related their features and functionality.

Madhavi et al. [15] proposed an effective model named TR for correcting and detecting slant angles of MTW. Telugu language is India's common language speaking around 80 million people. The difficult character is attached with further marks called "vatthus" and "maatras" it is stimulating for detecting a slant angle. The presented TR model performs pre-processing and identifies interconnected components within the provided MTW. Later, estimate the slant angle of every interconnected component by acquiring interconnected slant lines on the boundary of every interconnected component. BJ et al. [16] handle classification and recognition of confusing Kannada characters. The RF and SVM as classifiers to categorize the confusing character. The presented method attained 78% of classifiers accuracy. Lastly, this method identifies the confusing characters with template matching and feature value outcomes-based classifier.

III. The Proposed TCR Model

In this study, a new DLTCR-PHWC technique is derived to detect and recognize the printed as well as handwritten characters that exist in the same image. The DLTCR-PHWC technique encompasses different operations such as AFF based pre-processing, line/paragraph segmentation, fusion-based feature extraction, BiLSTM based recognition, and AO based hyperparameter optimization. Fig. 1 showcases the overall block diagram of proposed DLTCR-PHWC model. The detailed working of every module is offered in the following subsections.

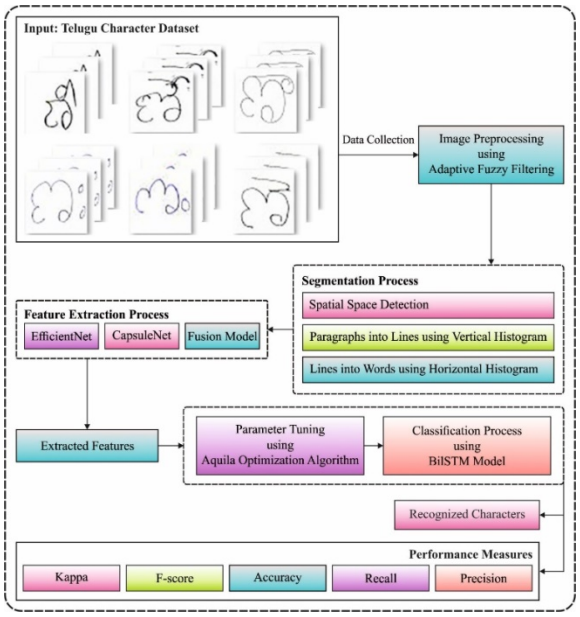


Fig. 1. Overall Process of DLTCR-PHWC Model.

A. Pre-Processing

In this model, the input receives image as input for removing the presence of noise. Once the image is reduced by high noise and the whole pixels in filter window are similar to the maximum values, it increases the window size up to a predefined maximal size $W_{\max}$, it changes $I(x,y)$ by the median values i.e., estimated from filter window. In the process of finding a median value, each pixel value is similar to the maximum values in filter window, through evaluated pixel value in filter window except the maximum values to obtain median values.

B. Segmentation Process

Segmentation is a major phase in detection method since it extracts significant fields for additional analyses. Typically, it is employed for verifying the boundary and objects such as lines, curves, and so on, in an image. The scanned images are segmented to the paragraph via spatial space recognition method, paragraph to lines through vertical histogram, lines to words through horizontal histogram [17]. Accurateness of character identification is based largely on segmentation performance. Mainly, this procedure has the subsequent steps:

- Recognize the text line on the page.
- Recognize single character in all the words.

The widely employed technique for line segment of the grayscale images is the projection profile model. By adding it to the horizontal course of the documents, gaps among the text line could be recognized through detecting the projection value. They employ horizontal projection profiles analyses since the text in most document images is arranged with the horizontal line. This method determines horizontal projection histogram, the amount of black pixel for every column of the raster images. Once these profiles are employed on M × N images, a column vector of M × 1 sizes were attained. Element of these column vectors is the amount of pixel value in every row of the document images. Afterward, the line segmentation the resulting outputs are fed into the character segments. Finally, the extracted line is segmented to the character. In order to detect the boundaries among the characters, they employ threshold values on the length of the space amongst the characters. Afterward detecting the position of the space among the characters they remove the part of the line segmentation. As per the abovementioned procedure, the character and lines are segmented from the pre-processed documents. Afterward the segmentation procedure the resulting outputs are fed into the feature extraction.

C. Fusion based Feature Extraction Technique

At this stage, the segmented images undergo feature extraction process using CapsNet and EfficientNet models. The underlying framework of the CapsNet, adapted in this study, is made up of 1 fully connected layer and two convolution layers. The 1st convolution layer is generated with 256 9 × 9 kernels using a stride of one and ReLU activation, deliver feature map which is additionally fed into the initial set of capsules in the abovementioned layers. The 2nd layers represent the initial capsule and accommodate thirty-two channels, therefore eight-dimensional convolution capsules of 9 × 9 kernels and a stride of two, in which every capsule captures each unit in the 1st convolution layer that receptive field overlaps with the center of capsules. Therefore, the set of initial capsules output an overall of 32 × 6 × 6 eight dimensional vectors, and the capsule of the similar grid share their weight with one another. The final layer of the CapsNet is an FC layer of 2 sixteen dimensional units which is interconnected to each capsule in the prior layer [18]. As the output of 1st convolution layer is 1D, it doesn't transfer the similar quality of data as capsule in the abovementioned layers (viz., the output of 1st layer doesn't offer orientation attribute to agree on), no routing is proposed with the initial capsule. It is notable that each logit Bij are initiated as zero that indicates the primary capsule output is transmitted to each possible parent capsule with equivalent probabilities Cij. In another word, previous to knowledge optimization, the initial capsule assumes an equivalent agreement with parent capsule (for example each entity tied with initial capsule is related to the entity tied to above parent capsule). Fig. 2 illustrates the framework of CapsNet model.

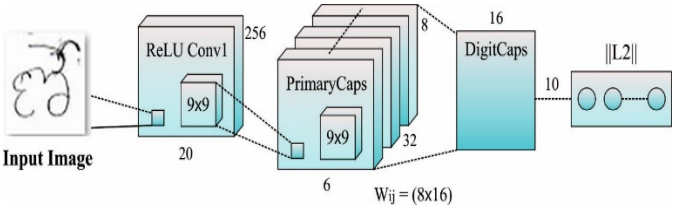


Fig. 2. Structure of CapsNet.

With respect to the loss function for the networks to learn, it consists in extending the length of instantiation vectors V_k for an entity class (car and solar panel in these cases) given that certain entity is witnessed in an image. This could be expanded for several entities through a single margin loss L_k for all entities k capsule :

$$L_k = T_k \max(0, m^+ - \|V_k\|)^2 + \lambda(1 - T_k) \max(0, \|V_k\| - m^-)^2 \quad (1)$$

Whereas T_k is fixed to 1 when the entity is existing.

In recent years, the rapid growth of DL method has spawned many excellent CNN models. From the initial simple network to the current complex network, the performance of the model is getting better and better in every aspect. EfficientNet combines the advantage of previous network, which summarize the development of network performances into 3D: (1) Deepen the network, i.e., use the skip connection to increase the depth of the neural networks, and attain feature extraction via deeper layer; (2) Widen the network, i.e., increase the amount of convolution layers to attain more functions and features; (3) By increasing the input image resolution, the network could express and learn more things, which is beneficial to improve accuracy. Then, use a compound coefficient ϕ to uniformly balance and scale the resolution, depth, and width of the networks, as well maximize the network performance on limited resources. Estimation of the compound coefficient is given in Eq. (1):

$$depth: d = \alpha^{\phi} \quad width: w = \beta^{\phi} \quad resolution: r = \gamma^{\phi}$$
$$s.t. \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \quad (2)$$

$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$

Whereas d, w, and r represent the coefficient used to scale the resolution, depth, and width of the networks [19]. The α, β, & γ denotes resource allocation for network depth, width, and resolution. EfficientNet largely consists of Stem, 16 Blocks, Conv2D, GlobalAveragePooling2D, and Dense layer. The design of Block is based mainly on the attention mechanism and residual structure, also other structures are similar to traditional CNN model.

The extracted features from both DL models are fused together. During the fusion process, it is needed to decide the value of λ, which calls the R feature fusion. The fusion of features can be defined as follows.

$$NF = \lambda \cdot LF + (1 - \lambda) \cdot HF, \quad (3)$$

where NF is the fusion feature, and LF and HF indicate the features derived by the CapsNet and EfficientNet models.

D. Design of AO based BiLSTM Model for Recognition Process

Finally, the BiLSTM model receives the feature vectors as input and carried out the recognition process. The LSTM refers to a special RNN model that resolves the problems of gradient vanishing of the RNN by presenting a threshold mechanism and memory unit [20]. But x represent the network input at distinct times, y indicates the network output, h represents the hidden layer, u signifies the weight from input to the hidden layers, w denote the weights of prior node hidden layer to the present node hidden layer, and v represents the weight from hidden to the output layers.

In the actual execution of the LSTM, the LSTM units are upgraded at time t as follows:

$$i_t = \sigma(w_i h_{t-1} + U_i x_t + b_t) \quad (4)$$

$$f_t = \sigma\left(w_j h_{t-1} + U_f x_t + b_f\right) \quad (5)$$

$$\tilde{c} = tanh(W_c h_{t-1} + U_c x_t + b_c) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (7)$$

$$o_t = \sigma(w_o h_{t-1} + U_o x_t + b_o) \quad (8)$$

$$h_t = o_{t-1} \odot tanh(c_t) \quad (9)$$

Now, σ represents the sigmoid function and $\odot$ indicates the equivalent product of the element. x_t denotes the input vector at time t. h_t signifies the hidden state vector, that is called output vector, and store each data at time t and the preceding time. U_i, U_f, U_c, U_o represent the weight of input vector xt for the input, forgotten, unit, and output gates, correspondingly. W_i, W_f, W_c, W_o denote the weight of dissimilar gates to the hidden state vector h_t. b_t, b_f, b_c, b_o denotes the off-set vector. Using the 3 gates structure, the LSTM allows the recurrent network to maintain the beneficial data for the task in the memory unit at the time of the training model, thus evading the problems of the RNN disappearing while obtaining long range data.

While processing sequence data, the BLSTM introduces further backward estimation processes, i.e., unlike normal LSTM case. This procedure could employ the succeeding data of the sequence. Finally, the reverse and forward evaluations were implemented. The value is output to the output layer concurrently; as a result, each data of a sequence is attained by two directions that is used for multiplying type of natural language processing task.

Followed by, the hyperparameter tuning of the BiLSTM model take place using AO technique. The Aquila is most famous bird of prey. Young Aquila generally attains entire assurance during the fall, subsequent that it can be moved extremely for building territory to themselves. Because of their hunting bravery, Aquila is a most considered bird globally. The Male Aquila become considerably further prey if the solo-hunting. An essential stimulus to the presented AO technique has been resultant in the techniques stated above. The subsequent subsections that define these procedures are modeled in AO. In AO, it can be population-oriented technique, the optimized rules start with population of candidate solution (X) as projected in Eq. (9) that has been created stochastically amongst the upper boundary (UB) and lower boundary (LB) of the provided issues. An optimum solution obtained during all iterations can be defined as follows.

$$x = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & x_{1,Dim-1} & x_{1,Dim} \\ x_{2,1} & \cdots & x_{2,j} & \cdots & x_{2,Dim} \\ \cdots & \cdots & x_{i,j} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N-1,1} & \cdots & x_{N-1,j} & \cdots & x_{N-1,Dim} \\ x_{N,1} & \cdots & x_{N,j} & x_{N,Dim-1} & x_{N,Dim} \end{bmatrix} \quad (10)$$

where X refers the group of present candidate solutions that are created arbitrarily with utilizing Eq. (10), X_i indicates the decision values (places) of i^{th} solutions, N represents the entire amount of candidate solution (population), and Dim signifies the dimensional of the issue.

$$X_{ij} = rand \times \left(UB_j - LB_j\right) + LB_j, i = 1,2, \ldots\ldots, Nj = 1,2, \ldots, Dim \quad (11)$$

where $rand$ denotes the arbitrary number, LB_j demonstrates the j^{th} lower bound, and UB_j implies the j^{th} upper bound of provided issue [21]. The AO technique is transmission in exploration to exploitation stages utilizing distinct performance dependent upon this form if $t \leq \left(\frac{2}{3}\right) * T$ the exploration stages were excited; else, the exploitation stages are implemented. The mathematical process of the AO has been presented. During the primary technique (X_1), the Aquila distinguishes the prey region and elects an optimum hunting region by great soar with vertical stoop. At this point, the AO extremely explorers in great soar for determining the region of search space in which the prey is. This performance has been mathematically projected as in Eq. (12)

$$X_1(t+1) = X_{best}(t) \times \left(1 - \frac{t}{T}\right) + (X_M(t) - X_{best}(t) * rand), \quad (12)$$

where, $X_1(t+1)$ refers the solution of subsequent round of t that is created by initial searching technique (X_1). $X_{best}(t)$ signifies the optimum attained solution until t^{th}

iteration, this reproduces the estimated prey's position. The formula $\left(\frac{1-t}{T}\right)$ has been utilized for controlling the extended search (exploration) with the amount of iterations. $X_M(t)$ defines the places mean value of present solutions associated at t^{th} iteration that is computed utilizing in Eq. (12). $rand$ indicates the arbitrary value amongst [0, 1]. r and T demonstrate the present round and the maximal rounds, correspondingly.

$$X_M(t) = \frac{1}{N}\sum_{i=1}^{N} X_j\,(t), \forall j = 1,2,\dots,Dim \quad (13)$$

where Dim implies the dimensional size of issue and N represents the population size.

During the second technique (X_2), if the prey area has been initiated in a great soar, the Aquila circle on the target, arranges the land, and next attack. This performance was mathematically processed as in Eq. (14).

$$X_2(t+1) = X_{best}(t) \times Levy(D) + X_R(t) + (y-x)^{*} rand, \quad (14)$$

where $X_2(t+1)$ signifies the solution of next iteration of r that has been created by the next searching technique (X_2). D implies the dimensional space, and $Levy(D)$ demonstrates the levy flight distribution function that was computed utilizing in Eq. (15). $X_R(t)$ represents the arbitrary solution obtained from the range of $[1N]$ at i^{th} iteration.

$$Levy(D) = s \times \frac{u \times \sigma}{|v|^{\frac{1}{\beta}}} \quad (15)$$

where s indicates the constant values set to 0.01, u and v denotes the arbitrary numbers amongst [0, 1]. σ signifies the computed utilizing in Eq. (16).

$$\sigma = \left(\frac{\Gamma(1+\beta) \times sine\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2\left(\frac{\beta-1}{2}\right)}\right) \quad (16)$$

In Eq. (14), y and x are utilized for presenting the spiral shape during the searching process that is computed below:

$$y = r \times cos(\theta) \quad (17)$$

$$x = r \times sin(\theta) \quad (18)$$

where,

$$r = r_1 + U \times D_1 \quad (19)$$

$$\theta = -\omega \times D_1 + \theta_1 \quad (20)$$

$$\theta 1 = \frac{3 \times \pi}{2} \quad (21)$$

r_1 gets the value amongst [1, 20] to set the amount of search cycles, and U refers the lesser value set to 0.00565. D_1 defines the integer number in 1 to the length of search spaces (Dim), and ω indicates the lesser value set to 0.005. During the third approach (X_3), if the prey area has been identified perfectly, and the Aquila has been prepared to land and attacks, the Aquila inclines vertically with initial attack for discovering the prey reaction. This technique is named minimum flight. At this point, the AO utilizes the chosen region of the target for getting nearby prey as well as attack. This performance has been mathematically processed in Eq. (22).

$$X_3(t+1) = \left(X_{best}(t) - X_M(t)\right) \times \alpha - rand + \left((UB - LB) \times rand + LB\right) \times \delta, \quad (22)$$

where $X_3(r+1)$ implies the solution of succeeding round of r that is created by 3rd searching manner (X_3). $X_{best}(t)$ signifies the estimated place of prey till i^{th} iteration (the optimum attained solution), and $X_M(t)$ represents the mean value of present solution at r^{th} iteration that has been computed utilizing in Eq. (12). $rand$ stands for the arbitrary value amongst [0, 1]. During the fourth technique (X_4), if the Aquila became nearby the prey, the Aquila attack the prey on the land based on its stochastic movement. This process was mathematically projected as in Eq. (23).

$$X_4(t+1) = QF \times X_{best}(t) - (G_1 \times X(t) \times rand) - G_2 \times Levy(D) + rand \times G_1, \quad (22)$$

where $X_4(t+1)$ implies the solution of next iteration of t that has been created by the fourth search technique (X_4). QF demonstrates the quality function utilized for equilibrium the search approaches that are computed.

IV. Performance Validation

The performance validation of the DLTCR-PHWC technique is tested using benchmark Telugu character dataset. The results are investigated under different folds and a comprehensive comparative analysis is also performed. Fig. 1 shows the sample segmented output of the image comprising handwritten and printed characters. The figure stated that the DLTCR-PHWC technique has offered effective identification of characters exist in the applied image shown in Fig. 3.

Table I and Fig. 4 offers the detailed recognition performance of the DLTCR-PHWC technique under ten folds. The results stated that the DLTCR-PHWC technique has gained effectual outcomes on all the applied folds. For example, with F_1, the DLTCR-PHWC approach has obtained a precision of 98.53%, recall of 98.37%, accuracy of 98.24%, F1-measure of 98.43%, and kappa of 97%. Also, with F_2, the DLTCR-PHWC approach has gained a precision of 99.14%, recall of 97.81%, accuracy of 99.30%, F1-measure of 99.24%, and kappa of 96.10%. Moreover, with F_4, the DLTCR-PHWC approach has attained a precision of 98.48%, recall of 98.68%, accuracy of 99.14%, F1-measure of 97.54%, and kappa of 97.29%. Furthermore, with F_6, the DLTCR-PHWC approach has achieved a precision of 99.05%, recall of 99.42%, accuracy of 98.91%, F1-measure of 98.70%, and kappa of 97.89%. Concurrently, with F_8, the DLTCR-PHWC approach has gained a precision of 99.31%, recall of 98.16%, accuracy of 99.57%, F1-measure of 97.97%, and kappa of 98.30%. Lastly, with F_1, the DLTCR-PHWC algorithm has provided a precision of 99.58%, recall of 99.82%, accuracy of 99.78%, F1-measure of 98.03%, and kappa of 99.34%.

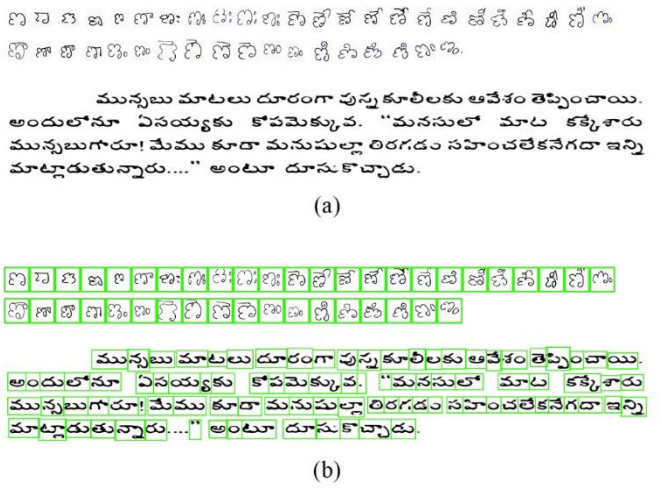


Fig. 3. Sample Segmentation Results (a) Original Handwritten & Printed Characters, (b) Segmented Handwritten & Printed Characters.

Finally, a detailed comparative results analysis of the DLTCR-PHWC technique takes place in Table II [22-26]. Fig. 5 provides a brief precision analysis of the DLTCR-PHWC technique with existing techniques. The figure reported that the MLP-HMM and DNN techniques have obtained worse outcomes with the least precision of 0.8467 and 0.9089 respectively. In addition, the CNN, KNN, MLP, and NN techniques have obtained moderately closer precision of 0.9567, 0.9415, 0.9572, and 0.9634 respectively. However, the DLTCR-PHWC technique has resulted in a higher precision of 0.9889.

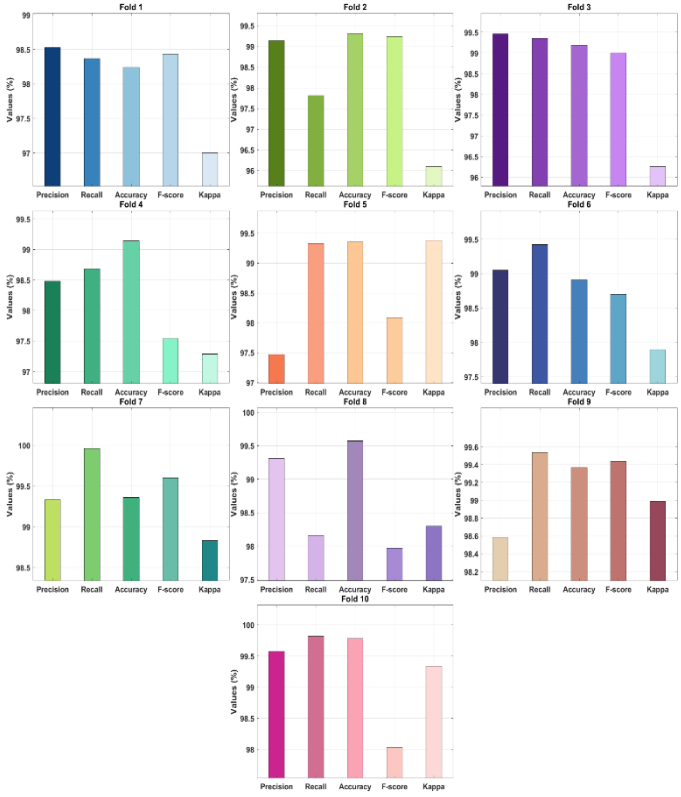


Fig. 4. Result Analysis of DLTCR-PHWC Model under Different Folds.

TABLE I. RESULT ANALYSIS OF PROPOSED MODEL IN TERMS OF DIFFERENT MEASURES

No. of Folds	Precision	Recall	Accuracy	F-score	Kappa
Fold 1	98.53	98.37	98.24	98.43	97.00
Fold 2	99.14	97.81	99.30	99.24	96.10
Fold 3	99.46	99.35	99.19	99.00	96.26
Fold 4	98.48	98.68	99.14	97.54	97.29
Fold 5	97.47	99.33	99.35	98.09	99.38
Fold 6	99.05	99.42	98.91	98.70	97.89
Fold 7	99.33	99.96	99.36	99.60	98.83
Fold 8	99.31	98.16	99.57	97.97	98.30
Fold 9	98.59	99.54	99.37	99.44	98.99
Fold 10	99.58	99.82	99.78	98.03	99.34
Average	98.89	99.04	99.22	98.60	97.94

TABLE II. RESULT ANALYSIS OF EXISTING WITH PROPOSED METHOD IN TERMS OF DIFFERENT MEASURES

Methods	Precision	Recall	F-score	Accuracy
DLTCR-PHWC	0.9889	0.9904	0.9860	0.9922
MLP-HMM	0.8467	0.8689	0.8512	0.8500
CNN	0.9567	0.9582	0.9572	0.9632
KNN	0.9415	0.9543	0.9512	0.9585
MLP	0.9572	0.9590	0.9578	0.9645
NN	0.9634	0.9685	0.9680	0.9750
DNN	0.9089	0.9144	0.9132	0.9210
CNN-RF	-	-	-	0.7140
CNN-MLP	-	-	-	0.7750
CNN-KNN	-	-	-	0.8160

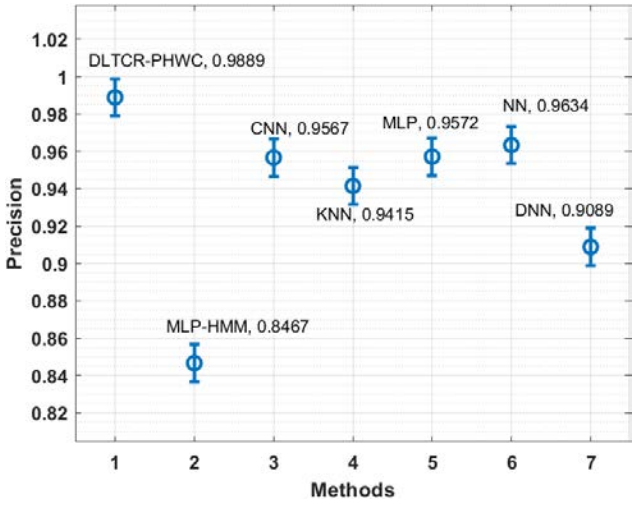


Fig. 5. Comparative Analysis of DLTCR-PHWC Model with respect to Precision.

Fig. 6 offers a detailed recall analysis of the DLTCR-PHWC algorithm with recent algorithms. The figure outperformed that the MLP-HMM and DNN manners have attained least outcome with the minimum recall of 0.8689 and 0.9144 correspondingly. Besides, the CNN, KNN, MLP, and NN approaches have gained moderately closer recall of 0.9582, 0.9543, 0.9590, and 0.9685 correspondingly. Finally, the DLTCR-PHWC methodology has resulted in an increased recall of 0.9904.

Fig. 7 showcases a brief F-score analysis of the DLTCR-PHWC manner with existing algorithms. The figure described that the MLP-HMM and DNN methods have reached minimum results with the reduced F-score of 0.8512 and 0.9132 correspondingly. Followed by, the CNN, KNN, MLP, and NN manners have achieved moderately closer F-score of 0.9572, 0.9512, 0.9578, and 0.9680 correspondingly. Eventually, the DLTCR-PHWC method has resulted in a superior F-score of 0.9860.

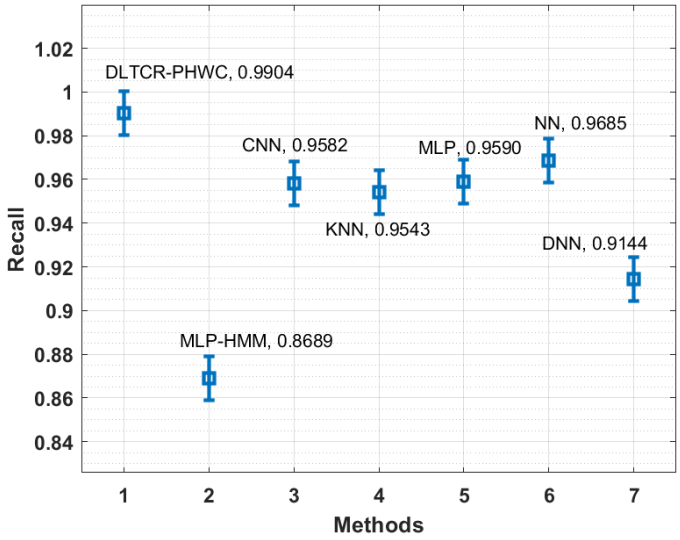


Fig. 6. Comparative Analysis of DLTCR-PHWC Model with respect to Recall.

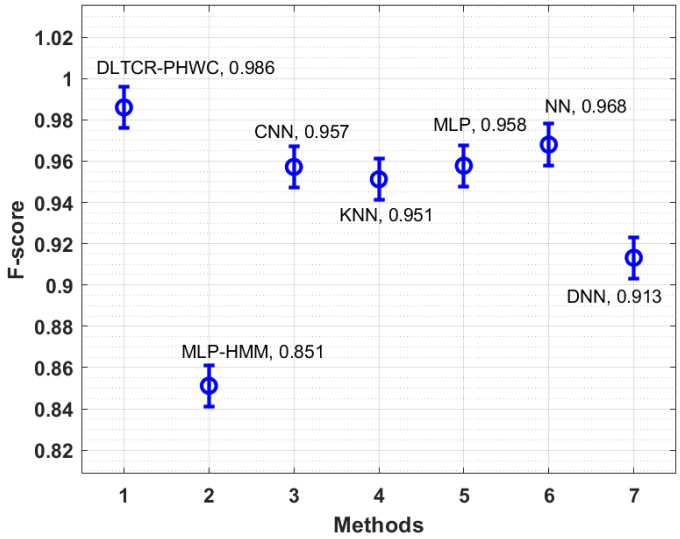


Fig. 7. Comparative Analysis of DLTCR-PHWC Model with respect to F-Score.

Fig. 8 demonstrates a brief accuracy analysis of the DLTCR-PHWC method with existing approaches. The figure stated that the CNN-RF, CNN-MLP, and CNN-KNN algorithms have gained worst outcome with the minimum accuracy of 0.7140, 0.7750, and 0.8160 respectively. At the same time, MLP-HMM and DNN techniques demonstrated a somewhat higher accuracy of 0.8500 and 0.9210 correspondingly. Along with that, the CNN, KNN, MLP, and NN techniques have obtained moderately closer accuracy of 0.9632, 0.9585, 0.9645, and 0.9750 respectively. At last, the DLTCR-PHWC methodology has resulted in a maximal accuracy of 0.9922.

After examining the above results and discussion, it is apparent that the DLTCR-PHWC technique has been found to be a proficient tool for effective TCR.

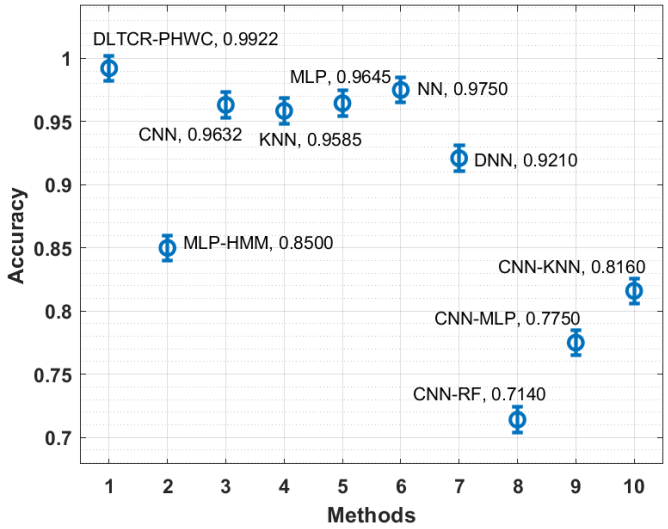


Fig. 8. Comparative Analysis of DLTCR-PHWC Model with respect to Accuracy.

V. Conclusion

In this study, a new DLTCR-PHWC technique is derived to detect and recognize the printed as well as handwritten characters that exist in the same image. The DLTCR-PHWC technique encompasses different operations such as AFF based preprocessing, line/paragraph segmentation, fusion based feature extraction, BiLSTM based recognition, and AO based hyperparameter optimization. The design of AO technique to fine tune the hyperparameters involved in the BiSLTM model helps to accomplish enhanced TCR outcomes. In order to showcase the supremacy of the DLTCR-PHWC technique, a wide range of simulations were performed against Telugu character dataset and the results have ensured the betterment of the DLTCR-PHWC technique. Therefore, the DLTCR-PHWC technique finds it useful to recognize both the handwritten and Telugu characters present in same image. In future, advanced DL models can be utilized for recognition process with hybrid metaheuristic-based parameter tuning process.

References

[1] Lin D, Lin F, Lv Y, Cai F, Cao D (2017) Chinese character CAPTCHA recognition and performance estimation via deep neural network. Neuro Computing 17:1–40.

[2] Guruprasad P, Majumdar J (2016) Multimodal recognition framework: an accurate and powerful Nandinagari handwritten character recognition Modelb. Procedia Comput Sci 89:836–844.

[3] Sampath AK, Gomathi N (2017) Fuzzy-based multi-kernel spherical support vector machine for effective handwritten character recognition, Research Article 10:1-13.

[4] Naz S, Hayat K, Razzak MI, Anwar MW, Madani SA, Khan SU (2013) The optical character recognition of Urdu-like cursive scripts, Research Article, 31:1229-1249.

[5] Chacko BP, Vimal Krishnan VR, Raju G, Babu Anto P (2012) Handwritten character recognition using wavelet energy and extreme learning machine. Int J Mach Learn Cybern 3:149–161.

[6] Sampath AK, Gomathi N (2017) Decision tree and deep learning based probabilistic model for character recognition. J Cent S Univ 24:2862–2876.

[7] Ajantha Devi V, Santhosh Baboo S (2014) Embedded optical character recognition on Tamil text image using raspberry pi. International Journal of Computer Science Trends and Technology (IJCST) 2(4):127– 132.

[8] Prameela, N., Anjusha, P. and Karthik, R., (2017), April. Off-line Telugu handwritten characters recognition using optical character recognition. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 2, pp. 223-226). IEEE.

[9] Cheekati, B.M. and Rajeti, R.S., (2020), October. Telugu handwritten character recognition using deep residual learning. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 788-796). IEEE.

[10] Lakshmi, K.M. and Babu, T.R., (2018), January. A Novel Telugu Script Recognition and Retrieval Approach Based on Hash Coded Hamming. In International Conference on Communications and Cyber Physical Engineering 2018 (pp. 571-582). Springer, Singapore.

[11] Burra, S., Patel, A., Bhagvati, C. and Negi, A., (2017), December. Improved symbol segmentation for telugu optical character recognition. In International Conference on Intelligent Systems Design and Applications (pp. 496-507). Springer, Cham.

[12] Hebbi, C., Mamatha, H.R., Sahana, Y.S., Dhage, S. and Somayaji, S., (2020). A convolution neural networks-based character and word recognition system for similar script languages Kannada and Telugu. In Proceedings of ICETIT 2019 (pp. 306-317). Springer, Cham.

[13] Rani, N.S., Vasudev, T. and Pradeep, C.H., (2017). An Enhanced Template Matching Technique for Recognition of Telugu Script. International Journal of Signal Processing, 2.

[14] Sarika, N. and Sirisala, N., (2021). Deep Learning Techniques for Optical Character Recognition. In Sustainable Communication Networks and Application (pp. 339-349). Springer, Singapore.

[15] Madhavi, G.B., Kumar, V. and Vakula, V.K., (2021). An Effective Slant Detection and Correction Method Based on the Tilted Rectangle Method for Telugu Manuscript Terms. International Journal of Information Technology Project Management (IJITPM), 12(4), pp.25-37.

[16] BJ, B.N., Athira, M.R. and Prajwal, M.L., (2021), May. Kannada Confusing Character Recognition and Classification Using Random Forest and SVM. In 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 537-541). IEEE.

[17] Kowsalya, S. and Periasamy, P.S., (2019). Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization. *Multimedia Tools and Applications*, *78*(17), pp.25043-25061.

[18] Mekhalfi, M.L., Bejiga, M.B., Soresina, D., Melgani, F. and Demir, B., (2019). Capsule networks for object detection in UAV imagery. *Remote Sensing*, *11*(14), p.1694.

[19] Wu, L., Ma, J., Zhao, Y. and Liu, H., (2021). Apple Detection in Complex Scene Using the Improved YOLOv4 Model. *Agronomy*, *11*(3), p.476.

[20] Ji, Z., Wang, X., Cai, C. and Sun, H., (2020). Power entity recognition based on bidirectional long short-term memory and conditional random fields. *Global Energy Interconnection*, *3*(2), pp.186-192.

[21] Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A.A., Al-qaness, M.A. and Gandomi, A.H., (2021). Aquila Optimizer: A novel meta-heuristic optimization Algorithm. *Computers & Industrial Engineering*, *157*, p.107250.

[22] Kummari, R. and Bhagvati, C., (2018), December. UHTelPCC: A Dataset for Telugu Printed Character Recognition. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 24-36). Springer, Singapore.

[23] Lakshmi, C.V., Jain, R. and Patvardhan, C., (2006). OCR of printed Telugu text with high recognition accuracies. In *Computer Vision, Graphics and Image Processing* (pp. 786-795). Springer, Berlin, Heidelberg.

[24] Pujari, A.K., Naidu, C.D., Rao, M.S. and Jinaga, B.C., (2004). An intelligent character recognizer for Telugu scripts using multiresolution analysis and associative memory. *Image and Vision Computing*, *22*(14), pp.1221-1227.

[25] Ganji, T., Velpuru, M.S. and Dugyala, R., (2021). Multi Variant Handwritten Telugu Character Recognition Using Transfer Learning. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1042, No. 1, p. 012026). IOP Publishing.

[26] Sujatha, P. and Lalitha Bhaskari, D., (2019). Telugu and Hindi Script recognition using deep learning techniques. *Int. J. Innov. Technol. Explor. Eng*, *8*, pp.2278-3075.

Industrial Revolution 5.0 and the Role of Cutting Edge Technologies

Mamoona Humayun
Department of Information Systems
College of Computer and Information Sciences, Jouf University
Al-Jouf, Saudi Arabia

***Abstract*—IR 4.0 emphasizes the interconnection of machines and systems to achieve optimal performance and productivity gains. IR 5.0 is said to take it a step further by fine-tuning the human-machine connection. IR 5.0 is more collaboration between the two: automated technology's ultra-fast accuracy combines with a human's intelligence and creativity. The driving force behind IR 5.0 is customer demand for customization and personalization, necessitating a greater human involvement in the production process. As IR 5.0 evolves, we may expect to see a slew of breakthroughs across various industries. However, just automating jobs or digitizing processes will not be enough; the finest and most successful businesses will be those that can combine the dual powers of technology and human ingenuity. IR 5.0 focuses on the use of modern cutting-edge technologies, namely, AI, IoT, big data, cloud computing, Blockchain, Digital twins, edge computing, collaborative robots, and 6G along with leveraging human creativity and intelligence. Wherever possible, IR 5.0 will change industrial processes worldwide by removing mundane, filthy, and repetitive activities from human workers. Intelligent robotics and systems will have unparalleled access to industrial supply networks and production floors. However, to understand and leverage the benefits of IR 5.0 better, there is a need to understand the role of modern CET in industrial revolution 5.0. To fill this gap, this article will examine IR 5.0 prospective, uses, supporting technologies, opportunities, and issues involved that need to be understood for leveraging the potentials of IR 5.0.**

Keywords*—*Industry 5.0; cutting-edge technologies; Internet of Things; Aartificial intelligence; big data

I. Introduction

From the industrial revolution to the digital transformation and beyond [1-3], the history of the IR traces the progress of the manufacturing industry. Each stage indicates a shift in the production process that has altered how we think about and operate in the business. When the major method of manufacturing shifted from people to machine power, the first IR began. The second IR, known as the Technological Revolution or IR 2.0, continued where the first left off, with improved electrical technology allowing for even more manufacturing and more complex machinery. The third IR, known as the digital revolution or IR 3.0, began with the invention of the first computers, which created the foundation for a society that is difficult to envision today without computer technology. We are currently in IR 4.0, which is defined by the amount of automation that has been accomplished. Machines can frequently govern themselves in many ways utilizing internet technology or the IoT. IR 4.0 also emphasizes the utilization of cloud technologies and the relevance of BGD. Although IR 4.0 has greatly improved the manufacturing industry by enabling security, scalability, control and visibility, customer happiness, and customization, the human drive for creativity still exists, and researchers are already predicting the next revolution that is IR 5.0 [4-6]. In short, if the current revolution focuses on converting factories into IoT-enabled smart facilities, IR 5.0 is expected to place a greater emphasis on the return of human skills and intelligence to the industrial framework. Fig. 1 illustrates all the IRs so far diagrammatically.

According to [7-9]; in IR 5.0, man and machine work together to optimize manufacturing process performance. Surprisingly, the fifth IR may already be underway among firms that are only starting to use IR 4.0 concepts. The reason for this is that when businesses adopt contemporary technology, they do not suddenly eliminate huge segments of their employees and shift to a fully automated operation. Thus, by delegating repetitive and boring activities to robots/machines and critical thinking tasks to people, IR 5.0 improves manufacturing quality. IR 5.0 supports more skilled occupations than IR 4.0 because experts handle the machines. Further, it focuses on improving consumer happiness by forming a collaborative relationship between humans and robots. Another advantage of IR 5.0 as mentioned by [10, 11] is that it provides greener solutions than traditional industrial transformations, which do not prioritize environmental protection.

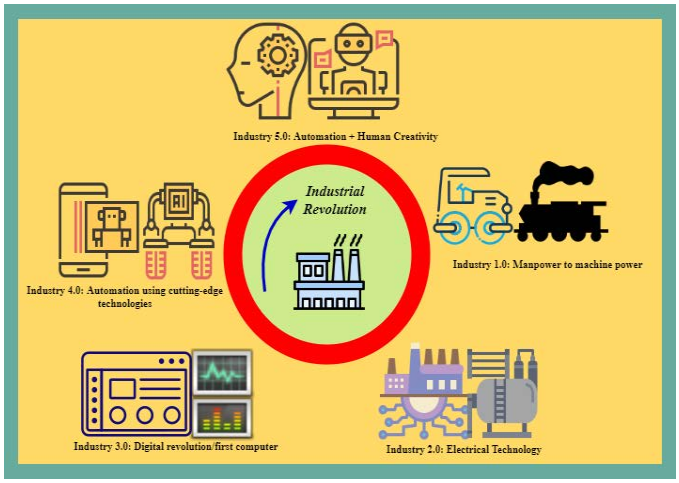


Fig. 1. Industrial Revolution.

Several current studies examine the enabling technologies (ETs), applications, and problems of prior industrial standards and supporting industrial technologies such as articles [10-15]. The authors of [14], for example, investigate the ETs of the IR 4.0 standard as well as the rationale for their inclusion in the standard. Simultaneously, [15] examines the ETs in the Web of Science database for IR 4.0, as well as the contributions of several forerunners in the development of this database. The authors of [16] have published a comprehensive review of the ETs and uses of virtual reality in IoT systems.

Industries can play an important role in addressing societal challenges such as resource preservation, climate change, and social stability. Businesses, employees, and society all benefit from IR 5.0. In addition to giving workers more control, it meets the ever-changing needs of employees in terms of skills and training. Competitiveness is boosted, and top talent can be attracted more easily. It is beneficial to our world since it encourages circular production models and supports technologies that increase the efficiency of natural resource utilization. Even though IR 5.0 is becoming more popular, we are unaware of any study that analyzes the role of modern CET in the revolution of IR 5.0. As a result of this observation, we intend to present the role of modern CET in the IR in the context of IR 5.0. In a nutshell, our work's contributions may be stated as follows:

1) Define IR 5.0 to have a comprehensive knowledge of the concept from several angles.
2) Comparing and contrasting IR 5.0 to earlier IR.
3) Discuss the most promising IR 5.0 applications
4) Discuss the role of modern CET in IR 5.0
5) Issues and challenges involved in IR 5.0

TABLE I. LIST OF ABBREVIATIONS USED

Abbreviations	**Used for**
CET	Cutting Edge Technologies
IoT	Internet of Things
AI	Artificial intelligence
BGD	Big Data
DT	Digital Twins
IR	Industrial Revolution
SCM	Supply Chain Management
BCT	Blockchain Technology
CPPS	Cyber Physical Production Systems
CAGR	Compound Annual Growth Rate
ITU	International Telecommunication Union
ZB	Zettabytes
CC	Cloud Computing
COBOTS	Collaboration Robots
ETs	Enabling Technologies
EC	Edge Computing
SME	Small and Medium size Organizations

The remaining paper is structured as: Section 2 will define IR 5.0 from various perspectives. Section 3 will summarize existing knowledge to compare and contrast IR 5.0 to the earlier IR. Section 4 will discuss the key applications of IR 5.0. Section 5 will highlight the role of modern CET in the fifth IR. Section 6 will discuss the associated issues and challenges. Section 7 will discuss the findings of the study. Finally, Section 7 will wrap up the paper by providing the conclusion of the paper and insights into future work.

Table I shows the list of abbreviations used in this research for better understanding.

II. LITERATURE REVIEW

This section will provide a general overview of IR 5.0 to provide a better understanding of the phenomenon. Further, it will also discuss the key pillars of IR 5.0.

A. IR 5.0 and State-of-the-Art

Since IR 5.0 is still evolving, several researchers and practitioners have presented varied definitions. Here are some of the definitions that are considered.

Definition1: By integrating work processes and intelligent systems, IR 5.0 brings the human workforce back to the factory, where humans and machines are partnered to boost process efficiency by harnessing human intelligence and creativity [4].

Definition2: IR 5.0 blends the inherent strengths of human intelligence and CPPS to build synergetic factories. In addition, authorities are searching for creative, ethical, and human-centered design to overcome the personnel shortages caused by IR 4.0 [17].

Definition3: IR 5.0 is a paradigm for the next phase of industrialization, which incorporates the return of labor to factories, distributed production, intelligent SCM, and hyper customization, all of which work together to create a tailored customer experience over time [18].

B. Pillars of IR 5.0

IR 5.0 focuses on stakeholder value rather than shareholder value, reinforcing the industry's role and contribution to society. Below are the key pillars of IR 5.0 (as shown in Fig. 2).

1) Human-centric: Human ingenuity and craftsmanship are combined with the speed, efficiency, and consistency of robots in IR 5.0. Thus it promotes human empowerment, talent, and diversity [4].

2) Sustainability: Additive manufacturing, often known as 3D printing, is one of the most notable elements of IR 5.0, and it is used to make manufacturing items more sustainable. In IR 5.0, additive manufacturing aimed to improve customer happiness by incorporating benefits into goods and services [19].

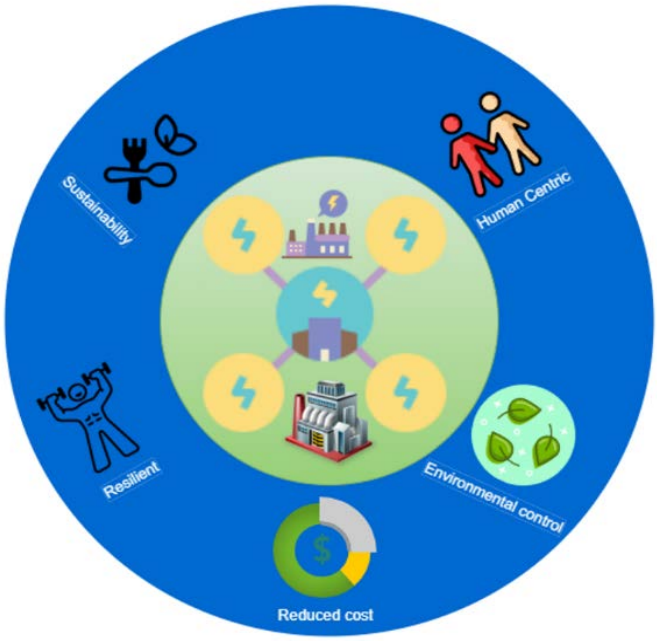


Fig. 2. Pillars of IR 5.0.

3) Resilient: The term "resilience" refers to the necessity to improve industrial production's robustness, equipping it better against interruptions and ensuring that it can offer and sustain key infrastructure during the crisis. High resilience can be attained when humans and robots operate together [20, 21].

4) Reduced cost and environmental control: Climate, humidity, temperature, and energy usage are all monitored in real-time and predicted using smart, networked sensors and specialized algorithms. This is especially beneficial in farms that are highly dependent on the weather. Knowing what to expect and where to act might help to avoid costly mistakes and boost the output [22].

C. Opportunities of IR 5.0

The prospects presented by IR 5.0 are as follows, these opportunities are discussed in the articles [8,23-26].

1) Better employment: Through the implementation of next-generation technologies, greater automation will have a favorable influence on employment in several areas.

2) Customization: Customers have more customization options with highly automated production methods.

3) Improved human efficiency: IR 5.0 opens up more options for creative individuals to come and work, allowing for improved human efficiency.

4) Employee safety: Employee safety on the work floor has improved since COBOTs can do hazardous tasks.

5) Customer satisfaction: More personalized products and services boost consumer happiness, loyalty, and attract new customers, resulting in higher profits and market share for businesses.

6) Better opportunities: It gives start-ups and entrepreneurs in creative and inventive fields, enormous opportunities to develop new goods and services related to IR 5.0, as long as enough money and infrastructure are available.

7) Human-machine interaction: IR 5.0 places a greater emphasis on human-machine interaction and gives a bigger platform for research and development in this industry.

8) Quality Service: With the support of IR 5.0, quality services may be offered in faraway regions, particularly in the healthcare business, such as surgical procedures performed by robots in rural areas.

9) Frequent follow-up: IR 5.0 will assist the customer digitally in handling frequent follow-up assignments by making machines adaptable according to employee demands.

10) Higher-value job: Because individuals are given the freedom to be responsible for building again, IR 5.0 provides higher-value jobs than before.

11) Better planning: In IR 5.0, the production cell operator is more involved in the planning approach than the more or less automated manufacturing method.

12) Creative freedom: It enables more custom-made and personalized items, as well as creative freedom.

13) Automation: With IR 5.0, it is easier to automate production processes.

In comparison to IR 4.0, the preceding discussion demonstrates that IR 5.0 is the next industrial revolution, in which human specialists and efficient, intelligent, and precise machines will work together to develop efficient and user-friendly manufacturing solutions. It will enhance the sector by bringing in human talents and competencies, as well as opening up new opportunities. We will further explore the strengths of IR 5.0 and the role of modern CET in IR 5.0 in the upcoming sections.

III. Comparison of IR 4.0 and 5.0

To better understand and visualize the benefits of IR 5.0, it's important to comprehend and see the differences between IR 4.0 and 5.0. This section will compare both of these IRs.

Today, we live in IR 4.0, which is rapidly evolving towards IR 5.0. The fourth IR, often known as IR 4.0, was constructed on top of the third to allow improved technology. Everything became "Smarter" during this period. IoT, CC, CPPS, and cognitive computing were among the most essential technologies discussed. IR 4.0 connects systems, components, and humans over a network as discussed in [27-30], making the production process more efficient and automated. Humans collaborating with these technologies to improve efficiency is the goal of the fifth Revolution, often known as IR 5.0. For optimal advantages, it is vital to strategize techniques of human-robot integration since it will seek to match the rising demand for individuals with unique customization and modifications. Improved human engagement with intelligent machines will lead to increased efficiency. Not to mention that, as a result of this shift, more high-paying positions will be created. Table II summarizes the comparison between IR 4.0 and 5.0 so that organizations could leverage the benefits of IR 5.0.

TABLE II. COMPARISON BETWEEN IR 4.0 AND 5.0

Industry 4.0	**Industry 5.0**
The goal is to automate processes.	The goal is to strike a balance between machine and human engagement.
The most crucial factor was technology	The most crucial collaboration is between people and robots.
The entire environment is virtual.	The shift back to the real world.
As new smart technologies were adopted, the number of personnel was reduced.	An increase in the number of people who come into contact with machines.
Machines that are smarter and more linked to the workplace.	Cognitive computers and human intelligence are being combined.
There is no way to personalize or customize the product.	Personalization and customization are available, allowing each product to be improved and tailored to the needs of the person.
It's still tossing back and forth between renewable and nonrenewable energy sources.	It is more environmentally friendly since renewable energy sources will be used more often.

IV. IR 5.0 APPLICATIONS

IR 5.0 is already in use in various areas, including healthcare, manufacturing, CPPS, SCM, education, disaster management, and so on, as mentioned in studies [23, 26, 31]. In addition to merging many CET with machines, such as AI, edge computing (EC), IoT, DT, COBOTS, 6G and beyond, BGD analytics, and so on, the intellect of people is also utilized when making judgments in IR 5.0. As a result, the personal human touch is added to the IR 4.0 pillars of efficiency and automation. Below we provide some core applications of IR 5.0.

A. IR 5.0 in Healthcare

IR 5.0 has the potential to revolutionize the healthcare industry [33], allowing for the production of individualized gadgets, implants, and other medical products. Routine occupations, such as routine checkups conducted by doctors, can be handled by COBOTS under IR 5.0. Similarly, human-robot collaboration can enhance the diagnosis and treatment process. Intelligent wearable gadgets, such as smartwatches and intelligent sensors, can, for example, continuously capture a patient's healthcare data in real-time and store it in the cloud [32]. The patients' medical condition may then be diagnosed using machine learning methods. These intelligent gadgets can interact with one another, and if a doctor's attention is necessary, these devices can feed the present state of the patient to the physicians and notify them to treat the patient. Doctors can use COBOTS to get assistance from robots that can interact with one another to perform surgery on patients.

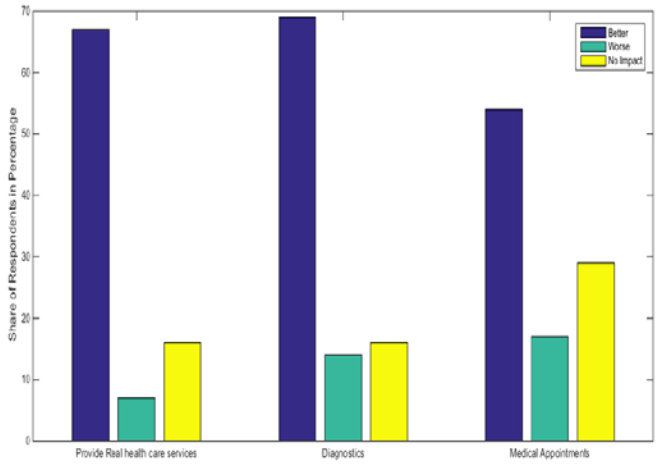


Fig. 3. Canadians' Predictions on the Influence of AI in Healthcare over the Next Ten Years, as of 2019[34].

In a poll conducted in Canada to assess the role of AI and robots in healthcare, approximately 70% of respondents said that AI programs or robots delivering genuine health care services, such as robot-aided surgery, early diagnosis, and so on, will make life better. As of August 2019, the data in Fig. 3 depicts Canadians' projections for AI-related health developments in the next ten years and their influence on life. Statista was used to compile the data for this research.

B. IR 5.0 in Manufacturing/Production

According to the co-founder of universal robots chief technology [35], IR 5.0 will transform the factory into a place where creative individuals can come and work, resulting in a more customized and human experience for both workers and consumers. The fifth IR will see substantially more complicated collaborative interactions between intellectuals, machinery, processes, and overall system for optimal performance optimization.

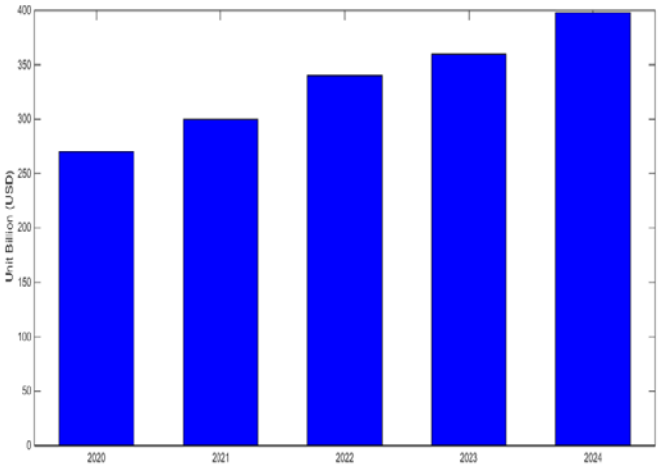


Fig. 4. Revenue of Global Manufacturing Industry.

According to the TrendForce investigation [36], as shown in Fig. 4, the manufacturing industry's income is predicted to improve in the near future owing to human and robot collaboration. This shows that the fifth IR might lead to paradigm shifts and profound changes in how we think about industry and production.

C. IR 5.0 and Supply Chain Management

IR 5.0 is expected to optimize the SCM by combining smart, linked digital environments with the human intelligence required to maximize their value. According to KBV research statistics [37], as shown in Fig. 5, the global SCM software market is estimated to reach $22.7 billion by 2024, growing at a 12.1 percent CAGR. SCM Software is a real-time analytical tool for managing the movement of products and other types of data throughout the SCM network. The software improves an organization's supply chain operations.

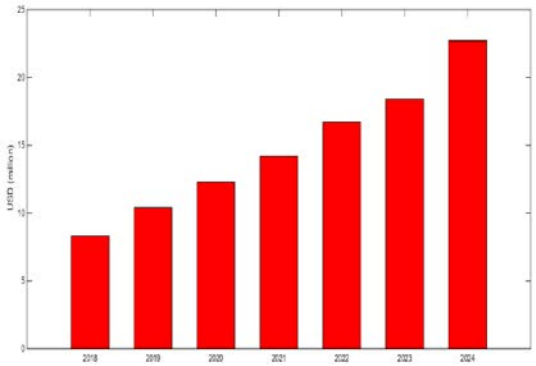


Fig. 5. Global SCM Software Market.

IR 5.0 will improve SCM in the following way [38]

- Increasing supply chain customization improves not just customer happiness but also efficiency and profits.
- Using more up-to-date data, reducing supply chain risks and waste.
- allowing supply chain and logistics units to spend more time on strategic innovation rather than putting out fires or dealing with fundamental execution issues.
- Improving supply chain integration to form more strategic alliances.
- Increasing the value of an organization's human capital by assisting in retaining and transferring knowledge about the features of a certain supply chain.

D. IR 5.0 and Education

IR 5.0 will also transform education [39, 40], the current education is fueled by information, and if it can be trained and equipped with digitally smart machines, or COBOTS, that are further supplemented with the human touch, it will take society along the path of personalized education. As a result, COBOTS will help to develop a human-centric society, which will be strengthened by human wisdom, enabling education 5.0, tailored education for everybody. Humans will be aided by many functioning COBOTS to assist them with day-to-day tasks and support their personal and professional development.

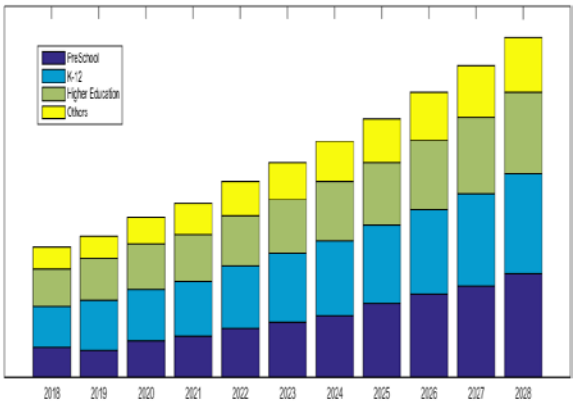


Fig. 6. Education Technology Market (Asia Pacific).

The worldwide education technology industry is expected to be worth USD 89.49 billion in 2020 [41], with a CAGR of 19.9% predicted between 2021 and 2028 as shown in Fig. 6. This shows that a blend of digital technology and human intelligence has the potential to expand educational opportunities.

V. Role of CET in the Revolution of IR 5.0

Edge computing, IoT, DT, Blockchain, BGD analytics, COBOTS, and 6G are just a few of the CET that, when combined with cognitive abilities and creativity, may help enterprises enhance output and offer customized goods more rapidly. IR 5.0 is an improved manufacturing model that focuses on the collaboration between humans and robots, and this collaboration makes human talents more productive and easier to automate for individuals and enterprises than they have ever been. According to the statistics provided by Statista in [42], as of 2020, new CET are projected to have the largest influence on businesses all around the world as shown in Fig. 7. The data of Fig. 7 shows that IoT is widely considered as one of the most crucial areas of current and future technology, the second CET that highly impacted organizations worldwide is AI robots, nearly used in every field to increase efficiency and complement our human skills [22]. Fig. 7 depicts the other CET that have had a significant influence on the global organization. Now we'll talk about how contemporary CET will play a part in the fifth IR.

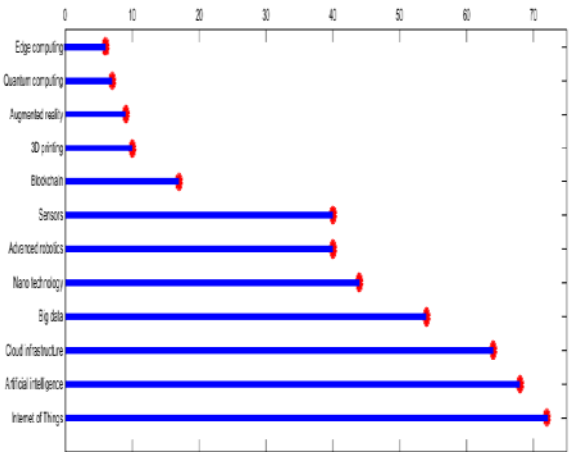


Fig. 7. CET Impact on Organizations Globally as of 2020.

A. IoT and IR 5.0

Humans define the strategy, give supervision, and contribute creative input in IR 5.0, while technology does the dull, repetitive, and error-prone activities. Businesses will benefit from this new division of labor not only in terms of cost savings but also in terms of tapping into new value streams provided by the human touch. The IoT is defined by two characteristics: automation and connection [43]. Given these characteristics, the IoT will need to leverage a variety of technologies to guarantee that data is sent, analyzed, and responded automatically across different devices. The IoT has changed the modern world and is a primary driver of the fifth IR [44]. The following are some of the most important advantages of IoT in terms of industrial advancement [45].

- Increased employee productivity and decreased human labor
- Effective management of operations
- Optimum utilization of resources and assets
- Cost-effective operation
- Workplace safety has improved.
- Marketing and business growth that is thorough
- Customer retention and service have improved.
- Better business prospects
- The company's image will be more trustworthy.

Because of IoT's widespread use and benefits, the number of IoT-connected devices is fast growing and is projected to continue to grow in the near future, as shown in Fig. 8. The data of Fig. 8 is taken from [46].

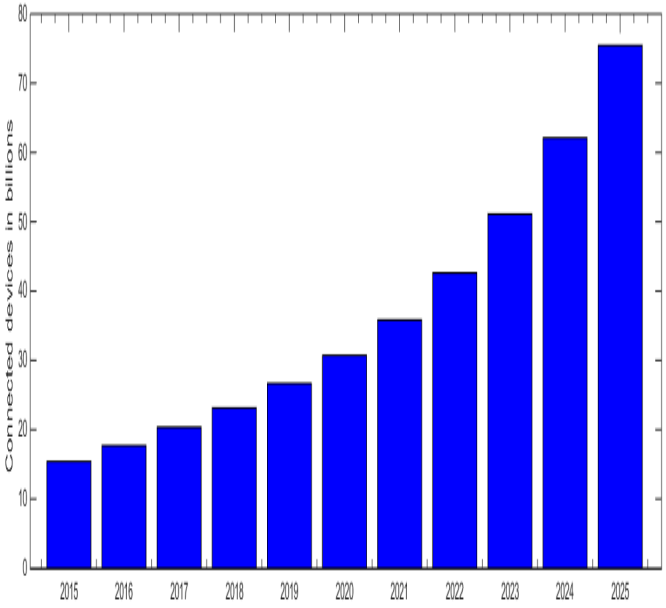


Fig. 8. Expected IoT Connected Devices Worldwide.

B. Cloud Computing and IR 5.0

CC is a concept that allows for the instantaneous leasing of computer resources with little or no communication with the provider. Cloud simplifies operations in this sense since it eliminates the need for rigorous resource dimensioning and planning, allowing for flexible usage without the user's previous commitment [47, 48]. Cloud users benefit from almost all the resources they need, and they may either utilize or supply everything as a service.

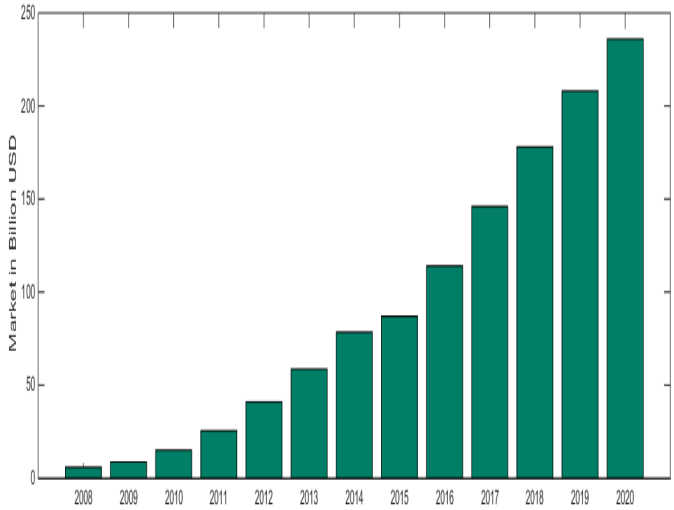


Fig. 9. Global Public Cloud Market Size Worldwide.

Cloud, in particular, is required to reap the benefits of IR 5.0 since it provides user mobility, distributed data analytics, resources' heterogeneity, and meets the needs of diverse applications with low latency. It also makes managing and developing computation, storage, and networking services easier between data centers and end devices. It's also a strong tool for processing BGD generated through IoT sensors, allowing IR 5.0 to realize its full potential [20, 47, 49]. Due to these potential benefits, the cloud market has grown at a tremendous pace as shown in Fig. 9. The statistics of Fig. 9 are taken from [50].

C. BGD and IR 5.0

In the realm of IR 5.0, BGD Analytics is expected to play a significant role. Some firms in IR 5.0 can utilize BGD Analytics to better understand customer behavior to optimize product pricing, improve manufacturing efficiency, and lower overhead expenses [51]. IR 5.0 apps can leverage BGD Analytics to make real-time choices to improve their competitive edge, with an emphasis on offering suggestions on predictive findings for significant events and customization [52]. Manufacturers can create and handle large amounts of data with the support of real-time analytical data. Another important problem in IR 5.0 is continuous process improvement, which frequently necessitates the collection of extensive data on the whole production cycle. To enhance predictability and explore new possibilities, BGD analytics approaches are utilized to identify and eliminate non-essentials. According to [53], the revenue from the BGD industry is growing and is likely to continue to grow in the future as shown in Fig. 10.

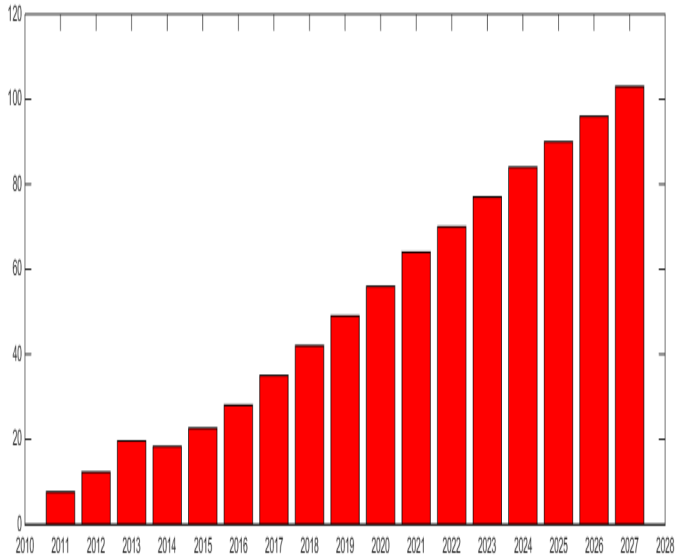


Fig. 10. Revenue Prediction for the Global BGD Market from 2011 to 2027.

D. Digital Twins and IR 5.0

DT can provide significant value to the creation of customized items on the market, improve business operations, fewer faults, and fast increase creative business models to make profits in IR 5.0. The DT may help IR 5.0 address technical challenges by finding them earlier, identifying configurable components, generating more accurate projections, forecasting future failures, and preventing enormous financial losses [54]. This form of smart architectural design enables businesses to get economic benefits more rapidly and sequentially than ever before. DT can be used for the access of real-time data and creating simulation models in IR 5.0, allowing organizations to edit and update physical things remotely. DT is also used in IR 5.0 for customization that allows clients to build virtual environments to realize the findings [55].

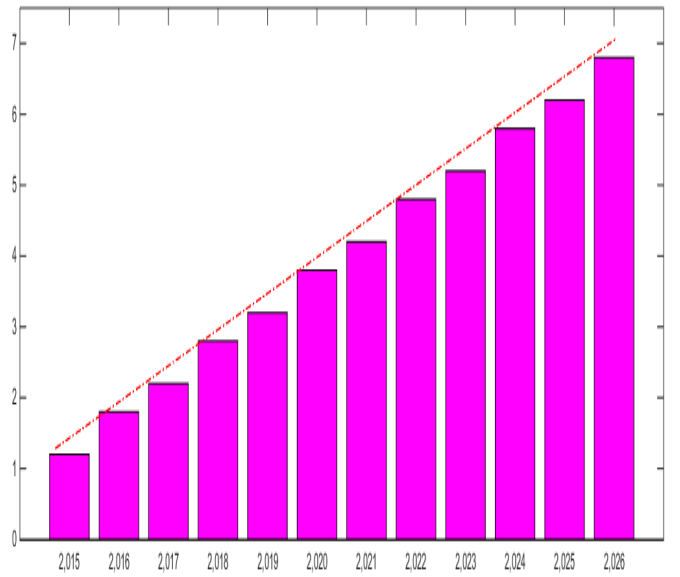


Fig. 11. DT Market 2015-2026.

As illustrated in Fig. 11 based on the statistics collected from [56], the global DT market is expected to reach USD 36.6 billion by 2025, rising at a CAGR of 38.5 percent from 2020 to 2026. Increased acceptance of new technologies such as IoT and cloud, increased demand for connected devices in the automotive and transportation industries, and expanding predictive maintenance usage are all driving market expansion.

E. COBOTS and IR 5.0

Robots are far superior to humans in the creation of high-volume items and are far more compatible. Robots are inefficient in critical thinking when compared to humans. When robots need to be guided, customizing or personalizing things may be a huge difficulty. As a result, managing human connections within manufacturing processes is critical. COBOTS have a lot of potential in IR 5.0. Robots can achieve their intended aim by collaborating with humans, allowing for the rapid and accurate delivery of mass customized and personalized items to clients [57]. Throughout IR 5.0, personalizing COBOTS may take various forms, including medical treatments. COBOTS aid with the improvement of safety and performance in IR 5.0 applications, while also providing more engaging tasks for human workers and enhancing product development. In highly competitive marketplaces, industries must recognize that COBOTS have the potential to increase corporate performance and minimize rising labor expenses [58]. According to the interact analysis report, material handling, assembly, and pick-and-place are expected to be the three most common uses of collaborative robots in the future [59]. In 2024, these three tasks will account for 62.7 percent of collaborative robot revenues, up from 71.9 percent in 2019 as depicted in Fig. 12. Pick and place is expected to remain the third most popular use for collaborative robots by 2024, according to Interact Analysis, but note the bar on the far right of each graph that denotes "other applications." Innovative COBOTS technologies are positioned to enter a variety of new products and non-manufacturing contexts settings for which industrial robots are unsuitable.

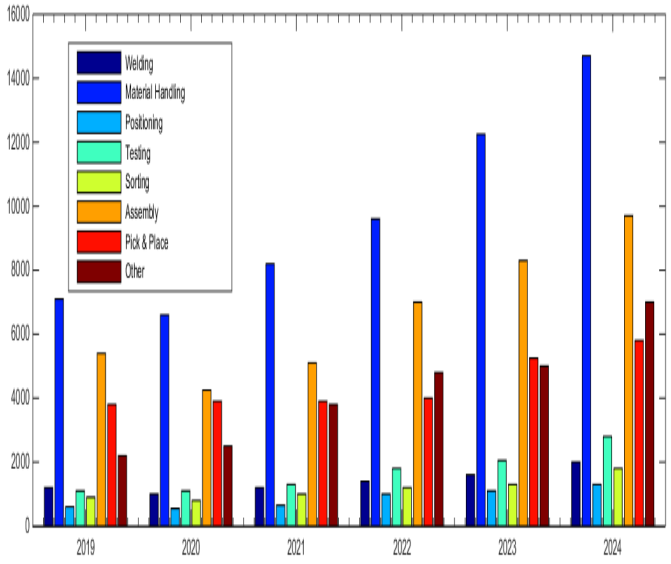


Fig. 12. Prediction of COBOTS Shipment by Application.

F. 6G & Beyond and IR 5.0

Due to the rapid rise of smart infrastructure and prospective applications, it will be impossible to quickly expand bandwidth requirements with present networks. The usage of 6G and beyond in the IR 5.0 revolution allows for lower latency, high-quality services, and vast IoT infrastructure, as well as integrated AI capabilities. 6G networks aid in the efficient and successful execution of IR 5.0 applications by enabling smart spectrum management and smart mobility. 6G networks are intended to satisfy the needs of future society by interconnecting the overall society [60]. IR 5.0 apps have a major challenge in terms of energy management due to a large number of connected smart devices and the significant quantity of energy they use. The employment of advanced energy efficiency algorithms and energy harvesting technology will make 6G networks more energy-efficient [61]. According to the ITU, the exponential growth trend will continue, and global mobile data traffic will reach a staggering 5ZB per month by 2030 [60], as seen in Fig. 13.

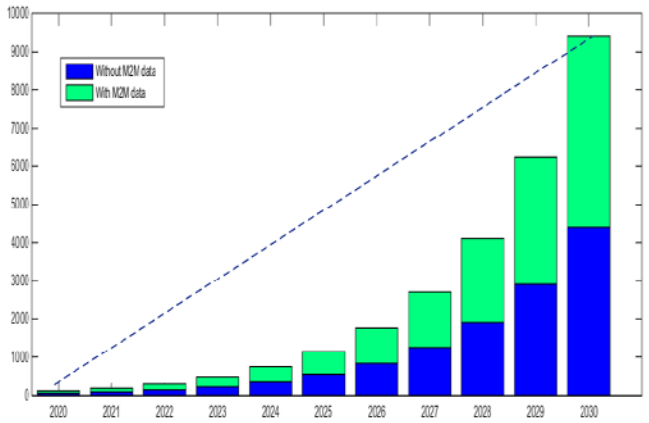


Fig. 13. Global Mobile Data Traffic Forecasting by ITU.

G. Blockchain Technology (BCT) and IR 5.0

BCT has the potential to provide significant value to IR 5.0 in the future. In IR 5.0, centralized administration of a high number of heterogeneous linked devices is a major difficulty. By enabling distributed trust, BCT may be utilized to construct decentralized and distributed management solutions. For effective subscriber management in IR 5.0, BCT may be utilized to generate digital identities for various persons and businesses. It's required for access management and authentication of stakeholders in any industrial activity that takes place via the internet [62].

Additionally, these digital identities may be used to manage properties, belongings, items, and services. BCT may also be used to catalog and save original work and register IP rights. By automating the agreement procedures between diverse parties, BCT can also assist to automate the contractual process [63]. According to [60], by 2028, the worldwide BCT industry is anticipated to be worth USD 394.60 billion. At the same time, financial institutions' increased interest in BCT is propelling industry expansion. According to [64], the market size of Blockchain is increasing with time as shown in Fig. 14, it shows that BCT is a widely used technology of the current time.

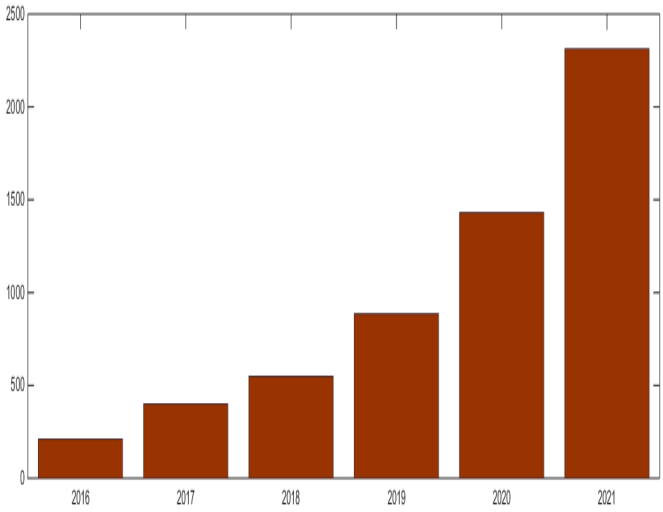


Fig. 14. Blockchain Market Size Worldwide.

H. Other Enabling Technologies

In addition to the above-discussed technologies, other CET such as EC, network slicing, augmented reality, and 3D printing also plays a vital role in IR 5.0. EC reduces communication costs and ensures that programs run smoothly even at faraway locations. Additionally, EC can process data without sending it to the public cloud, reducing security concerns for IR 5.0's major events [65]. Because IR 5.0 supports a broad collection of applications, a single physical infrastructure will be unable to meet the needs of heterogeneous networks. In this situation, network slicing can provide several virtualized networks at a low cost [66]. The way information is accessed, used, and transferred is changing because of augmented reality. In truth, augmented reality is data that broadens people's perceptions. This technology enhances our sensory perception by allowing us to engage with data. Because augmented reality can actively contribute to the success and transformation of industrial production processes, it will be useful as IR 5.0 focuses on human-machine collaboration [67]. The usage of 3D technology is increasing in a variety of industries, from food printing to the construction of Mars settlements. Innovations in the healthcare, automotive, construction, and manufacturing sectors are among the more practical applications for 3D printing. The 3D printer is being dubbed the forerunner of the 5th IR by scholars and practitioners for these reasons [68].

VI. Challenges of IR 5.0

Despite the benefits offered by IR 5.0, it also faces several challenges. Some of the most significant challenges of IR 5.0, as taken from available research [23-26, 69], are as follows:

1) Initial cost: IR 5.0 necessitates a significant amount of expenditure to completely execute all of its pillars, which is challenging for the industry, particularly SMEs, to accomplish.

2) Lack of precision and accuracy: For example, IR 5.0 has a lot of potential in the healthcare business, but it requires a lot of precision and accuracy. The research in this area is still in its infancy, and it necessitates a significant amount of investment and infrastructure.

3) Technology requirements: This presents a challenge for startups and entrepreneurs, as IR 5.0 necessitates a significant amount of investment and infrastructure, as well as CET requirements.

4) Skill-gap: This tendency exacerbates job polarization, as middle-skill employment declines and the workforce is divided into two groups: highly trained and qualified personnel and low-paid and unqualified workers. This may help to bridge the gap between the skilled and unskilled in society.

5) Need training: Due to highly automated manufacturing systems, skill development is a massive task that includes training employees to adopt advanced CET as well as inducing behavioral changes to interact with them.

6) Risk: Collaborative robotics is a type of technology that, along with human coworkers, poses a significant risk on the factory floor.

7) Data integration: It is difficult to obtain high quality and integrity data from industrial systems, and it is also challenging to accommodate several data sources.

8) Regulatory system: Due to the high level of automation in IR 5.0, it is difficult to develop regulatory systems. For example, who should be held accountable and to what extent in the event of a failure.

9) Process tailoring: The old company strategy and business models must be adjusted and tailored to match the requirements of IR 5.0 due to a greater degree of automation in the industries.

10)Mass personalization: As a result of mass personalization, company strategy will become more customer-centric. Customer subjectivity shifts throughout time, making it tough to adapt corporate strategies and models regularly.

VII. Discussion

Humans have recognized the possibility of using technology as a tool of advancement since the first IR. Steam machines, assembly lines, and computers are just a few of the technological developments that have occurred over the previous several centuries, all with the goal of producing more powerful technology and enhancing productivity and effectiveness. IR 5.0 shifts the paradigm and ushers in a revolution by putting less emphasis on technology and assuming that the ultimate potential for advancement resides in human-machine cooperation. IR 5.0 is not a passing trend, but rather it's a manufacturing paradigm change with ramifications for productivity, economics, and commerce. Due to the competitive benefits that the IR 5.0 model offers, organizations that do not adapt their production to this model will quickly become outdated.

IR 5.0 acknowledges that industry can fulfill social objectives beyond employment and development, such as becoming a dependable source of wealth, by making sure that production takes into account the limitations of our planet and places a premium on the well-being of industry employees. Keeping in view the benefits of IR 5.0, this study provides a full description of IR 5.0, as well as the functions of CET, so that scholars and practitioners may appreciate the significance of this revolution and recommend ways for maximizing its benefits.

VIII. Conclusion

IR 5.0 expands on the established IR 4.0 paradigm by placing a premium on research and innovation as critical drivers of the transition to a better industry. It refocuses attention on stakeholder value, which helps everyone. It places a premium on worker well-being throughout the manufacturing process and leverages new technologies to create wealth beyond employment and development, all while keeping conscious of the planet's production restrictions. In the upcoming days, the industries which will not follow IR 5.0 will obsolete from the market. Keeping in view the importance of IR 5.0, this conceptual paper provides a detailed overview of IR 5.0. The key topics covered in this research article include: providing a comprehensive knowledge of the IR 5.0 concept from several angles, promising applications of IR 5.0, the role of modern CET in the fifth IR, and opportunities and challenges faced by IR 5.0. This paper will help researchers and industry practitioners to better understand the role of IR 5.0 in the upcoming era.

CET's involvement in real-time IR 5.0 settings will be further examined in the future. Additionally, we'll combine human expertise with CET to study the real-world effects of IR 5.0.

Acknowledgment

The author would like to acknowledge the support provided by the Deanship of Scientific Research at Jouf University, Saudi Arabia.

References

[1] Humayun, M., Industry 4.0 and Cyber Security Issues and Challenges. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021. 12(10): p. 2957-2971.

[2] Humayun, M., et al., Privacy protection and energy optimization for 5G-aided industrial Internet of Things. IEEE Access, 2020. 8: p. 183665-183677.

[3] Ragulina, Y.V., et al., Methodology of criterial evaluation of consequences of the industrial revolution of the 21st century, in Industry 4.0: Industrial Revolution of the 21st Century. 2019, Springer. p. 235-244.

[4] Nahavandi, S., Industry 5.0—A human-centric solution. Sustainability, 2019. 11(16): p. 4371.

[5] Alferidah, Dhuha Khalid, and N. Z. Jhanjhi. "Cybersecurity Impact over Bigdata and IoT Growth." In 2020 International Conference on Computational Intelligence (ICCI), pp. 103-108. IEEE, 2020.

[6] Almusaylim, Zahrah A., and Noor Zaman. "A review on smart home present state and challenges: linked to context-awareness internet of things (IoT)." Wireless networks 25, no. 6 (2019): 3193-3204.

[7] Leong, Y.K., et al., Significance of industry 5.0, in The Prospect of Industry 5.0 in Biomanufacturing. 2021, CRC Press. p. 95-114.

[8] Demir, K.A., G. Döven, and B. Sezen, Industry 5.0 and human-robot co-working. Procedia computer science, 2019. 158: p. 688-695.

[9] Xu, X., et al., Industry 4.0 and Industry 5.0—Inception, conception and perception. Journal of Manufacturing Systems, 2021. 61: p. 530-535.

[10] Maddikunta, P.K.R., et al., Industry 5.0: a survey on enabling technologies and potential applications. Journal of Industrial Information Integration, 2021: p. 100257.

[11] Chai, Y.H., et al., State-of-the-Art Technologies in Industry 5.0, in The Prospect of Industry 5.0 in Biomanufacturing. 2021, CRC Press. p. 257-286.

[12] Li, J.-Q., et al., Industrial internet: A survey on the enabling technologies, applications, and challenges. IEEE Communications Surveys & Tutorials, 2017. 19(3): p. 1504-1526.

[13] Ruppert, T., et al., Enabling technologies for operator 4.0: A survey. Applied Sciences, 2018. 8(9): p. 1650.

[14] Martinelli, A., A. Mina, and M. Moggi, The enabling technologies of industry 4.0: Examining the seeds of the fourth industrial revolution. Industrial and Corporate Change, 2021. 30(1): p. 161-188.

[15] Knudsen, M.S., J. Kaivo-oja, and T. Lauraeus. Enabling Technologies of Industry 4.0 and Their Global Forerunners: An Empirical Study of the Web of Science Database. in International Conference on Knowledge Management in Organizations. 2019. Springer.

[16] Hu, M., et al., Virtual reality: A survey of enabling technologies and its applications in IoT. Journal of Network and Computer Applications, 2021: p. 102970.

[17] Longo, F., A. Padovano, and S. Umbrello, Value-oriented and ethical technology engineering in industry 5.0: a human-centric perspective for the design of the factory of the future. Applied Sciences, 2020. 10(12): p. 4182.

[18] Sułkowski, Ł., Kolasińska-Morawska, K., Seliga, R. and Morawski, P., 2021. Smart Learning Technologization in the Economy 5.0—The Polish Perspective. Applied Sciences, 11(11), p.5261.

[19] Farsi, M., R.K. Mishra, and J.A. Erkoyuncu, Industry 5.0 for Sustainable Reliability Centered Maintenance. Available at SSRN 3944533, 2021.

[20] Aslam, F., et al., Innovation in the era of IoT and industry 5.0: absolute innovation management (AIM) framework. Information, 2020. 11(2): p. 124.

[21] Rachmawati, I., et al., Prevalence of Academic Resilience of Social Science Students in Facing the Industry 5.0 Era. International Journal of Evaluation and Research in Education, 2021. 10(2): p. 676-683.

[22] Al Faruqi, U., Future Service in Industry 5.0. Jurnal Sistem Cerdas, 2019. 2(1): p. 67-79.

[23] ElFar, O.A., et al., Prospects of Industry 5.0 in algae: Customization of production and new advance technology for clean bioenergy generation. Energy Conversion and Management: X, 2021. 10: p. 100048.

[24] Paschek, D., A. Mocan, and A. Draghici. Industry 5.0-The expected impact of next Industrial Revolution. in Thriving on Future Education, Industry, Business, and Society, Proceedings of the MakeLearn and TIIM International Conference, Piran, Slovenia. 2019.

[25] Skobelev, P. and S.Y. Borovik, On the way from Industry 4.0 to Industry 5.0: From digital manufacturing to digital society. Industry 4.0, 2017. 2(6): p. 307-311.

[26] Ngo, L., The influence of ICT on the accommodation industry in the upcoming industry 5.0. 2019.

[27] Lasi, H., Fettke, P., Kemper, H.G., Feld, T. and Hoffmann, M., 2014. Industry 4.0. Business & information systems engineering, 6(4), pp.239-242.

[28] Potočan, V., M. Mulej, and Z. Nedelko, Society 5.0: Balancing of Industry 4.0, economic advancement and social problems. Kybernetes, 2020.

[29] Zengin, Y., et al., An investigation upon industry 4.0 and society 5.0 within the context of sustainable development goals. Sustainability, 2021. 13(5): p. 2682.

[30] Polat, L. and A. Erkollar. Industry 4.0 vs. Society 5.0. in The International Symposium for Production Research. 2020. Springer.

[31] Javaid, M., et al., Industry 5.0: Potential applications in COVID-19. Journal of Industrial Integration and Management, 2020. 5(04): p. 507-530.

[32] Ullah, A., et al., Secure healthcare data aggregation and transmission in IoT—A survey. IEEE Access, 2021. 9: p. 16849-16865.

[33] Irujo Aizcorbe, J., A review on human robot collaboration and its application in the health care sector. 2020.

[34] Gong, B., Nugent, J.P., Guest, W., Parker, W., Chang, P.J., Khosa, F. and Nicolaou, S., 2019. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: ANational survey study. Academic radiology, 26(4), pp.566-577.

[35] Majid, Mahardhika Ishlah, Cattleya Khansa Darmawan, Suharto Abdul Majid, and Yuda Yulianto. "Anticipating the Entry of Industry 5.0 in Transportation Sector." Advances in Transportation and Logistics Research 2 (2019): 103-115.

[36] Leea, Wen - Chieh, and Shinn-Shyr Wangb. "Misallocations and Policy Constraints on Mergers in the Modern Manufacturing Sector."

[37] Rupa, C., Midhunchakkaravarthy, D., Hasan, M.K., Alhumyani, H. and Saeed, R.A., 2021. Industry 5.0: Ethereum blockchain technology based DApp smart contract. Mathematical Biosciences and Engineering, 18(5), pp.7010-7027.

[38] Frederico, G.F., From Supply Chain 4.0 to Supply Chain 5.0: Findings from a Systematic Literature Review and Research Directions. Logistics, 2021. 5(3): p. 49.

[39] Taranenko, N.Y., et al., Education as socio-cultural and economic potential of the global information society. Journal of History Culture and Art Research, 2019. 8(1): p. 136-145.

[40] Saxena, A., et al., Emergence of Educators for Industry 5.0-An Indological Perspective.

[41] Begum, Salma. "A Study on growth in Technology and Innovation across the globe in the Field of Education and Business." International Research Journal on Advanced Science Hub 3 (2021): 148-156.

[42] https://www.statista.com/statistics/1200006/industry-40-technology-greatest-impact-organizations-worldwide/.

[43] Ullah, A., et al., Secure Critical Data Reclamation Scheme for Isolated Clusters in IoT enabled WSN. IEEE Internet of Things Journal, 2021.

[44] Alkinani, M.H., et al., 5G and IoT Based Reporting and Accident Detection (RAD) System to Deliver First Aid Box Using Unmanned Aerial Vehicle. Sensors, 2021. 21(20): p. 6905.

[45] Özdemir, V. and N. Hekim, Birth of industry 5.0: Making sense of big data with artificial intelligence,"the internet of things" and next-generation technology policy. Omics: a journal of integrative biology, 2018. 22(1): p. 65-76.

[46] Alam, T., A reliable communication framework and its use in internet of things (IoT). CSEIT1835111| Received, 2018. 10: p. 450-456.

[47] Humayun, M., Role of emerging IoT big data and cloud computing for real time application. Int. J. Adv. Comput. Sci. Appl., 2020. 11(4): p. 1-13.

[48] Mishra, S.K., et al., Energy-aware task allocation for multi-cloud networks. IEEE Access, 2020. 8: p. 178825-178834.

[49] Alayda, S., et al., A Novel Hybrid Approach for Access Control in Cloud Computing.

[50] Islam, T. and HASAN, M., 2017. A Performance Analysis of a Typical Server running on a Cloud.

[51] Fukuda, K., Science, technology and innovation ecosystem transformation toward society 5.0. International journal of production economics, 2020. 220: p. 107460.

[52] Majeed, A., et al., A big data-driven framework for sustainable and smart additive manufacturing. Robotics and Computer-Integrated Manufacturing, 2021. 67: p. 102026.

[53] Gagan, B. R., S. K. Majumdar, and S. Menon. "Application of Data Science in Transforming the Digital Economy: Evidence From Global Big Data Analytics Service Providers." In 2nd International Conference on Digital Entrepreneurship (ICDE 2019) Conference Proceedings, Bangalore, India (8, vol. 9, p. 2.

[54] Fei, T., et al., Digital twin-driven product design, manufacturing and service with big data. The International Journal of Advanced Manufacturing Technology, 2018. 94(9-12): p. 3563-3576.

[55] Teng, S.Y., et al., Recent advances on industrial data-driven energy savings: Digital twins and infrastructures. Renewable and Sustainable Energy Reviews, 2021. 135: p. 110208.

[56] Zubarev, A. E., O. V. Vatolina, and A. M. Kolesnikov. "Information And Communication Technologies Of Digital Transformation Of The Economy." In European Proceedings of Social and Behavioural Sciences EpSBS, pp. 435-442. 2020.

[57] Simões, A.C., A.L. Soares, and A.C. Barros, Factors influencing the intention of managers to adopt collaborative robots (cobots) in

manufacturing organizations. Journal of Engineering and Technology Management, 2020. 57: p. 101574.

[58] Sowa, K., A. Przegalinska, and L. Ciechanowski, Cobots in knowledge work: Human–AI collaboration in managerial professions. Journal of Business Research, 2021. 125: p. 135-142.

[59] Calitz, Andre P., Paul Poisat, and Margaret Cullen. "The future African workplace: The use of collaborative robots in manufacturing." SA Journal of Human Resource Management 15, no. 1 (2017): 1-11.

[60] Tariq, F., et al., A speculative study on 6G. IEEE Wireless Communications, 2020. 27(4): p. 118-125.

[61] Lu, Y. and X. Zheng, 6G: A survey on technologies, scenarios, challenges, and the related issues. Journal of Industrial Information Integration, 2020: p. 100158.

[62] Alamri, M., Jhanjhi, N.Z. and Humayun, M., 2019. Blockchain for Internet of Things (IoT) research issues challenges & future directions: A review. Int. J. Comput. Sci. Netw. Secur, 19, pp.244-258.

[63] Humayun, M., Jhanjhi, N.Z., Hamid, B. and Ahmed, G., 2020. Emerging smart logistics and transportation using IoT and blockchain. IEEE Internet of Things Magazine, 3(2), pp.58-62.

[64] Singh, A.P., Pradhan, N.R., Luhach, A.K., Agnihotri, S., Jhanjhi, N.Z., Verma, S., Ghosh, U. and Roy, D.S., 2020. A novel patient-centric architectural framework for blockchain-enabled healthcare applications. IEEE Transactions on Industrial Informatics, 17(8), pp.5779-5789.

[65] Du, A., et al., CRACAU: Byzantine Machine Learning meets Industrial Edge Computing in Industry 5.0. IEEE Transactions on Industrial Informatics, 2021.

[66] Xu, L., et al., AF Relaying Secrecy Performance Prediction for 6G Mobile Communication Networks in Industry 5.0. IEEE Transactions on Industrial Informatics, 2021.

[67] Ramalho, F. and A.L. Soares, Augmented reality in complex manufacturing systems as an informational problem: a human-centered approach. iConference 2020 Proceedings, 2020.

[68] Martynov, V.V., D.N. Shavaleeva, and A.A. Zaytseva. Information Technology as the Basis for Transformation into a Digital Society and Industry 5.0. in 2019 International Conference" Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS). 2019. IEEE.

[69] Zambon, I., et al., Revolution 4.0: Industry vs. agriculture in a future development for SMEs. Processes, 2019. 7(1): p. 36.

Detecting Distributed Denial of Service Attacks using Machine Learning Models

Ebtihal Sameer Alghoson, Onytra Abbass
Department of Information Technology
University of Tabuk, KSA

***Abstract*—The Software Defined Networking (SDN) is a vital technology which includes decoupling the control and data planes in the network. The advantages of the separation of the control and data planes including: a dynamic, manageable, flexible, and powerful platform. In addition, a centralized network platform offers situations that challenge security, for instance the Distributed Denial of Service (DDoS) attack on the centralized controller. DDoS attack is a well-known malicious attack attempts to disrupt the normal traffic of targeted server, network, or service, by overwhelming the target's infrastructure with a flood of Internet traffic. This paper involves investigating several machine learning models and employ them with the DDoS detection system. This paper investigates the issue of enhancing the DDoS attacks detection accuracy using a well-known DDoS named as CICDDoS2019 dataset. In addition, the DDoS dataset has been preprocessed using two main approaches to obtain the most relevant features. Four different machine learning models have been selected to work with the DDoS dataset. According to the results obtained from real experiments, the Random Forest machine learning model offered the best detection accuracy with (99.9974%), with an enhancement over the recent developed DDoS detection systems.**

Keywords*—Cybersecurity; distributed denial of service (DDoS); machine learning (ML); Canadian institute cybersecurity - distributed denial of service (CICDDoS2019) dataset

I. Introduction

SDN stands for Software Defined Network Technology, a new technology in the network world, in which the network management and control function is separated from the data routing function, through which engineers attempt to rearrange the parts and roles of all network infrastructure components that have not been modified since the 1980s. It is the transition from NCP to TCP / IP and since then no change has occurred. A change in the level of the network infrastructure to keep pace with the great development that takes place in the field of information technology, especially in virtual computing, which made virtualization of all layers, and the infrastructure is still intractable to this technology, so SDN technology is a successful attempt to separate the data layer from the control layer [1].

Denial of Service (DOS) It is one of the types of electronic attacks, and it is a very powerful technology that has been launched to attack network devices and services, and this type can separate different services from the Internet. Distributed Denial of Service (DDOS) is a more powerful type of DOS and uses multiple distributed attack points [2].

DoS was originally appeared by Gligor in an operating system context [3, 4], where DoS became widely employed. In general, DoS attack tries to reach more than one computer to reach a victim in a coordinated manner is called a DDoS attack.

Software Defined Network (SDN) infrastructure is vulnerable to several security threats. Among the DDoS attacks are the most dominant one. The DDoS attacks are considered as one of the most destructive attacks in the Internet. In general, most website hacking are probably a DDoS attack. The DDoS attack aims to disrupting the normal operation of the system through making services and resources unavailable to legitimate users by overloading the system with unnecessary superfluous traffic from distributed source. In addition, DDoS attack aims to increase in strength and frequency day-by-day. Therefore, the new systems which have been developed should be able to enhance the performance requirements and improve scalability of modern data centers, and provide maximum protection against the DDoS attacks.

This paper aims to mitigate denial of service attacks in software-defined networks through developing an efficient DDoS detection system based on machine learning models. The main contributions of this paper includes the following:

1) Research and analyze the recent developed DDoS detection systems.

2) Adopt several feature selection methods before processing the training stage.

3) Employ various machine learning models in the training process in order to enhance the efficiency of the DDoS detection system.

4) Test the developed machine learning model using real datasets, and real experiments, in order to assess the efficiency of the DDoS detection system.

II. Related Work

This section discusses the recent developed DDoS detection systems employed using the CICDDoS2019 dataset. Authors of [5] proposed a hybrid machine learning-based system to detect DDoS attacks. The proposed system involves combining the Extreme Learning Machine (ELM) algorithm and the black-hole optimization algorithm. Authors conducted several experiments through adopting various datasets to assess the performance of the proposed hybrid machine learning system. The proposed hybrid system has been employed in

detecting the DDoS attacks in cloud computing, and achieves 99.80% detection accuracy using the CICDDoS2019 dataset.

On the other hand, authors of [6] proposed an Intrusion Detection System against DDoS attacks (DDoSNet) in SDN environments. The proposed system is based on the Deep Learning (DL) technique, integrating the Recurrent Neural Network (RNN) with autoencoders. The developed system has been evaluated using the CICDDoS2019 dataset. Authors obtained a significant enhancement in attack detection compared to the existing methods. Therefore, the proposed system offers great confidence in securing SDN environments.

The work presented in [7] includes examining the impact of data balancing algorithm in the network traffic classification problem on several types of DDoS attacks using the CICDDoS2019 dataset, which consists of various information about the reflection-based and exploitation-based attacks. The obtained results showed that the effectiveness of data balancing algorithms such as synthetic minority sampling, naïve random, and adaptive synthetic sampling in classifying network attacks.

Authors of [8] proposed a detection system which was able to detect the different types of DDoS attacks based on several classification algorithms using the CICDDoS 2019 dataset. In addition, authors captured packets from SDK environment, apply preprocessing function for the dataset, and then apply classification algorithm to detect the DDoS attacks. Authors revealed that the decision tree offers the better performance compared to SVM and Naïve Bayes machine learning models.

The work presented in [9] involves analyzing the success rate in the intrusion detection system through adopting several machine learning methods. The CICDDoS2019 dataset was employed, where several machine learning models were investigated, including: the ANN, Support Vector Machine (SVM), Gaussian Naïve Bayes, Multinomial Naïve Bayes, Bernouli aïve Bayes, Logistic Regression, K-nearest neighbor (KNN), Decision Tree, and Random Forest algorithms. Authors showed that the K-nearest neighbor, logistic regression, and Naïve Bayes offers the best prediction accuracy.

Authors of [10] employed the Deep Neural Network (DNN) as a deep learning method to detect the DDoS attacks on the sample of packets captured from network traffic. The DNN model can work rapidly and with high detection accuracy even with small samples, since it contains feature extraction and classification methods. Authors preformed their experiments using the CICDDoS2019 dataset which contains several DDoS attack types created in 2019. The proposed system achieves 94.57% accuracy rate using the deep learning model.

The work presented in [11] surveys the recent developed DDoS detection approaches using the machine learning models. Authors of [12, 13, 14] proposed a DDoS detection system using Naïve Bayes model. On the other hand, the support vector machine model has been adopted in this works [15, 16, 17] to detect the present of DDoS attacks. In addition, Decision Tree algorithm has also been adopted to detect the DDoS attacks, as presented in [18, 19].

TABLE I. A COMPARISION BETWEEN THE EXISTING SYSTEMS THAT EMPLOYED THE CICDDOS2019 DATASET

Research work	Algorithm	Detection Accuracy
[5]	Extreme Learning Machine & blackhole algorithms	99.80%
[6]	Recurrent Neural Network with autoencoders	92.54%
[7]	SMOTE	93.51%
[8]	Decision Tree	92.15%
[9]	Artificial Neural Network, Support Vector Machine, Guassian Naïve Bayes, Random Forest Algorithm & K-Nearest Neighbor	Naïve Bayes offers the best detection accuracy
[10]	Deep Neural Network	94.57%

As presented above, several DDoS detection systems have been developed recently based on the employment of the CICDDoS2019 dataset. Table I presents a comparison between the existing developed systems based on the algorithm used and the detection accuracy.

III. SYSTEM DESIGN

The Distributed Denial of Service (DDoS) attacks include transmitting multiple requests to the attacked web resource, with the goal of exceeding the website's capacity to handle multiple requests, and hence prevent the website from functioning correctly. Several researchers have discussed the DDoS attacks and analyzed the major security threats and the corresponding solutions. This section discusses the main methods which have been employed in order to develop the DDoS system. In addition, this section presents the experimental setup including: the development environment, the selected DoS datasets, and the experimental setup.

A. System Methodology

Fig. 1 shows the development process of the DDoS detection system. As presented below, the first stage includes searching an efficient DDoS dataset, that are being developed recently by several research works. The second stage involves cleaning up the dataset and apply feature extracting methods, in order to pick the most significant features. Next, several machine learning models will be implemented to test the performance of the developed machine learning system, and then obtain the model's accuracy after conducting the training and testing processes.

B. DDoS Dataset

For each single machine learning model, a training and testing processes are needed to be implemented in order to assess the performance of the developed machine learning model. An extensive research has been carried out in order to identify the best DDoS datasets, which will be employed later in the training and testing process. Several datasets are available online such as the CIC-DDoS2019 dataset, where it contains benign and the most up-to-date common DDoS attacks, which resembles the true real-world data.

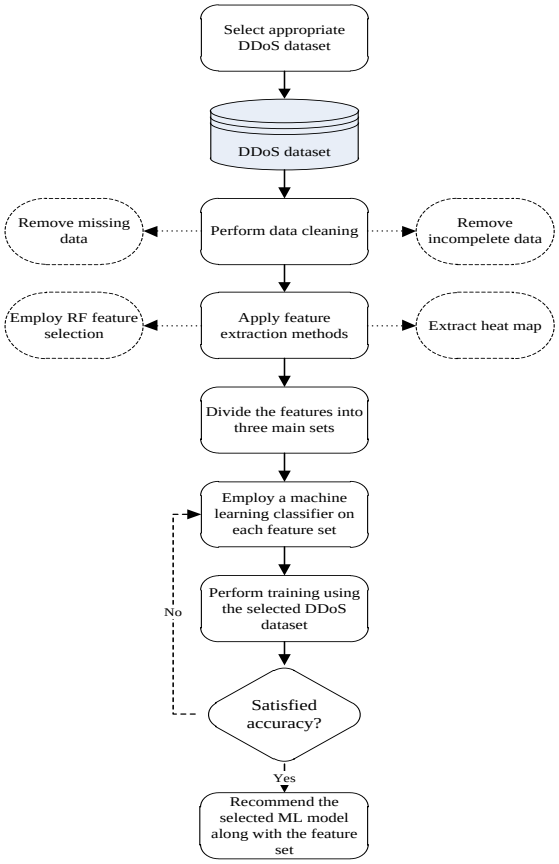


Fig. 1. The Development Process for the DDoS System.

Authors of [20] generated the CICDDoS2019 dataset that remedies several shortcomings and limitation which are presented in the existing datasets. CICDDoS2019 is labeled with 80 network traffic features that were extracted and calculated for all benign and denial of service flows. CIC-DDoS2019 dataset contains the results of the network traffic analysis with labeled flows based on the time stamp, source and destination IPs, source and destination ports, protocols and attack.

Therefore, the CICDDoS2019 dataset will be divided into two subsets: the training subset, and the testing subset. The training subset contains samples of data used to fit the machine learning models, whereas the testing subset is a gold standard employed to assess the performance of the trained machine learning model.

TABLE II. CICDDOS2019 DATASET GENERAL STATISTICS

Parameter name	**Total #**
Total number of records	12,794,627
Total number of features	82
Total number of labels	1
Total number of normal records	6,398,925
Total number of attack records	6,395,702
% of normal records	50.02%
% of attack records	49.98%

Table II shows the general statistics for the CICDDoS2019 dataset. CICDDoS2019 dataset is a large dataset in size and records, it consists of (12,794,627) records with a total memory size (6.3 gigabyte). The CICDDoS2019 is a balanced dataset, where the total number of normal records is (6,398,925) with the percentage of (50.02%), and the total number of fraud records is (6,395,702) with the percentage of (49.89%).

C. Data Preparation

In general, data preparation is considered as the most difficult stage in machine learning, and includes: data cleaning, data pre-processing, data wrangling, and feature engineering. Data preparation involves transforming raw data into a format where the machine learning algorithms can deal with, in order to uncover insights or make predictions. The data preparation process may consist of several steps, however, the most significant one involves processing the missing or incomplete data in the CICDDoS2019 dataset.

Data cleaning includes identifying and correcting errors or mistakes in the CICDDoS2019 dataset. Dropping columns that include missing or incomplete data, since missing and incomplete data affect the efficiency of the machine learning model. Therefore, it is important to process the missing and incomplete data in the dataset. For the CICDDoS2019 dataset, we noticed several attributes (columns) that contain zero values, and this will affect the machine learning model in negative way. For instance, *Fwd Byts/b Avg, Fwd Pkts/b Avg, Fwd Blk Rate Avg, Bwd Byts/b Avg, Bwd Pkts/b Avg, and Bwd Blk Rate Avg* attributes contain zero values in most of the records. Therefore, an important stage is required to remove these attributes from the CICDDoS2019 dataset.

The CICDDoS2019 dataset consists of several categorical data which are unsuitable for machine learning model. Therefore, there is a significant demand to remove these attributes from the CICDDoS2019 dataset in order to be able to train the machine learning model in a proper way. Moreover, the columns (attributes) that contain missing values more than 50% will be dropped from the CICDDoS2019 dataset. In addition, the rows where their columns contain more than 5% missing values are dropped. And finally, the faulty data in the CICDDoS2019 dataset are required to be considered. For instance, all records that contain negative values will be removed from the dataset.

The new shape for the dataset is presented in Fig. 2 after considering several data preparation methods. As noticed, the 19 attributes (columns) have been removed from the dataset, and 48,187 records have been removed from the CICDDoS2019 dataset.

On the other hand, the feature selection methods are considered next. According to [21], there are more than 2.5 quintillion bytes of data is produced every day. However, most of the generated data is required first to be pre-processed before starting any statistically analysis with the selected data, moreover, the produced data needs to be analysed using machine learning techniques in order to provide insights and to create predictions.

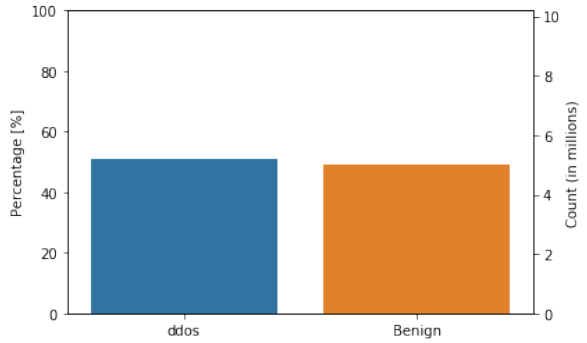


Fig. 2. Distribution of Records in the CICDDoS2019 Dataset.

As presented earlier in Table II, there are 82 features in the CICDDoS2019 dataset, and this makes the training and prediction tasks are very difficult. Therefore, it is important to minimize the number of features in CICDDoS2019 dataset through adopting several feature extraction models. This section discusses several methods which are used in order to extract the most significant features in the CICDDoS2019 dataset.

Minimizing the number of features may lead to several benefits, including: accuracy improvement, speed up in the training process, reducing the overfitting, and improve data visualization. Therefore, there are several different feature selection methods which can be applied to select the most significant feature in a given dataset, some of the most significant methods are: Filter method, and embedded method.

The first feature selection method is the filter method. Filter method involves filtering the dataset and take only a subset containing the most relevant features. This can be done using correlation matrix using Pearson Correlation. In general, the heat map (correlation matrix) is a graphical representation where individual values of matrix are represented as colours in order to display the correlation between attributes in a certain dataset and hence perform better prediction. The heat map for several features are shown below. For instance, Fig. 3 presents the heat map for 16 features, in order to show the relation among them. As seen in below, there is a high correlation between BWD IAT Std feature and FWD IAT Tot feature, and Bwd IAT Tot and FwdIAT Tot.

An embedded method is adopted next in order to enhance the prediction results. Embedded method includes examining the different training iterations of the machine learning model and then ranks the importance of the input features on how much each of the features contributed to the machine learning model through the training process.

For this stage, the Decision Tree model has been selected to rank the importance of CICDDoS2019's features. The Decision Trees models that are based on ensembles, can be used to rank the significance of the input features in the dataset. Since, extruding the most significant features offer vital importance on training the machine learning model, and hence obtaining efficient prediction accuracy. In addition, the features which will not offer any benefits to the machine learning model will be removed from the selected dataset.

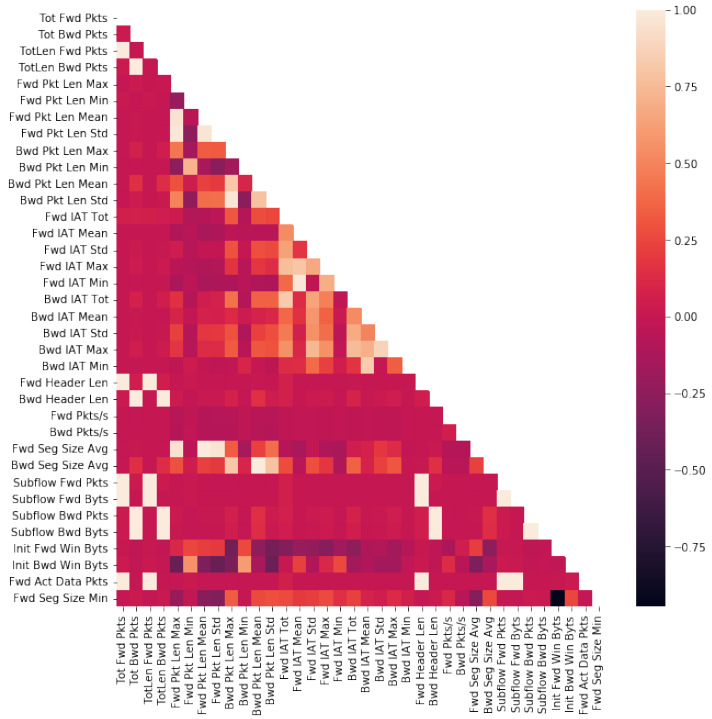


Fig. 3. The Heat Map for 36 Features.

In the Decision Tree, the CICDDoS2019 dataset was divided into two subsets: training subset, and testing subset, with 80% for training and 20% for testing. After completing the training process of the Random Forest Classifier, a set of feature importance plot is established according to the results obtained from the training stage. Fig. 4 shows the most 30 significant features in the CICDDoS2019 dataset.

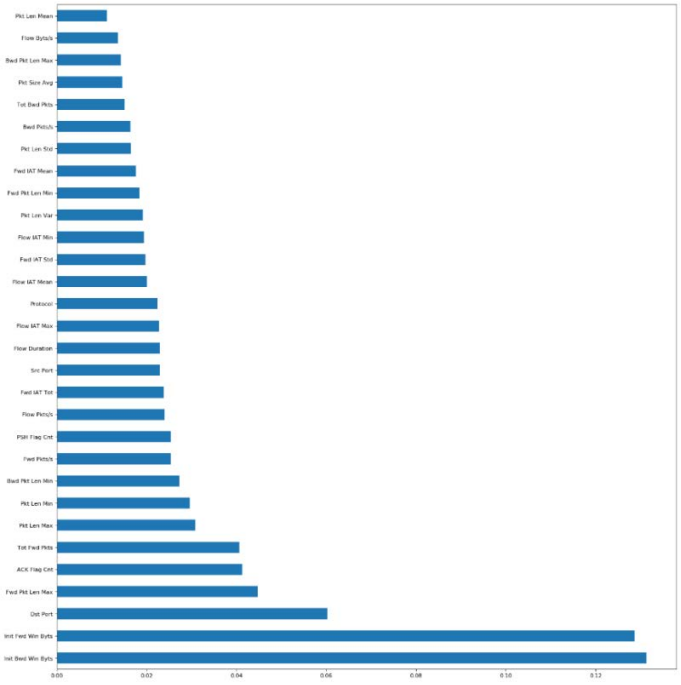


Fig. 4. The Feature Importance Plot for the 30 Most Significant Features.

In this paper, we investigate the efficiency of different features sets (10, 20, and 30), where each set is employed with every machine learning model and then assess and analyze the system's performance. The next section discusses the machine learning models which will be employed for the detection purposes.

D. DDoS Detection Models

In this project, several supervised machine learning models will be investigated, implemented, and tested, including: random forest, Light Gradient Boosting, CatBoost, and Convolutional Neural Networks.

- Random Forest (RF): it is also known as random decision forests that are ensemble learning method for classification and regression. RF operates through constructing multitude of decision trees at the training time, and producing the class which is the mode of the classes (classification) or average prediction (regression) of the individual trees.
- Light Gradient Boosting: is a light, fast, distributed, and high-performance gradient boosting framework, which is based on the decision tree algorithm, used for classification, ranking, and several machine learning tasks. It works by splitting up the tree leaf wise with the best fit, however, other boosting algorithms split the tree depth wise rather than the leaf-wise.
- CatBoost: CatBoost is an algorithm for gradient boosting on decision trees. CatBoost can be easily integrated with deep learning architectures. In addition, it can work with several data types to help solving a wide range of problem. CatBoost provides the best-in-class accuracy.
- Convolutional Neural Network (CNN): CNN is a deep neural networks, and is multilayer perceptron, which means that the CNN network is a fully connected. In any layer, each neuron is connected to all neurons in the next layer. CNN employs a mathematical operation named as convolution, where convolution is a specialized kind of linear operation.

IV. Expreimental Results

This section discusses the results obtained from several experiments conducted to assess the efficiency of different machine learning models. Several experiments have been conducted using the developed environment discussed earlier, in order to assess the DDoS systems' efficiency. Moreover, this section includes analyzing the obtained results and compares the system's efficiency with the recent developed systems.

A. Perforamcne Analysis

Several parameters are considered in order to assess the performance of the implemented DoS detection system; the parameters include:

- Average Training Time: this refers to the total time required to train the machine learning model.
- Accuracy: is the total of transactions that were correctly predicted over the total number of transactions.
- Precision: this indicates the total number of cases that were correctly classified among that class. Precision is the percentage of correctly predicted cases over the total predicted.
- Recall: is the ability of the classifier to correctly find all the positive instances. Recall is the ratio of true positives to the sum (total) of true positives and false negatives.
- Misclassification rate (error rate): this refers to how often the classifier is wrong.

B. Results of Average Training Time

The average training time is estimated for each machine learning model. As shown in Fig. 5, the RF model requires the largest training time (19,078 seconds), this is because the RF builds multiple decision trees and combines them together to obtain more accurate and stable prediction, and this makes the RF is a slow algorithm compares to others. Next, the average training time for the CNN model is (13,785 seconds), since the CNN training time depends on the training subset, batch size, and number of epochs.

On the other hand, the Light GB offers the minimum training time (150 seconds), since the Light GB is considered as a fast machine learning model for three main reasons: First, it splits the data based on their histogram, Second, it is gradient-based one-side sampling, and Third, the Light GB is used to deal with sparse features. Therefore, the Light GB machine learning model is best in terms of training time.

C. Evaluation of Essential ML Metrics

According to [22], there are three main metrics used to assess the machine learning classification model which are: accuracy, precision, and recall. As discussed earlier in the previous section, four different machine learning models were evaluated, where each machine learning model was evaluated through employing three sets of features. The best RF model was with 20-features set which offers (99.99740%) accuracy. On the other hand, the best LGB machine learning model was with the 20-features set which offers (99.99146%). The best accuracy result for the CatBoost model was with the 30-features set with accuracy (99.98592%). And finally, the best accuracy results for the CNN model was with 20-features set. Table III shows the detection accuracy for the best 4 machine learning models.

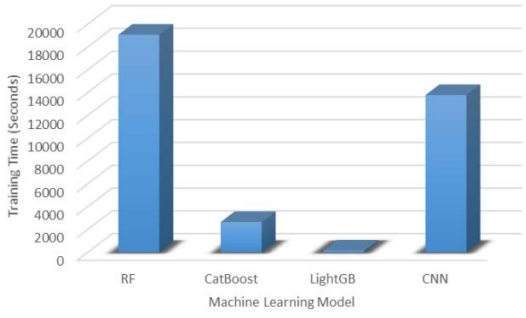


Fig. 5. Average Training Time (in Seconds) for 4 Machine Learning Models.

TABLE III. EVALUATION OF ACCURACY METRIC FOR 4 MACHINE LEARNING MODELS

	Accuracy
RF 20-features set	99.99740%
Light GB 20-features set	99.99146%
CatBoost 30-featues set	99.98592%
CNN 30-features set	98.29388%

TABLE IV. EVALUATION OF PRECISION METRIC FOR 4 MACHINE LEARNING MODELS

	Precision
RF 20-features set	99.99681%
Light GB 20-features set	99.98889%
CatBoost 30-featues set	99.97837%
CNN 20-features set	98.38997%

TABLE V. EVALUATION OF RRECALL METRIC FOR 4 MACHINE LEARNING MODELS

	Recall
RF 30-features set	99.99816%
Light GB 20-features set	99.99430%
CatBoost 30-featues set	99.99391%
CNN 20-features set	99.52368%

TABLE VI. EVALUATION THE FALSE NEGATIVE RATE FOR 4 MACHINE LEARNING MODELS

	False Negative Rate
RF 30-features set	19
Light GB 20-features set	59
CatBoost 30-featues set	63
CNN 30-features set	4,932

The Precision metric is discussed next, where the precision metric was assessed for every machine learning model. Precision refers to how often the machine learning model is able to predict the correct answer. The RF model with 20-feature set offers the best precision result (99.99681%). However, the best precision result using the Light GB model was through adopting 20-feature set with (99.9889%) result, whereas the CatBoost model achieves the best precision result with 30-feature set (99.97837%), and finally, the CNN model offers the best precision result with 20-feature set with (98.3899%) result. Consequently, as presented in Table IV, the machine learning model with best precision result was the Random Forest with 20-feature set.

Finally, the Recall metric is studied in this section. As discussed earlier in the previous section, the RF model with 30-feature set offers the best recall accuracy (99.99816%) among all the RF models (the three trained RF models using different number of features), whereas the Light GB model with 20-features set achieves the best recall accuracy (99.99430%) between all the LGB models. The CatBoost 30-features set offers the best recall result (99.99391%) amongst all the CatBoost models. And finally, the CNN 20-features set offers the best recall results among all the trained models with various number of features. Table V shows the recall results for 4 different machine learning models, and presents that the RF model with 30-features set offers the best recall results (99.99816%).

D. Results of False Negative Rates

This section evaluates the False Negative Rate (FNR) for each machine learning model employed above. FNR is a significant factor and refers to incorrectly predict the absence of DDoS attack when it is actually present, and this is the most significant metric in DDoS attack detection systems. Therefore, it is important to deal with machine learning model with the minimum FNR.

In this section, 14-different experiments were conducted to assess the efficiency of various machine learning models using 3 different DDoS subsets. Table VI presents the best FNR for 4 machine learning models. As presented in the Table below, the Random Forest classifier with 30-features set offers the best FNR with only 19 DDoS records which were misclassified and predicted as normal DDoS packets, whereas a large difference arises when adopting the CNN model. Fig. 4 depicts the false negative records for each machine learning model.

V. DISCUSSION

This section discusses the results obtained in this report with the results obtained from the previous research works, considering the CICDDoS2019 dataset. Most of the existing works evaluated the efficiency of the DDoS prediction model using the accuracy metric. Therefore, this section compares the accuracy metric obtained in this paper, with the existing works developed recently.

In this work, the detection accuracy that was achieved equal to 99.99740% using the random forest machine learning model with 20-features set. The high detection accuracy refers to the pre-processing methods which have been employed on the CICDDoS2019 dataset before applying the machine learning model. Two different feature selection methods were employed in this paper: filter method and feature extraction methods, in order to extract the most important feature which affect the machine learning model.

Therefore, in this paper, as shown in the previous section, the obtained detection accuracy results are greater than the results obtained from the recent developed works. Table VII presents the overall detection accuracy for various machine learning models used to detect the DDoS attacks. Fig. 6 shows the detection accuracy for several machine learning models, with different classification accuracy.

TABLE VII. DETECTION ACCURACY FOR SEVERAL MACHINE LEARNING MODELS

	Detection Accuracy
[5]	99.80
[6]	92.54
[7]	93.51
[8]	92.15
[10]	94.57
This system	99.99

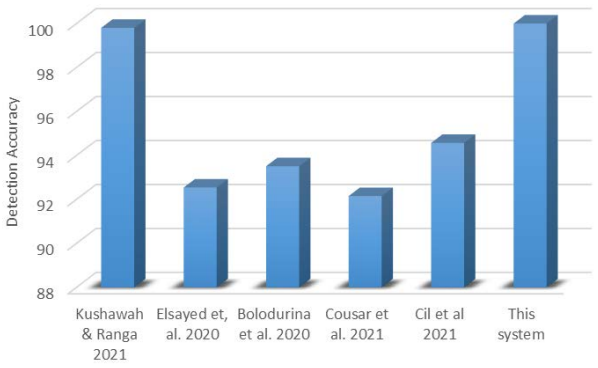


Fig. 6. The Detection Accuracy for Several Machine Learning Models.

VI. Conclusion and Future Work

Recently, the DDoS attack is considered as one of the most significant attack, which is a very powerful technology that has been launched to attack network devices and services. Therefore, in this paper, we consider the DDoS attack to be studied, analysed, and develop a machine learning model to detect such attacks. In this paper, we employed several feature selection methods in order to select the most significant features that can be used to predict the DDoS attacks in an efficient way. Three sets of features have been chosen from the selected dataset, and employed with four machine learning models. According to the obtained results, the RF-machine learning model with 20-features set offers the best precision, accuracy, recall, and false negative rate. For future work, we aim to work with real-time DDoS detection systems which will be able to detect the DDoS attack in real-time situations. Therefore, in this paper, we offered significant improvement in the detection of DDoS attacks using the CICDDoS2019 dataset.

References

[1] Faujdar, N., Sinha, A., Sharma, H., & Verma, E. (2020, October). Network Security in Software defined Networks (SDN). In *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)* (pp. 377-380). IEEE.

[2] Iqbal, M., Iqbal, F., Mohsin, F., Rizwan, M., & Ahmad, F. (2019). Security Issues in Software Defined Networking (SDN): Risks, Challenges and Potential Solutions. *International Journal of Advanced Computer Science and Applications*, *10*(10).

[3] K. Lakshminarayanan, D. Adkins, A. Perrig, and I. Stoica, "Taming ip packet flooding attacks," ACM SIGCOMM Computer Communication Review, vol. 34, no. 1, pp. 45–50, 2004.

[4] V. D. Gligor, "A note on denial-of-service in operating systems," IEEE Transactions on Software Engineering, no. 3, pp. 320–324, 1984.

[5] Kushwah, G.S. and Ranga, V., 2021. Optimized extreme learning machine for detecting DDoS attacks in cloud computing. *Computers & Security*, *105*, p.102260.

[6] Elsayed, M.S., Le-Khac, N.A., Dev, S. and Jurcut, A.D., 2020, August. Ddosnet: A deep-learning model for detecting network attacks. In *2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)* (pp. 391-396). IEEE.

[7] Bolodurina, I., Shukhman, A., Parfenov, D., Zhigalov, A. and Zabrodina, L., 2020, November. Investigation of the problem of classifying unbalanced datasets in identifying distributed denial of service attacks. In *Journal of Physics: Conference Series* (Vol. 1679, No. 4, p. 042020). IOP Publishing.

[8] Kousar, H., Mulla, M.M., Shettar, P. and Narayan, D.G., 2021, June. Detection of DDoS Attacks in Software Defined Network using Decision Tree. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 783-788). IEEE.

[9] Aytaç, T., Aydın, M.A. and Zaim, A.H., 2020. Detection DDOS Attacks Using Machine Learning Methods.

[10] Cil, A.E., Yildiz, K. and Buldu, A., 2021. Detection of DDoS attacks with feed forward based deep neural network model. *Expert Systems with Applications*, *169*, p.114520.

[11] Arshi, M., M. D. Nasreen, and Karanam Madhavi. "A Survey of DDOS Attacks Using Machine Learning Techniques." In *E3S Web of Conferences*, vol. 184, p. 01052. EDP Sciences, 2020.

[12] A. Bivens, C. Palagiri, R. Smith, B. Szymanski, M. Embrechts, et al, "Networkbased intrusion detection using neural networks," Intelligent Engineering Systems through Artificial Neural Networks, vol. 12, no. 1 , pp. 579–584, 2002.

[13] Jasreena Kaur Bains ,Kiran Kumar Kaki ,Kapil Sharma,‟ Intrusion Detection System with Multi-Layer using Bayesian Networks‟, International Journal of Computer Applications (0975 – 8887) Volume 67– No.5, April 2013.

[14] M. Alkasassbeh, G. Al-Naymat et.al," Detecting Distributed Denial of Service Attacks Using Data Mining Technique," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, pp. 436-445, 2016. Science and Information Technologies, Vol. 6 (2), pp. 1096-1099, 2015.

[15] Mangesh Salunke, RuhiKabra, Ashish Kumar.‟ Layered architecture for DoS attack detection system by combine approach of Naive Bayes and Improved Kmeans Clustering Algorithm‟, International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 03, June-2015.

[16] T. Subbulakshmi et.al, ‟A Unified Approach for Detection and Prevention of DDoS Attacks Using Enhanced Support Vector Machine and Filtering Mechanisms‟, ICTACT Journal on Communication Technology, June 2013.

[17] Yogeswara Reddy B, Srinivas Rao J, Suresh Kumar T, Nagarjuna A, International Journal of Innovative Technology and Exploring Engineering, Vol.8, No. 11, 2019, pp: 1194- 1198.

[18] HodaWaguih, "A Data Mining Approach for the Detection of Denial of Service Attack", International Journal of Artificial Intelligence, vol. 2 pp. 99106(2013).

[19] Dewan Md. Farid, Nouria Harbi, EmnaBahri, Mohammad Zahid ur Rahman, Chowdhury Mofizur Rahman,‟ Attacks Classification in Adaptive Intrusion Detection using Decision Tree "International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:4, No:3, 2010.

[20] Sharafaldin, I., Lashkari, A.H., Hakak, S. and Ghorbani, A.A., 2019, October. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In *2019 International Carnahan Conference on Security Technology (ICCST)* (pp. 1-8). IEEE.

[21] Subramaniam, A. What is Big Data? — A Beginner's Guide to the World of Big Data. Accessed at: https://www.edureka.co/blog/what-is-big-data/.

[22] Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J. and Asadi, H., 2019. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, *212*(1), pp.38-43.

A Patient Care Predictive Model using Logistic Regression

Harkesh J. Patel, Jatinderkumar R. Saini*
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University), Pune, India

***Abstract*—Medical treatments and operations in hospitals are divided into in-patient and out-patient procedures. It is critical for patients to know and understand the differentiation between these two forms of treatment since it will affect the time of a patient's stay in a hospital or a medical institution as well as the cost of a treatment. In today's era of information, a person's talents and expertise may be put to good use by automating activities wherever possible. A medical service will be termed inpatient care if a doctor issues an order and the patient is admitted to the hospital on that order whereas a patient seeking outpatient care do not need to spend the night in a hospital. Choosing between in-patient and out-patient care is usually a matter of how involved the doctor wants to be with the patient's treatment. With the aid of numerous data points regarding the patients, their illnesses, and lab tests, our main objective is to develop a system as part of the hospital automation system that predicts and estimates whether the patient should be given an in-patient care or an out-patient care. The main idea of the paper is to understand and develop a logistic regression model to predict whether a patient needs to be treated as an in-patient or an out-patient depending on the results of laboratory tests. Furthermore, this study also focuses on how logistic regression performs for this dataset. In addition, research on how logistic regression performs for this dataset was also not done. From the study, the results show that logistic regression gives an accuracy of 75%, F1-score of 73%, precision of 74% and recall of 74%.**

***Keywords*—Health-care; inpatient care; logistic regression; machine learning model; outpatient care; stacking classifier**

I. Introduction

Inpatient and outpatient procedures can be used to classify medical treatments and operations. A patient should be aware of this distinction since it affects how long he or she stays in a hospital and how much the operation costs. The difference between an in-patient and an out-patient treatment is the duration of time a patient must spend time in a medical institution where the operation or treatment is performed. In-patient treatment compulsorily requires an overnight stay in the hospital or the medical institution.

Patients must spend time or stay at least for one night in the medical facility that is provided to them where their operation or treatment was performed, typically a hospital. During this period, kids are normally under the care of a nurse or a doctor.

Patients seeking out-patient care need not have to spend the night in a hospital. These patients are allowed to depart the hospital once their treatment is completed. In certain situations, they must wait while the anesthetic wears off or to ensure that there are no problems. Patients are not required to spend the night under supervision if there are no significant problems. In addition, the out-patient treatment is much more cost-effective in comparison to in-patient treatment [12].

To categorise patients' medical requirements into an in-patient or an out-patient treatment, doctors rely heavily on the results of lab tests. This time-consuming approach requires doctors to exert considerable effort in order to determine if the patient is required to be in the hospital and closely watched or not. Also, the patient's life may be in risk if the wrong decision is made [1]. Clinical machine learning research is largely restricted to proof-of-concept studies. Machine learning applications in clinical medicine are currently hindered by a number of obstacles. A major shift in medical practise may result from overcoming obstacles to potential deployment, with the use of specialised technologies helping the healthcare team provide better, more customised patient care [2].

Machine learning algorithms, it is widely accepted, find and extract information based on the data available. Yet a vast quantity of information is available in machine-readable format, ready to be incorporated into machine learning algorithms and models [4]. For this study, Logistic Regression, a supervised machine learning algorithm, is used. The main objective of this paper is to see how logistic regression works on such a problem. Furthermore, we will check the performance metrics, precision, recall and accuracy to get an idea of how useful logistic regression is for such problems.

II. Literature Review

Melhem et al. [1] built four models depending on the patient's circumstances and lab test results: support vector machine model, decision tree model, random forest model and k-nearest neighbours model. The major aim of their study was to make use of ML algorithms to categorize the patient treatment as an in-patient or out-patient, in order to lessen the time and effort expended by the healthcare experts, which reflects the kind of services provided to the patient. Furthermore, this research assists in the reduction of human errors, which can result in hazards to the patient's life as well as an increase in the overall bill amount. The best model out of four was picked based on its accuracy, sensitivity, specificity, and precision scores, as well as its low false-negative & false-positive rates. To construct and evaluate these models, the EHR dataset was utilised, which comprises of patients' laboratory test results from a private hospital that is in Indonesia. The outcomes of their study say that random forest algorithm had the best accuracy (77 %), precision-rate (72%) & sensitivity (65%) as well as the model had the lowest false-

*Corresponding Author

negative rate (35%) and almost the lowest false-positive rate (16%).

Ben-Israel et al. [2] conducted a review in keeping with the prisma criteria and concentrating on human studies which utilised machine learning to directly treat a hard-headed scenario. The studies were performed between 1st Jan, 2000 & 1st May, 2018 and offered data on the performance of the used machine learning technique. Reviewers looked over 1909 distinct publications and found 378 retrospective papers and 8 prospective ones that met eligibility requirements. 61% of papers published in the past four years were retrospective. Few articles met our inclusion criteria, with just 2% of them being prospection articles. When it comes to clinical medicine, a majority of the literature is retrospective and focuses on proof-of-concept ways to improving patient care. A major transformation in medical practise will be enabled by recognizing the key translational hurdles, including instantaneous access to hard-headed data, data reliability, medical practitioner approval of "black box" generated findings and performance evaluation.

Beaulieu-Jones et al. [3] have talked about the applications of ML in health-care which is increasing quickly and might have a major effect on the profession. Using ML for health-care research, the study was aimed to provide quantitative and qualitative assessments of the current status. In order to assess the present status of research in ML for health-care, including areas of methodological and medical focus and limits, as well as areas that are underexplored, they analysed contributions. Results showed that the clinical collaborators were involved in 58 (34.9%) of the 166 accepted entries, and in 83 (50.0%) of the submissions that focused on clinical practise. On average, (97 datasets) 58% of the data sets utilised were publicly available or needed registration. (70 articles (42.2%)) of them were in clinical practise, with brain & mental health (25(15.1%)), cancer (21 (12.7%)) and cardiovascular (19 (11.4%)) being the most prevalent specialities. Data that is well-annotated and freely accessible is critical to the development of translational implementations in ML for health-care research, according to current trends.

Radovanović et al. [4] proposed an approach to logistic regression that incorporates domain information in the form of ontologies/hierarchies via layered generalisation. Because ontology/hierarchy relations are stacked, they may be combined to create higher, abstract ideas. In this case, they were able to tackle the problem of unexpected 30-day hospital readmissions. The proposed framework outperforms ridge, lasso and tree lasso logistic regression in terms of accuracy. This framework increases AUC by up to 9.5% for children and up to 4% for severely over-weight patients. Also, it increases the AUPRC up to 5.7% for children and up to 2.6% for ghastly over-weight patients, the researchers found.

Mu-Yen Chen [5], when using Decision Tree (DT) classification, the accuracy decreased the more PCA was applied, according to this paper's findings. According to his findings in the study of financial hardship, the accuracy of their DT classification technique increased as time passed, with an accuracy rate of 97.01% for two seasons previous to financial difficulty. They found that PCA increases the error while attempting to identify firms in a financial crisis as "normal" enterprises, and that DT classification has a higher short-term prediction accuracy than the Logistic Regression (LR) classification technique (less one year). Instead, the LR method improves long-term prediction accuracy (above one and half year). A short-term financial distress prediction model using AI, rather than standard statistical methods, is proposed in this study as a possible alternative to traditional statistical methods.

El-Rashidy et al. [6] have come up with a new way to forecast ICU patient death using stacking ensemble. Compared to the literature study, their method is more accurate and intuitive from a medical point of view in collaboration with an ICU domain specialist, data were produced and features selected. On the basis of the expert's judgments, six categories of data were created. When it came to the prediction procedure, each modality was assigned a distinct classifier depending on its performance. Our classifiers included linear discriminant analysis, decision tree algorithm, multilayer perceptron, k-nearest neighbour (knn) and logistic regression, among others. A stacking ensemble classifier was then created and tuned using the five classifier decisions. The system was validated with the help of a benchmark dataset of over ten thousand patients from MIMIC III. Patients' time series data of varied durations was used to undertake extensive studies in order to predict death. The first six, twelve, and twenty-four hours of a patient's initial stay were tested. On the basis of the results, their model surpassed the current techniques in terms of accuracy i.e. 94.4%, f1-score i.e. 93.7%, precision i.e. 96.4%, recall i.e. 91.1% & ROC curve i.e. 93.3%. As a result of these findings, it's clear that their technique of predicting ICU mortality works.

Polikar [7] studied situations where ensemble-based systems are superior to single-classifier systems, techniques for producing separate parts of ensemble systems and methods for combining separate classifiers. Many ensemble-based algorithms like bagging and boosting have been discussed as well as generalisation and hierarchical mixtures of experts. They have also discussed typically used combination rules such as algebraic blend of outputs, voting-based approaches and behaviour knowledge space as well as decision templates. A last look at future exploration prospects for ensemble systems was conducted. In addition, ensemble systems have showed significant promise in a variety of other fields like as feature selection and learning with lost features, confidence estimation and fault-correcting output codes. It has been demonstrated that ensemble-based systems generate better outcomes than expert systems for a broad range of implementations and circumstances. In their paper, they have talked about how to design, build, and use such systems.

Saini et al. [8] in their study seek to present the importance of artificial intelligence in magnetic learning algorithm and examines their function in different sectors of health, such as bioinformatics, cancer gene identification, epileptic seizure, brain computer interface. It also examines the medical imaging of illnesses, including diabetic retinopathy, gastro-intestinal disease, and tumour via extensive learning. Finally, this essay highlights the real barriers to AI approaches which need to be addressed. They examined the reason why ML was used in healthcare in this work. The main category of ML, is also

discussed. They concentrated on deep learning, its architecture and explore various health data analysing and examining deep learning. However, ML technologies draw considerable attention in the field of medical research. There are still difficulties with real-time implementation. Regulations are one such difficulty. Recent rules lack safety, evaluation and efficiency criteria for the ML system. The US FDA provides advice for the evaluation of ML systems to preserve security and efficiency in order to solve this challenge. The existing health care environment does not encourage the exchange of information on the system. It is also a limitation. The ML training was therefore compromised before implementation. In many nations, the Health Care Revolution encourages data exchange.

Kirasich et al. [13] addresses the challenge of model selection by assessing the overall classification performance for datasets with diverse underlying structures between the random forest and logistic regression by increasing the variance in the explanatory as well as the noise variables, number of explanatory variables, noise variables and observations. They created a model evaluation tool which can simulate classification models for such data and performance indicators as real positive rates, false positives and accuracy in certain circumstances. They observed that logistic regression has continuously exercised a better overall precision than random forests by increasing the variance not only in the explanatory but also the noise factors. The true positive-rate for random forest algorithm was, however, greater than the logistic regression algorithm & the data set with rising noise factors showed higher false-positive rates. Each and every case study included thousand simulations and the model executions in it consistently demonstrated that the false positive-rate for random forest with hundred trees was scientifically different from logistic regression. Under varied simulated dataset circumstances, logistic regression algorithm & random forest algorithm produced variable corresponding classification scores in all four situations.

Maroof [14] says that based on the continuous predictor, logistic regression seeks to classify or predict a discrete, categorical variable from among continuous or discrete predictors. Clinical neuropsychology's preference for using this paradigm in research is connected to the discipline's fundamental structure, which includes the use of scientific terminology to explain cognition and behaviour, as well as the compartmentalization of syndromes into diagnostic entities.

Goldarag et al. [15] developed, tested and compared forest fire risk prediction models based on logistic regression and neural networks. The findings show that the neural network model is more accurate at categorising fire points than logistic regression, which is sensitive to fire point samples. The percentage of fire and non-fire samples must be matched to obtain high accuracy in logistic regression. A neural network with two hidden layers, twenty-eight neurons, and a logarithmic-sigmoid transfer function in both hidden layers was also tested and the best architecture was found to be a neural network with two hidden layers, 28 neurons, and a logarithmic-sigmoid transfer function in both hidden layers.

Pepe & Thompson [16] developed strategies for maximising the accuracy of routinely used diagnostic measures by discovering linear combinations of markers. The approaches were non-distributional, appeared to have strong statistical features, and could account for heterogeneity defined by variables.

Sadikin and Mujiono [17] created an electronic health record predicting dataset obtained from a private hospital in Indonesia. It comprises the findings of the patient's laboratory tests, which are used to determine the next patient treatment, whether the patient is in or out of the hospital.

In order to determine the allocation of staff care in community-acquired pneumonia, España et al. [18] created a new prediction algorithm based on the 5 risk classifications described by the Pneumonia Severity Index. There was no question about the decision to hospitalise low risk (I-III) classes, when one or more of the following was evident, namely: tension in arterial oxygen < 8.0 kPa (60 mm Hg), shock, coexisting decompensating diseases, plural effusion, unable to maintain oral intake, a social problem and a lack of reaction to prior empirical antibiotic therapy. The findings are presented in a number of 616 patients after 18 months following application of this new prediction criteria. In 221 patients treated as ambulatory patients, the death rate was 0.5% vs 8.9% in 395 patients treated as hospitals. Of the 178 low risk individuals treated as hospital patients, 106 were given the specific extra requirements for hospitalisation, while the other 72 evidently did not justify the decision to admit hospitalisation under the predictive criterion. These 72 patients had better results than high-risk patient and low-risk patient who fulfilled the extra particular criteria for admitting to hospital (substantially shorter admission, antibiotic days, death and complex course) their results were better. Therefore, rigorous adherence to the new prediction criteria might have prevented admission in these low-risk individuals. Another significant finding was that not all patients admitted to the hospital had been identified in the Pneumonia Severity Index alone.

Blais et al. [19], in their study identified factors related with length of stay (LOS) and included measurement of these variables into their normal preadmission evaluation. A retrospective research of 80 discharged patients looked at the relationship between LOS and 25 factors representing a combination of patient/demographic characteristics, disease variables, and therapy variables. According to multivariate analysis, ten factors independently accounted for 62% of the variance in LOS. The information utilised was largely gathered during the pre-admission screening. In the prospective research, the factors' predictive ability decreased. However, fewer individual factors were substantially related to LOS; the total of the variables' scores predicted 17% of the LOS variation. The findings showed that significant criteria for predicting LOS are accessible at the time of admission, and these variables may be systematically examined and incorporated into clinical decision making.

Cuffel et al. [20] examined the correctness of different models for projecting a rehospitalization in a maintained mental health organisation, as well as the efficacy of various

care or treatment management methods for improving out-patient treatment follow-up. In a randomised controlled trial, patients that had in-patient mental health or substance use admissions were assigned to one out of 3 types of treatment supervision based on the level of involvement of treatment executives in the release planning & post-release outreach i.e. usual (N=31), enhanced (N=94) or intensive (N=74). Here, the classes that were formed were compared to each other and to a cohort hospitalised the year before the research and given usual treatment (N=192) to see if there were any changes in time to out-patient check out, quantity of post-release care and rehospitalization at 30, 60 & 180 days. There weren't any differences between the classes found. The larger part of number of patients i.e. 69% got out-patient treatment within 30 days of being released. The number of hard-headed and socio-demographic risk variables reported by care executives was associated to the probability of rehospitalization at 60 and 180 days, according to logistic regression prediction models. Patients who were approved to receive intermediate treatment i.e. partial hospitalisation and those who did not attend the intermediate treatment, if it was authorised were more suitable to be re-hospitalized at 30, 60 & 180 days than other patients. With increasingly extensive release planning and outreach, outpatient follow-up following mental hospitalisation did not improve. Improved prediction of re-hospitalization risk may improve possibilities to deliver intensive treatments to difficult-to-engage patients.

III. Methodology

Most of the data science implementations depend heavily on ML models. Other expert knowledge exists outside of the given data, which may theoretically assist the ML algorithms better recognize the conditions and circumstances of the data that is provided [4].

Machine Learning employs three types of learning: supervised, unsupervised and semi-supervised learning.

Supervised learning is a type of training in which we educate or train the machine using well-labelled data, i.e. data that already has the correct answer. Following that, the machine is given a fresh collection of instances, i.e. the test data, so that the supervised learning algorithm may analyse the training data and generate a proper result from labelled data. Further, supervised learning approaches include regression and classification, which are further classified. Regression is a helpful statistical prediction approach that aims to establish a meaningful link between dependent and independent variables by attempting to find correlations between them. To forecast a continuous output, the regression technique is employed in machine learning (ML). The predicted result is a real number. The output of classification is discrete, but the output of categorization is continuous [8].

Unsupervised learning, on the other hand, is the training of a computer utilising input that has not been categorised or labelled and enabling the algorithm to operate on that information without supervision. Clustering and principal component analysis (PCA) are two of the most important techniques in unsupervised learning. PCA is typically used to reduce the size of an object. With numerous dimensions, PCA reduces the data to a few principle component directions without losing much of the data. PCA is often used before clustering to minimise the number of dimensions of the data before it is clustered. Instead of using output information, the clustering approach is used to create a collection of variables that exhibit similarities or commonalities. As a result of these algorithms, the cluster labels for the variable with the highest degree of similarity within and between the clusters are generated and displayed on a graph [8].

Semi-supervised learning, on the other hand, includes function estimation on both labelled and unlabelled data. This method is driven by the fact that labelled data is frequently expensive to create, but unlabelled data is cheap.

Here, we have used Logistic Regression as it a binary classification task i.e. 0 or 1. A discrete, categorical variable is classified or predicted using logistic regression using continuous or discrete predictor, such as yes/no depending on the continuous predictor [14]. In order to achieve high accuracy in logistic regression, the percentage of in-patient and out-patient care samples must be balanced [15]. It produces a linear score that clearly distinguishes between two outcomes [16]. Moreover, the research on how logistic regression model performs on such scenario was also not done. General approaches such as ensemble learning can be used to improve the accuracy of prediction or classification models such as decision trees and artificial neural networks [11]. Hence, Stacking Classifier is used for improving the accuracy of prediction.

A. Dataset and Experimental Discussion

The dataset comprises predictions from an Electronic Health Record gathered from a private hospital which in Indonesia. It comprises the laboratory test results of different patients, which are used to decide the next patient's treatment, whether in or out of the hospital. The dataset contains 4412 rows and 11 columns. The total number of features are 10, where number of numerical features are 9 and number of categorical feature is 1 [17].

B. Attribute Information

- HAEMATOCRIT: It is the patient's laboratory test result of haematocrit.
- HAEMOGLOBINS: It is the patient's laboratory test result of haemoglobins.
- ERYTHROCYTE: It is the patient's laboratory test result of erythrocyte.
- LEUCOCYTE: It is the patient's laboratory test result of leucocyte.
- THROMBOCYTE: It is the patient's laboratory test result of thrombocyte.
- MCH: It is the patient's laboratory test result of MCH.
- MCHC: It is the patient's laboratory test result of MCHC.
- MCV: It is the patient's laboratory test result of MCV.
- AGE: It is the patient's age.

- SEX: It is the patient's gender.
- SOURCE: Target i.e. Binary: in-patient/out-patient – 0/1.

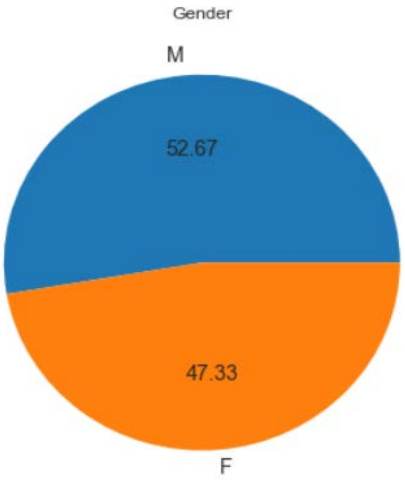


Fig. 1. Percentage of Males (M) and Females (F) in the Dataset.

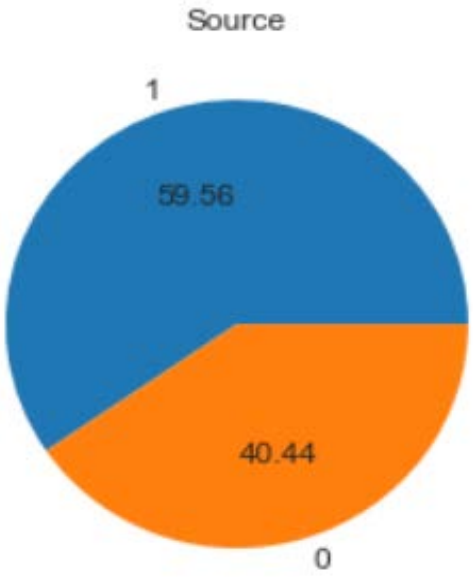


Fig. 2. Percentage of Inpatient (0) and Outpatient (1) in the Dataset.

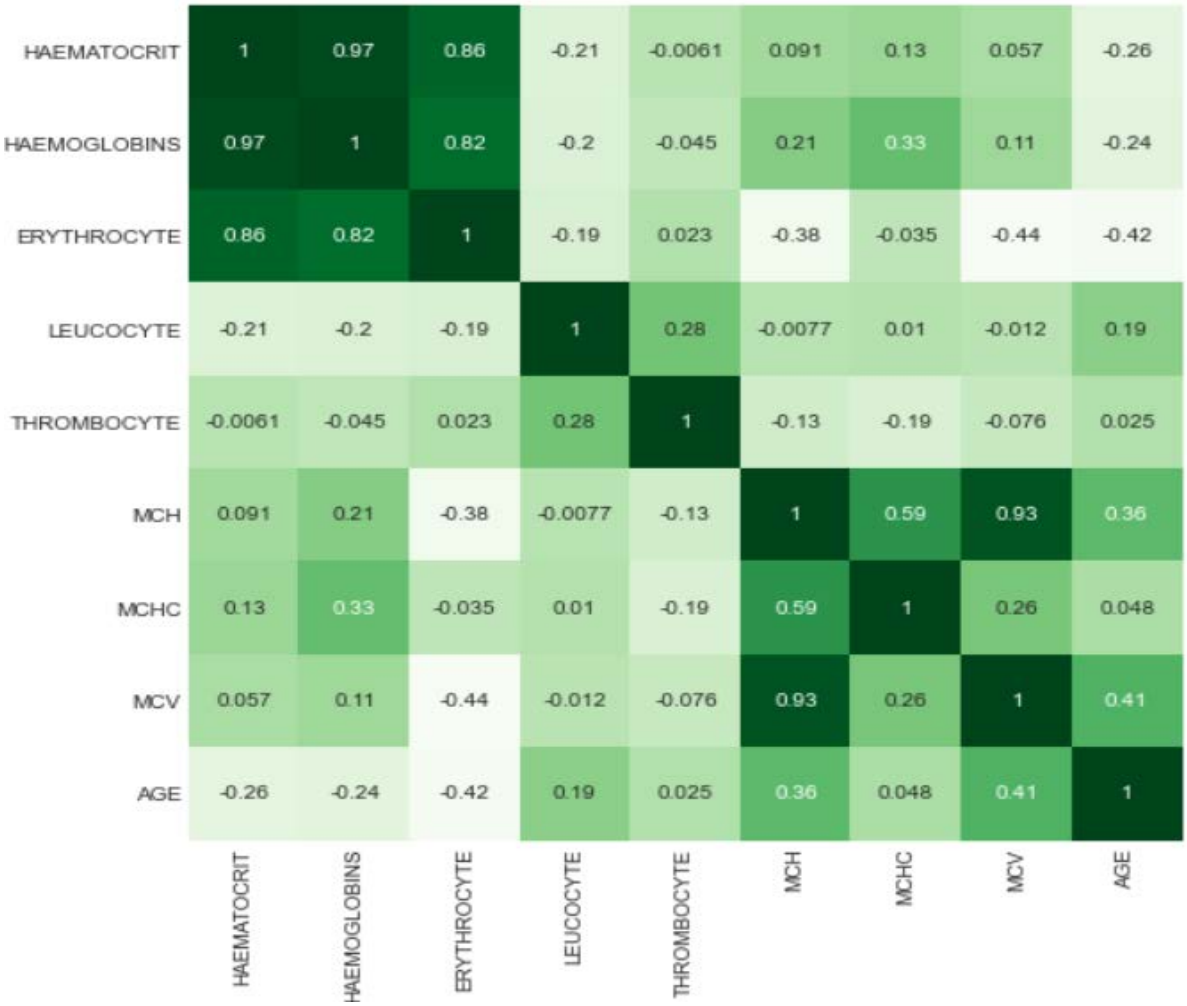


Fig. 3. Pearson's Correlation of Features with respect to each other.

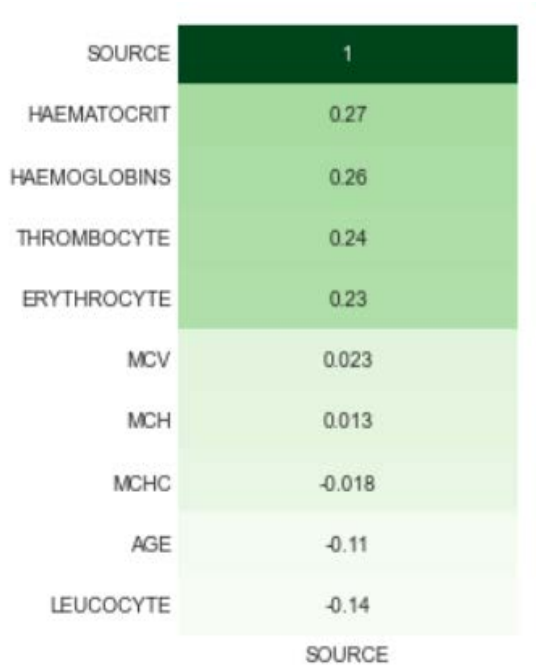


Fig. 4. Pearson's Correlation of Features with respect to Target.

TABLE I. FIRST FEW RECORDS OF THE DATASET

HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
35.1	11.8	4.65	6.3	310	25.4	33.6	75.5	1	F	1
43.5	14.8	5.39	12.7	334	27.5	34.0	80.7	1	F	1
33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	1	F	1

TABLE II. PERCENTAGE OF MISSING VALUES IN EACH FEATURE

Attribute	Missing %
HAEMATOCRIT	0.0
HAEMOGLOBINS	0.0
ERYTHROCYTE	0.0
LEUCOCYTE	0.0
THROMBOCYTE	0.0
MCH	0.0
MCHC	0.0
MCV	0.0
AGE	0.0
SEX	0.0
SOURCE	0.0

The pie chart in Fig. 1 shows that there are 52.67% Males (M) and 47.33% Females (F) in the dataset. While the pie chart in Fig. 2 shows that 40.44% people were treated as inpatient (0) and 59.56% people were treated as outpatient (1) in the dataset. Fig. 3 shows the Pearson's Correlation of features with respect to each other whereas Fig. 4 shows the Pearson's Correlation of features with respect to the target i.e. SOURCE.

In the dataset, the 80% data was taken as training data and 20% data was taken as test data. The Tables I, II and III show the first few records of the dataset, percentage of missing values in each column and number of unique values in each column respectively. During the feature engineering part, we replaced the labels of sex column with binary numbers i.e. F = 0 and M = 1. Thereafter, MinMaxScaler was used to scale the features to a range of [0, 1]. At the end, we removed the least correlated features i.e. [MCH, MCHC, MCV] from the dataset. After performing feature engineering, my dataset was ready for the next step i.e. training the model. The exploratory data analysis and feature engineering was performed in order to increase the accuracy and precision of the model. Efficient and effective feature selection techniques allow better generalization of predictive models and improved interpretability, which is a very important property for applications in health care [10]. Furthermore, Hyperparameter tuning was performed for obtaining best case scenario.

Solving issues with a decision boundary which extends outside of the space of the function that is implemented by the specified classifier model is exceedingly difficult for a single classifier to do it successfully. This non-linear boundary may be learned by combining ensemble classifiers in the right way. To avoid bad selection of a single classifier that cannot generalise performance, combine many classifiers and average their output to lower the chance of poor performance of the single classifier that is picked. As a result, the chance of making a bad choice is reduced as well [6]. Thus, we have used Stacking Classifier in order to combine the skills of the models on the regression problem to produce predictions that outperform any single model in the ensemble.

It has been demonstrated that ensemble-based systems generate better outcomes than single-expert systems for a wide range of applications and circumstances [7].

TABLE III. NUMBER OF UNIQUE VALUES IN EACH FEATURE

Attribute	Number of unique values
HAEMATOCRIT	326
HAEMOGLOBINS	128
ERYTHROCYTE	433
LEUCOCYTE	276
THROMBOCYTE	554
MCH	189
MCHC	105
MCV	406
AGE	95
SEX	2
SOURCE	2

IV. Result Analysis

After using Logistic Regression, we got the train accuracy as 70.9% and test accuracy as 71.5%. For increasing the accuracy, we performed hyperparameter tuning using RandomizedSearchCV and got best parameters for logistic regression i.e. {'penalty': 'none', 'max_iter': 300, 'fit_intercept': True, 'class_weight': {0: 1, 1: 1}, 'C': 0.01}.

After retraining the model, we got the train accuracy as 71.6% and test accuracy as 73%. The Table IV shows the classification report of the model.

Stacking is one of the most efficient methods for solving classification and regression issues. The concept of stacking is using the predictions of machine learning models from the previous level as the input variables in the following level's machine learning models [9].

And hence, we have used Stacking classifier. We got the train accuracy as 72% and test accuracy as 75%. It could be seen that there was a minor increase in the accuracy. Furthermore, I have used cross-validation but stacking classifier gave a better accuracy and so, we went for stacking classifier. The Table V shows the classification report of the model.

TABLE IV. Classification Report of the Logistic Regression Model

	Precision	**Recall**	**F1-score**	**Support**
0 (In-patient Care)	0.75	0.52	0.61	357
1 (Out-patient Care)	0.73	0.88	0.80	526
Accuracy			0.73	883
Macro Average	0.74	0.70	0.70	883
Weighted Average	0.74	0.73	0.72	883

TABLE V. Classification Report of Logistic Regression using Stacking Classifier

	Precision	**Recall**	**F1-score**	**Support**
0 (In-patient Care)	0.74	0.53	0.62	357
1 (Out-patient Care)	0.73	0.88	0.80	526
Accuracy			0.75	883
Macro Average	0.74	0.70	0.72	883
Weighted Average	0.74	0.73	0.73	883

The findings of Melhem et al. [1] reveal that out of four models i.e. Support Vector Machine (SVM) model, Decision Tree model, Random Forest model and K-Nearest Neighbors (KNN) model, Random Forest model had the best accuracy (77%), precision (72%) and sensitivity (65%). So comparing with that model, our model gave nearly a similar accuracy (75%) but gave better precision (74%). Moreover, Kirasich et al. [13] observed that logistic regression has continuously exercised a better overall precision than random forest model by increasing the variance in the explanatory factors as well as noise factors. Under varied simulated dataset circumstances, logistic regression model and random forest model had produced variable relative classification scores in all the four situations they observed.

V. Conclusion and Future Work

The results show that the logistic regression gives nearly 75% accuracy, 73% recall, 73% f1-score on the test data. It gives a decent result on the dataset. Furthermore, the main objective and idea behind the research was fulfilled.

Moreover, for logistic regression model, independent characteristics may be used to predict accurate probabilistic outcomes based on statistical analysis. The model may over-fit on the training set if the dataset has a high number of dimensions, and hence may not be able to predict correct outcomes on the test set if the dataset has a large number of dimensions. Sometimes this happens if a little amount of data is used to train the model, but the data has a large number of features. Regularization strategies should be explored for high-dimensional datasets in order to avoid over-fitting but this makes the model complex.

Using Stacking classifier, there was a slight increase in the accuracy. Moreover, using multiple machine learning algorithms with stacking classifier or using regularization techniques on logistic regression with stacking classifier on a huge dataset could be thought of.

References

[1] S. Melhem, A. Al-Aiad and M. S. Al-Ayyad, "Patient care classification using machine learning techniques," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021, pp. 57-62, doi: 10.1109/ICICS52457.2021.9464582.

[2] David Ben-Israel, W. Bradley Jacobs, Steve Casha, Stefan Lang, Won Hyung A. Ryu, Madeleine de Lotbiniere-Bassett, David W. Cadotte, The impact of machine learning on patient care: A systematic review, Artificial Intelligence in Medicine, Volume 103, 2020, 101785, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2019.101785.

[3] Beaulieu-Jones B, Finlayson SG, Chivers C, et al. Trends and Focus of Machine Learning Applications for Health Research. JAMA Netw Open. 2019;2(10):e1914051. doi:10.1001/jamanetworkopen.2019.14051.

[4] Sandro Radovanović, Boris Delibašić, Miloš Jovanović, Milan Vukićević and Milija Suknović A Framework for Integrating Domain Knowledge in Logistic Regression with Application to Hospital Readmission Prediction, International Journal on Artificial Intelligence Tools , VOL. 28, NO. 06, https://doi.org/10.1142/S0218213019600066.

[5] Mu-Yen Chen, Predicting corporate financial distress based on integration of decision tree classification and logistic regression, Expert Systems with Applications, Volume 38, Issue 9, 2011, Pages 11261-11272, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2011.02.173.

[6] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek and H. M. El-Bakry, "Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model," in IEEE Access, vol. 8, pp. 133541-133564, 2020, doi: 10.1109/ACCESS.2020.3010556.

[7] R. Polikar, "Ensemble based systems in decision making," in IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[8] Saini, Akanksha and Meitei, A J and Singh, Jitenkumar, Machine Learning in Healthcare: A Review (April 26, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3834096 or http://dx.doi.org/10.2139/ssrn.3834096.

[9] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018, pp. 255-258, doi: 10.1109/DSMP.2018.8478522.

[10] S. Radovanovic, M. Vukicevic, A. Kovacevic, G. Stiglic, and Z. Obradovic, “Domain knowledge Based Hierarchical Feature Selection for 30-Day Hospital Readmission Prediction,” Lecture Notes in Computer Science, pp. 96–100, 2015.

[11] King, Michael A. Ensemble learning techniques for structured and unstructured data. Diss. Virginia Polytechnic Institute and State University, 2015.

[12] Justin Oh, Anahi Perlas, Johnny Lau, Rajiv Gandhi, Vincent W.S. Chan, Functional outcome and cost-effectiveness of outpatient vs inpatient care for complex hind-foot and ankle surgery. A retrospective cohort study, Journal of Clinical Anesthesia, Volume 35, 2016, Pages 20-25, ISSN 0952-8180, https://doi.org/10.1016/j.jclinane.2016.07.014.

[13] Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: Vol. 1 : No. 3, Article 9.

[14] Maroof D.A. (2012) Binary Logistic Regression. In: Statistical Methods in Neuropsychology. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3417-7_8.

[15] Jafari Goldarag, Y., Mohammadzadeh, A. & Ardakani, A.S. Fire Risk Assessment Using Neural Network and Logistic Regression. J Indian Soc Remote Sens 44, 885–894 (2016). https://doi.org/10.1007/s12524-016-0557-6.

[16] Margaret Sullivan Pepe, Mary Lou Thompson, Combining diagnostic test results to increase accuracy , Biostatistics, Volume 1, Issue 2, June 2000, Pages 123–140, https://doi.org/10.1093/biostatistics/1.2.123.

[17] Sadikin, Mujiono (2020), “EHR Dataset for Patient Treatment Classification”, Mendeley Data, V1, doi: 10.17632/7kv3rctx7m.1.

[18] P.P. España, A. Capelastegui, J.M. Quintana, A. Soto, I. Gorordo, M. García-Urbaneja, A. Bilbao, European Respiratory Journal 2003 21: 695-701; DOI: 10.1183/09031936.03.00057302.

[19] Blais, M.A., Matthews, J., Lipkis-Orlando, R. et al. Predicting Length of Stay on an Acute Care Medical Psychiatric Inpatient Service. Adm Policy Ment Health 31, 15–29 (2003). https://doi.org/10.1023/A:1026044106172.

[20] Cuffel, Brian J., Martin Held, and William Goldman. "Predictive models and the effectiveness of strategies for improving outpatient follow-up under managed care." Psychiatric Services 53.11 (2002): 1438-1443.

Vision based 3D Gesture Tracking using Augmented Reality and Virtual Reality for Improved Learning Applications

Zainal Rasyid Mahayuddin[1]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM, Bangi, Selangor, Malaysia

A F M Saifuddin Saif[2]
Faculty of Science
University of Helsinki, Finland

***Abstract*—3D gesture recognition and tracking based augmented reality and virtual reality have become a big interest of research because of advanced technology in smartphones. By interacting with 3D objects in augmented reality and virtual reality, users get better understanding of the subject matter where there have been requirements of customized hardware support and overall experimental performance needs to be satisfactory. This research investigates currently various vision based 3D gestural architectures for augmented reality and virtual reality. The core goal of this research is to present analysis on methods, frameworks followed by experimental performance on recognition and tracking of hand gestures and interaction with virtual objects in smartphones. This research categorized experimental evaluation for existing methods in three categories, i.e. hardware requirement, documentation before actual experiment and datasets. These categories are expected to ensure robust validation for practical usage of 3D gesture tracking based on augmented reality and virtual reality. Hardware set up includes types of gloves, fingerprint and types of sensors. Documentation includes classroom setup manuals, questionaries, recordings for improvement and stress test application. Last part of experimental section includes usage of various datasets by existing research. The overall comprehensive illustration of various methods, frameworks and experimental aspects can significantly contribute to 3D gesture recognition and tracking based augmented reality and virtual reality.**

Keywords*—Augmented reality; virtual reality; 3D gesture tracking

I. Introduction

Dexterity is one of the most important driving forces of human intelligence. Our hands have enabled us to create arts and crafts as well as build massive constructions alike. In this age of technology, human use their hands to interact with electronic devices everyday. Touch and gesture input have become part of common life. Nowadays, people interact with digital devices by using touch screen displays or trackpads. Two dimensional (2D) touch screens are basically offered by the latest technology. However, mobile phone touch screens constrain us within a small space of the device. In this context, virtual reality refers convincing visual rendering of the simulated objects in lieu with manipulating them in a fast, precise, and natural way [1], augmented reality indicates machine generated image on user's view of the real world to enable composite view [56] and hand gesture refers movement of the hand to enable meaningful interpretation [60]. To extend the interaction space from 2D surface to real 3-dimensional (3D) space, this research investigates currently existing vision based 3D gestural architectures for augmented reality and virtual reality. The core goal of this research is to present comprehensive investigation on tracking and predicting hand gestures from RGB camera images and interact with virtual objects in mobile phones. Huge investments of the big technology companies on augmented reality and virtual reality has broadened the applications of this case where a relatively new branch has been introduced towards the direction of gesture tracking and recognition in the domain of computer vision and pattern recognition.

Tracking the hands of an user makes it a difficult task when the tracking camera is also moving, which is the case for augmented reality and virtual reality. While there have been numerous previous research works for tracking hands, the problem still remains as a research interest. Because of the progress in the area of computer vision and machine learning, computers are now capable of tracking hand gestures and pose through various techniques. This research analyzed those previous works to track hand gestures and interact with objects with hands in augmented reality and virtual reality. In the early days of this field, there were numerous approaches made with traditional image processing techniques to detect and track hands. But in the recent days, most of the research has been performed with respect to machine learning approaches. This research presents the benefits and shortcomings of these approaches in lieu with presenting existing experimental analysis with different approaches.

II. Core Background Study

Augmented reality and virtual reality have been a research interest for a long time. Researchers have been trying to do camera tracking based augmented reality and virtual reality for a long time. Recently these two domains have become a big interest of research because of advanced technology in mobile phones and smartphones. In recent years, there have been a lot of progress in the processing capability of smart phones. Because of this, handheld mobile phones can compute the necessary amount of data to perform camera tracking as well as rendering on corresponding display. In recent years, there has been a lot of progress in recognition and tracking of the

human face and we have seen a rise in user interest in such kinds of applications.

This research also relies on the research that has been done to track plane surfaces as well as wall or vertical objects and render digital objects on them to deliver the experience of augmented reality and virtual reality. There has been enormous research on this area and there have been massive breakthroughs in the area [56][60]. There have been games based on augmented reality and virtual reality where a planar surface such as floor or road is tracked and on the tracked surface a game object is drawn and players use their camera feed to see the object augmented on the video feed of that planar surface. There have also been a lot of applications to visualize shopping products in augmented reality and virtual reality. Big technology companies like Google and Apple have their own libraries to implement augmented reality and virtual reality. Furniture vendors such as IKEA has launched their app in which users are able to place a virtual 3d object in their living room and judge which product will look good in their room as well as the environment. These usages and demands of such applications have turned lots of active research interest in the area of augmented reality and virtual reality.

In the past, there has been research on recognizing human hands and tracking them. In the early days of the research, recognizing the hand and tracking the hand in an image was done by offline image processing as this was a computationally heavy task. After some successful research, it has also been possible to track human hands in real time using a web camera and a desktop pc as technology progressed. Now with the advancement in neural networks, feature tracking and modeling the network has been done by neural networks, and because of the easiness of modeling a human hand as an object for tracking has been easier. Now, there has been active ongoing research for tracking human hands via mobile phone camera as well as recognizing the pose of the hands in a camera frame. Research interest of this proposed investigation relies within the research of tracking and recognizing human hand pose and tracking it in continuous frames.

III. Qualitative Analysis on Methods

Early researchers used augmented reality and interaction with personal interaction panels which includes creating menu system and 3D interface [17] [39]. The Positive side of these researches were that there was instant feedback from users. However, in accuracy and unreliability were the main area of concern in these researches. Collaborative geometry learning based object constructions shown in Fig. 1 is an active area for the researchers which includes augmented reality by hybrid hardware and software setup [18][47]. However, requirement of comprehensive evaluation of the practical value such as development of substantial educational content demands for further investigation for real use in the classroom for teachers or students. Hand gesture recognition and AR marker recognition were designed previously for geometry learning [19][46]. However, applications for learning 3D geometry by these researches could also be expanded for multi-device environments. Observation on classes and assignments for the students by letting them work with augmented reality is an interesting broad domain which mostly depends on educational design research [20] [43] [44] [46] [48]. However, existing curriculums in the schools at junior level, more passiveness, less constructive contents in the context of teaching media used by researchers, limited understanding of the roles by teachers, lack of ability to create interesting teaching contents demand of extensive investigation in this area of research. Marker based positional tracking as well as picture based positional tracking is another common method among previous researchers due to easiness of implementation for these systems [21][22]. However, due to the limited validation inside classroom environments instead of real world scenarios, these researches could not provide expected satisfactory results. Some augmented reality researchers took gamification a step further in the gaming industry [23]. However, gamification hugely depends on users and participants feedback due to new technology.

For virtual reality based education systems are more reliable to design and construct geometries [24]. Unity3D engine and Vuforia plugin made the tasks easier using camera tracking [25] although previously tested to a small group of students demands further investigation. QR code based tracking method was an old method in this context still attracts researchers since these are quick and easy to implement [26]. However, with the aspects of language, contents, and design QR based tracking requires further validation. Surface tracking, geometry modification and structure from motion pipeline (SFM) has been considered as another potential subsets of augmented reality and virtual reality [27] [28] [29] [30]. Among these methods, structure from motion has been holding higher research interest among researchers. However, due to recent advanced computing systems and user friendly process to construct and visualize point clouds make these research challenging for practical usage although there have been significant improvements in visual quality which still needs deep focus in terms with reducing computational complexity [28]. Perspective geometry [29] is another option by the researchers which has been suffering significantly from complexity and dependency on external hardware. In addition, there are significant errors observed in these kinds of systems and video rendering through augmentation takes a lot of computational cost [29]. Planar surface tracking with the inclusion of background and foreground subtraction is one of the successful method in the area of surface tracking [30]. However, planar surface tracking method did not attain good frame rate although later occlusion management has been improved significantly by the recent advancement in machine learning technology. In the design industry, AR and VR has the most usage observed in the previous research. Some degree of success were found for spatial skill learning where specialized applications were designed and developed in order to put them into training purposes [31]. Existing technology as in research by [32] and [41] were used to build 3D models for visualizing augmented reality which was developed for improving students understanding capability. However, users for these applications requires some degree of geometry and 3D modeling skills. User experience on the existing applications is also another area for improvement observed by the researchers during implementation of gesture control [33].

In this context, hand pose estimation and shape recognition from single hand image or video stream were focused most by the researchers of such gestures and has many applications in augmented reality and virtual reality. Convolutional neural network and 3D hand meshing using LBS were used to recognize hand shape and pose estimation from depth images [34] and initially from RGB images. As part of hybrid approach, hand pose estimation was done from RGB images by adopting multiview RGB images and depth data [49]. However, reliable size of the dataset in lieu multidimensionality, hand to objects and hand to hand interactions were not considered in these researches in order to make the hand shape and pose estimation more robust. Deep learning [40] and monocular RGB cameras advances estimation of hand shape from RGB images [34]. In terms with datasets, researchers put effort to make hand gesture datasets for acceptable validation, however, models trained on one dataset did not perform well on other datasets due to lack of generalization in the training data [35] because researchers found significantly improved results in indoor and outdoor by exploring validation in generalized datasets. In this context, vision based Deep Convolutional Neural Networks (CNNs) attributes the success for hand pose estimation [36] [54] using depth cameras, depth map data with the analysis of effectiveness based on detection based method versus regression based methods. From different viewpoint perspectives, neural network was deployed to model human hands and pose in virtual reality, i.e. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking [38]. In MEgATrack, depth based approaches was used and generated training data from model based tracking system. Research by [38] used DetNet to track hands KeyNet to predict 21 key-points in hand from the cropped image based on the bounding box provided by DetNet in the previous step. However, hand scale and distance recognition were their main challenge to achieve accuracy. MediaPipe pipeline shows prominent progress for handheld mobile phone based hand tracking and hand gestures recognition [37]. Research by [37] used single RGB camera consists of palm detector to provide bounding box of the hand and hand landmark model to predict hand skeleton. However, for multidimensional data, performance of MediaPipe is still not resolved.

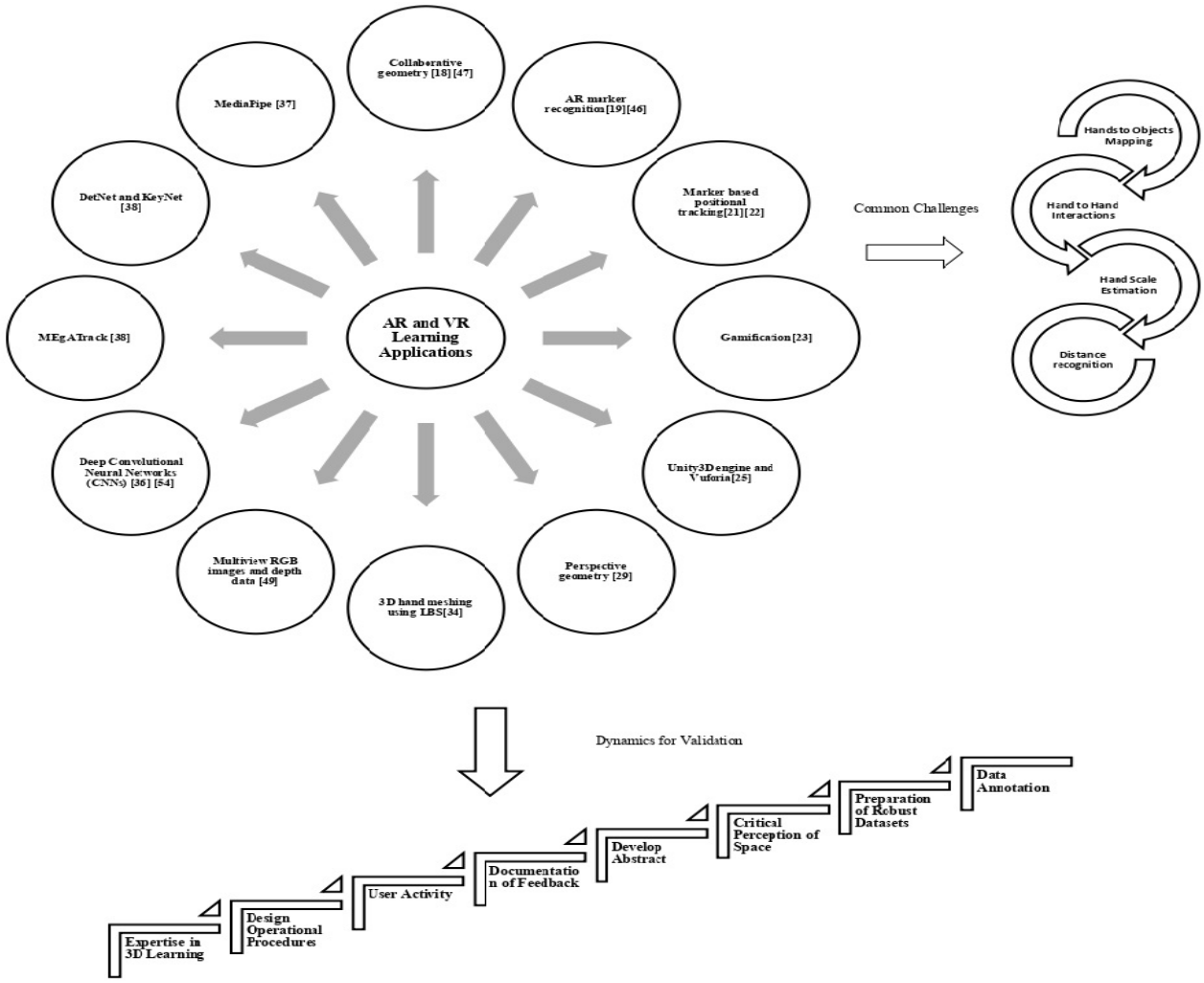


Fig. 1. Methods and Challenges for Augmented and Virtual Reality based Learning Applications.

IV. Step Wise Frameworks Analysis

Majority of the research done in the earlier days of augmented reality and virtual reality consisted of using specialized hardware. Frameworks using Construct3D involves creating user interface and then tracking user hand movement and giving input using pen-like tools [17]. Other similar methods also consist of similar frameworks. Researcher's implemented 3D layer based interaction systems [18]. These methods also require extensive hardware setups and screen projection. For Marker based augmented reality tracking, initial step is to acquire the markers and recognize them and track them in the scene. When combined with software design methodologies, applications become easier to develop and get feedback from users and participants. In these cases, first step is to choose the target participants, users or students, then making prototypes and improving it further [20] [21] [23] [25] [45]. Testing into small groups first and then to a broader range of participants also brings out better results for research [26].

Tracking surfaces for augmented reality was adopted for similar approaches by the second group of researchers. For 3D reconstruction, steps are to acquire a large number of images of the target object and then reconstruct it using different structures-from-motion pipelines [27] followed by rending these objects in 3D and shown on top of video camera feed [28]. In this context, most of the pipelines use offline processing step followed by real time camera tracking and rendering. In the same focus, foreground and background separation techniques provide better results for occlusion management [30]. Applying virtual reality and augmented reality for practical purposes have seen significant increase in recent times. By interacting with 3D objects in virtual reality, users get a better understanding of the subject matter. The primary requirement for these projects is expertise in 3D geometry modeling and then designing operational procedures for users or participants and observing user activity and documentation of feedback [32] [33] [47]. Investigating students understanding of the topic before going into development of the project is suggested by researchers, since that allows researchers to get to know their participants perspective better. Then allowing students to learn and practice by themselves is another good way to get feedback [31]. Researchers have also tried developing abstract and critical perception of space before putting students to use the applications [32].

Preparation of robust datasets and annotation for hand tracking were core research focus for one subset of researchers which not an easy tasks where most researchers prefer training models on synthesized hand gesture data. Straightforward solutions like human pose tracking and applying them to track hand and estimate pose were preferred by some researchers [34] where usage of multiview geometry was a popular approach to attain hand images [35] [41] required heavy hardware set up and accurate annotation manually [50]. Single shot detection is a preferred for image based tracking [58] whereas multicamera tracking in virtual reality is another favorable approach where computational complexity was a big concern.

V. Experimental Analysis

Several works were done in the past for 3D gesture tracking and recognition. A robust gesture detection system by using a single camera is a challenging issue in the area of computer vision [58]. The camera quality is also a challenging part in this context. Most of the researchers that are based on augmented reality uses marked gloves shown for accurate and reliable fingertip tracking [1] [2]. Few methods depend on the object segmentation for the shape or temperature [3] [4] [5]. In most devices, thermal based approaches and expensive infrared cameras are needed but not given in the previous research work. Most of the gesture tracking devices like Kinect are based on depth sensors. These gesture tracking devices are only available for stationary systems because of size and power limitations [6]. In few of the systems, color-based techniques were used but color-based techniques are sensitive for the lights and degrade the quality of gesture recognition and tracking process. Template matching and contour-based techniques suits well for these types of specific hand gesture recognition and tracking [7]. Few systems were designed based on the syntactic analysis of hand gestures by using syntactic pattern recognition paradigm [8]. Few approaches that were designed for smartphones and tablets use accelerometer-based methods with the device's acceleration sensor [9] [10]. For detecting the fingertips, in some gesture based interaction, visual color markers are used [11]. Several researches were proposed for recognition and tracking system of hand gesture which was based on low level edge orientation features and can be implemented by using the hierarchical scoring of the similarity between the query and database images. In these researches, fingertips and all the hand joints that consist of the finger joints are marked from the database. Then, overall system saves the exact position of the marked points with the help of the image coordinates and finds out the relation between the joints in the form [12]. Some researches were conducted to track 3D photos of the human body by using sensor-fusion algorithms [13] [14]. A sensor-fusion method that can track the articulation of the hand in the presence of excessive motion blur was proposed using HPF framework [15]. Gyroscopes are very popularly used when it comes to human body pose estimation but the investigation for the use of gyroscope for hand pose estimation is not completed yet. IMU sensor was used to assist model-based tracking to get more robust performance [16].

Research that was performed in the education domain requires multiple accounts of user questionnaire and feedback. Researchers let their participants use the application and later ask questions to acquire feedback from participants [17] [18] [19]. Researchers made attempts by giving primary knowledge about the subject before exposing them to the real application and later tested again their knowledge level to measure the improvement after using their applications [20] [43]. Extensive data documentation was required for these experiments. In classroom setups, researchers observed students using their application and later asking questions to get feedback. Making a prototype before the final application testing was used by some researchers [23] [45]. In every case, user questionnaires and recording student improvement in learning is the mandatory step for these researches.

Researches also included expert assessment to validate their research [26].

Surface tracking was experimented by researchers for a long time. Researchers experimented with different methods and choose suitable approaches for their research where it is common to experiment with multiple methods [27] [29]. In addition, it is also emphasized to stress test the application under different kinds of motions [39]. A subset of researchers utilizes a preprocessing stage for achieving better performance [30].

In recent years, augmented and virtual reality have been a subject of major experimentation to explore possibilities of different use cases. Researchers investigated students understanding before and after using the applications. Also, letting the participants find how to use the application is one more approach that returns good results [31]. One more investigation was done that lets users acquire little knowledge about a topic beforehand and then let them, use the application and quiz on how much they have improved [32]. Putting augmented reality as well as virtual reality into practical training purposes have brought out better results among technicians. Researchers verified their research by evaluating the participants knowledge and skill after each subsequent phase [33].

Usage of multiple datasets is common for researchers using different parameters and including or excluding different levels of details while annotating the data [34,51-53,59-63]. or FreiHAND, researchers performed cross-dataset generalization to achieve improved results [35]. Evaluating different levels of details and annotation was performed to optimize research effort. Numerous datasets are currently available for hand gesture recognition training [36, 55-57, 64-68] mentioned in Fig. 2. Availability of datasets is a major reason for the increase of research interest for augmented reality and virtual reality domain. For Google MediaPipe, researchers created their own annotated dataset of hand images in the wild, in-house hand images and synthetic hand gesture dataset [37]. By utilizing different datasets for their specific purposes, MediaPipe achieves greater efficiency in terms of performance.

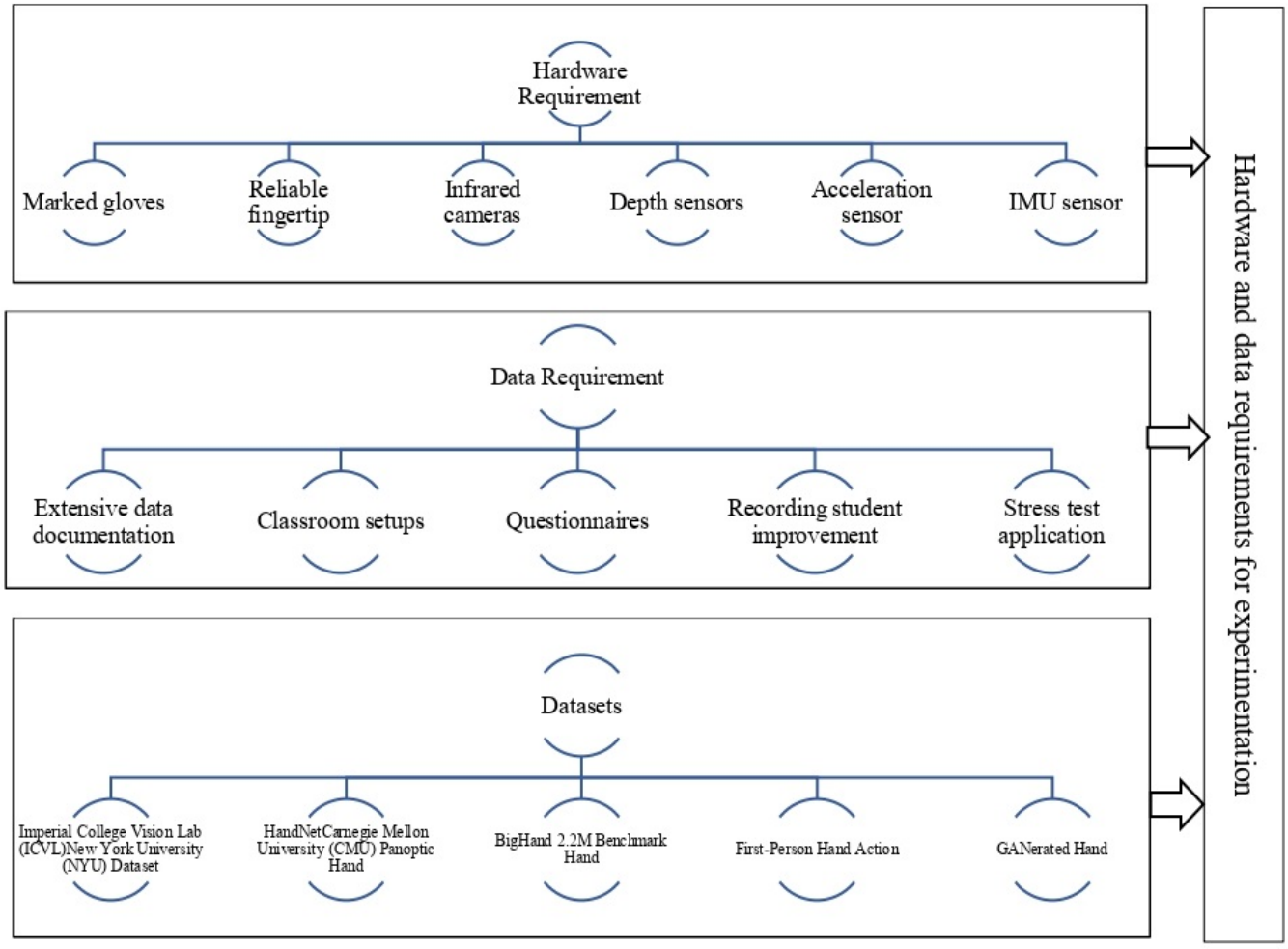


Fig. 2. Experimental Requirements.

VI. Observation and Discussion

This research observes that in recent years there has been a significant amount of research was done in computer vision in the context of 3D gesture tracking based on augmented reality and virtual reality. With the advancement of computer processing power, it is possible to compute and find solution quickly and more accurately. In the earlier days of research on augmented reality and virtual reality there have been requirements of customized hardware support and even so the performance and quality was not satisfactory. At a later stage there has been significant research on tracking planar surfaces as well as tracking objects. By tracking a surface camera position, efficient detection was proposed. On top of those technology, occlusion handling was performed. With greater computer processing power tracking have been more stable in recent times. While there were requirements for heavy hardwares and sensors, after a decade, now it can be performed on a handheld mobile device.

By introducing augmented reality and virtual reality into media, their popularity and research interest have been increased. Comprehensive investigation by this research has observed significant amounts of research as well as practical usage of augmented reality and virtual reality. In the early stage of research for these domains, there was requirement for specialized input pens as well as specialized input panels for interacting with the system. Naturally humans are used to use hands for performing day to day tasks, for this reason tracking hands has been a great interest in recent times. While the technology is not perfect, yet there has been a significant amount of improvement and technological advancement in the past few years.

This research also observed that big technology companies like Microsoft have been working with HoloLens technology which is a high-end mixed reality platform [42]. Google and Apple have their own augmented reality platforms named ARCore and ARKit respectively and there has been a rise in augmented reality applications ever since. Facebook has shown interest in Virtual Reality and with Oculus Virtual Reality systems they have been outperforming themselves every year.

While there have been significant hardware improvements, there has not been improvement in user experience as per with time. Human computer interaction research domain has been working on improving user experience for a few decades now a days and result is improving day by day. There have been touch input display for mobile phones and gesture tracked hand controller for gaming console systems. However, it is high time to investigate more on making the interactions more meaningful for human beings by making interaction between humans and computer more natural.

VII. Conclusion

Hand gesture detection based on augmented reality and virtual reality is an active and ongoing research field which are attracting a lot of research towards the topic. The vastness of both topics makes it interesting to pursue research problems further. In this research, investigation of different usages and implementations of augmented reality and virtual reality based systems was elaborated and discussed in detail. Besides, possibility of neural network based hand palm tracking and hand gesture tracking was illustrated comprehensively. This research found that hand interaction in augmented reality and virtual reality can be achieved with acceptable accuracy based on improved user experience. In addition, this research also emphasizes to focus on usages of augmented reality and virtual reality, tracking surfaces as well as tracking and 3D reconstruction of real life objects in the context of hand palm detection, hand tracking and detecting symbolic gestures from finger shapes. With the keypoint from hand landmark points, 3D mesh can be rendered and that 3D mesh can also be used to interact with augmented and virtual objects in future.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the “Geran Universiti Penyelidikan” research grant, GUP-2020-064.

References

[1] K. Dorfmuller-Ulhaas and D. Schmalstieg, "Finger tracking for interaction in augmented environments," *Proceedings IEEE and ACM international symposium on augmented reality*, pp. 55-64, 2001.

[2] C. Maggioni, "A novel gestural input device for virtual reality," *Proceedings of IEEE Virtual Reality Annual International Symposium*, pp. 118-124, 1993.

[3] C. Von Hardenberg and F. Bérard, "Bare-hand human-computer interaction," *Proceedings of the 2001 workshop on Perceptive user interfaces*, pp. 1-8, 2001.

[4] D. Iwai and K. Sato, "Heat sensation in image creation with thermal vision," Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology, pp. 213-216, 2005.

[5] M. Kolsch and M. Turk, "Fast 2d hand tracking with flocks of features and multi-cue integration," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 158-158, 2004.

[6] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, "Robust hand gesture recognition with kinect sensor," *Proceedings of the 19th ACM international conference on Multimedia*, pp. 759-760, 2011.

[7] H. Zhou and Q. Ruan, "Finger contour tracking based on model," IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOM'02. Proceedings., vol. 1, pp. 503-506, 2002.

[8] M. Flasiński and S. Myśliński, "On the use of graph parsing for recognition of isolated hand postures of Polish Sign Language," *Pattern Recognition*, vol. 43, no. 6, pp. 2249-2264, 2010.

[9] F. Arce and J. M. G. Valdez, "Accelerometer-based hand gesture recognition using artificial neural networks," *Soft Computing for Intelligent Control and Mobile Robotics*: Springer, pp. 67-77, 2010.

[10] J. Choi, K. Song, and S. Lee, "Enabling a gesture-based numeric input on mobile phones," *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 151-152, 2011.

[11] W. Hürst and C. Van Wezel, "Gesture-based interaction via finger tracking for mobile augmented reality," *Multimedia Tools and Applications*, vol. 62, no. 1, pp. 233-258, 2013.

[12] S. Yousefi and H. Li, "3D hand gesture analysis through a real-time gesture search engine," *International Journal of Advanced Robotic Systems*, vol. 12, no. 6, p. 67, 2015.

[13] A. Gilbert, M. Trumble, C. Malleson, A. Hilton, and J. Collomosse, "Fusing visual and inertial sensors with semantics for 3d human pose estimation," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 381-397, 2019.

[14] T. Helten, M. Muller, H.-P. Seidel, and C. Theobalt, "Real-time body tracking with one depth camera and inertial sensors," in *Proceedings of the IEEE international conference on computer vision*, pp. 1105-1112, 2013.

[15] G. Park, A. Argyros, J. Lee, and W. Woo, "3d hand tracking in the presence of excessive motion blur," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 5, pp. 1891-1901, 2020.

[16] H.-i. Kim and W. Woo, "Smartwatch-assisted robust 6-DOF hand tracker for object manipulation in HMD-based augmented reality," *IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 251-252, 2016.

[17] H. Kaufmann, D. Schmalstieg, and M. Wagner, "Construct3D: a virtual reality application for mathematics and geometry education," *Education and information technologies*, vol. 5, no. 4, pp. 263-276, 2000.

[18] H. Kaufmann and D. Schmalstieg, "Mathematics and geometry education with collaborative augmented reality," *ACM SIGGRAPH 2002 conference abstracts and applications*, pp. 37-41, 2002.

[19] H.-Q. Le and J.-I. Kim, "An augmented reality application with hand gestures for learning 3D geometry," *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 34-41, 2017.

[20] R. Johar, "A need analysis for the development of augmented reality based-geometry teaching instruments in junior high schools," *Journal of Physics: Conference Series*, vol. 1460, no. 1: IOP Publishing, p. 012034, 2020.

[21] B. Cahyono, M. B. Firdaus, E. Budiman, and M. Wati, "Augmented reality applied to geometry education," *2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pp.299-303, 2018.

[22] R. Fernández-Enríquez and L. Delgado-Martín, "Augmented Reality as a Didactic Resource for Teaching Mathematics," *Applied Sciences*, vol. 10, no. 7, p. 2560, 2020.

[23] I. Radu, E. Doherty, K. DiQuollo, B. McCarthy, and M. Tiu, "Cyberchase shape quest: pushing geometry education boundaries with augmented reality," *Proceedings of the 14th international conference on interaction design and children*, pp. 430-433, 2015.

[24] H. Kaufmann and D. Schmalstieg, "Designing immersive virtual reality for geometry education," *IEEE Virtual Reality Conference (VR 2006)*, pp. 51-58, 2006.

[25] M. Nazar *et al.*, "Development of Augmented Reality application for learning the concept of molecular geometry," *Journal of Physics: Conference Series*, vol. 1460, no. 1: IOP Publishing, p. 012083, 2020.

[26] R. Auliya and M. Munasiah, "Mathematics learning instrument using augmented reality for learning 3D geometry," in *Journal of Physics:Conference Series*, vol. 1318, no. 1: IOP Publishing, p. 012069, 2019.

[27] Y. Nakashima, Y. Uno, N. Kawai, T. Sato, and N. Yokoya, "AR image generation using view-dependent geometry modification and texture mapping," *Virtual Reality*, vol. 19, no. 2, pp. 83-94, 2015.

[28] A. Samini and K. L. Palmerius, "A perspective geometry approach to user-perspective rendering in hand-held video see-through augmented reality," in *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, pp. 207-208, 2014.

[29] Y. Nakashima, T. Sato, Y. Uno, N. Yokoya, and N. Kawai, "Augmented reality image generation with virtualized real objects using view-dependent texture and geometry," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1-6, 2013.

[30] H. L. Wang, K. Sengupta, P. Kumar, and R. Sharma, "Occlusion handling in augmented reality using background-foreground segmentation and projective geometry," *Presence*, vol. 14, no. 3, pp. 264-277, 2005.

[31] Y. Tang, K. Au, and Y. Leung, "Comprehending products with mixed reality: Geometric relationships and creativity," *International Journal of Engineering Business Management*, vol. 10, p. 1847979018809599, 2018.

[32] N. A. A. González, "How to include augmented reality in descriptive geometry teaching," *Procedia Computer Science*, vol. 75, pp. 250-256, 2015.

[33] I.-J. Lee, T.-C. Hsu, T.-L. Chen, and M.-C. Zheng, "The Application of AR Technology to Spatial Skills Learning in Carpentry Training," *International Journal of Information and Education Technology*, vol. 9, no. 1, 2019.

[34] L. Ge *et al.*, "3d hand shape and pose estimation from a single rgb image," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10833-10842, 2019.

[35] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 813-822, 2019.

[36] B. Doosti, "Hand pose estimation: A survey," *arXiv preprint arXiv:1903.01013*, 2019.

[37] F. Zhang *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[38] S. Han *et al.*, "MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 87: 1-87: 13, 2020.

[39] Y. Chang, "sur. faced. io: augmented reality content creation for your face and beyond by drawing on paper," *ACM SIGGRAPH 2019 Appy Hour*, pp. 1-2, 2019.

[40] T. Huang and Y. Liu, "3d point cloud geometry compression on deep learning," *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 890-898, 2019.

[41] Z. Li, Y. Wang, J. Guo, L.-F. Cheong, and S. Z. Zhou, "Diminished reality using appearance and 3D geometry of internet photo collections," *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 11-19, 2013.

[42] A. Sherstyuk, A. Treskunov, and B. Berg, "Fast geometry acquisition for mixed reality applications using motion tracking," *7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp.179-180, 2008.

[43] S. Sudirman, R. Yaniawati, M. Melawaty, and R. Indrawan, "Integrating ethnomathematics into augmented reality technology: exploration, design, and implementation in geometry learning," *Journal of Physics: Conference Series*, vol. 1521, no. 3: IOP Publishing, p. 032006, 2020.

[44] R. Andrea, F. Agus, and R. Ramadiani, " Magic Boosed" an elementary school geometry textbook with marker-based augmented reality," 2019.

[45] A. Buchori, P. Setyosari, I. W. Dasna, and S. Ulfa, "Mobile augmented reality media design with waterfall model for learning geometry in college," *Int. J. Appl. Eng. Res*, vol. 12, no. 13, pp. 3773-3780, 2017.

[46] M. Flores-Bascuñana, P. D. Diago, R. Villena-Taranilla, and D. F. Yáñez, "On Augmented Reality for the learning of 3D-geometric contents: A preliminary exploratory study with 6-Grade primary students," *Education Sciences*, vol. 10, no. 1, p. 4, 2020.

[47] K. Olalde, B. García, and A. Seco, "The importance of geometry combined with new techniques for augmented reality," *Procedia Computer Science*, vol. 25, pp. 136-143, 2013.

[48] D. Rohendi, S. Septian, and H. Sutarno, "The use of geometry learning media based on augmented reality for junior high school students," *IOP conference series: Materials science and engineering*, vol. 306, no. 1: IOP Publishing, p. 012029, 2018.

[49] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1106-1113, 2014.

[50] S. Gattupalli, A. R. Babu, J. R. Brady, F. Makedon, and V. Athitsos, "Towards deep learning based hand keypoints detection for rapid sequential movements from rgb images," *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, pp. 31-37, 2018.

[51] A. S. Saif and Z. R. Mahayuddin, "Robust Drowsiness Detection for Vehicle Driver using Deep Convolutional Neural Network," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020.

[52] Z. R. Mahayuddin and A. S. Saif, "A COMPARATIVE STUDY OF THREE CORNER FEATURE BASED MOVING OBJECT DETECTION USING AERIAL IMAGES," *Malaysian Journal of Computer Science*, pp. 25-33, 2019.

[53] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Moment feature based fast feature extraction algorithm for moving object detection using aerial images," *PloS one*, vol. 10, no. 6, p. e0126212, 2015.

[54] M. Yasen and S. Jusoh, "A systematic review on hand gesture recognition techniques, challenges and applications," *PeerJ Computer Science*, vol. 5, p. e218, 2019.

[55] A. S. Saif and Z. R. Mahayuddin, "Moment Features based Violence Action Detection using Optical Flow," *Moment*, vol. 11, no. 11, 2020.

[56] Z. R. Mahayuddin and A. Saif, "Augmented Reality Based Ar Alphabets Towards Improved Learning Process In Primary Education System," *Journal of Critical Reviews*, vol. 7, no. 19, pp. 514-521, 2020.

[57] Z. R. Mahayuddin and A. Saif, "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality," *TEST Engineering and Management*, vol. 83, pp. 4020-4027, 2020.

[58] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, pp. 4903-4911, 2017.

[59] Z. R. Mahayuddin, H. Arshad, and C. H. C. Haron, "Pengintegrasian VRML dengan Java dalam simulasi sistem masa nyata proses kisar hujung maya," *Sains Malaysiana*, vol. 38, 2009.

[60] Z. R. Mahayuddin and A. Saif, "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality," *TEST Engineering and Management*, vol. 83, pp. 4020-4027, 2020.

[61] Z. R. Mahayuddin and A. Saif, "Augmented Reality Based AR Alphabets Towards Improved Learning Process In Primary Education System," *Journal of Critical Reviews*, vol. 7, no. 19, pp. 514-521, 2020.

[62] Z. R. Mahayuddin and A. Saif, "A Comprehensive Review towards Segmentation and Detection of Cancer Cell and Tumor for Dynamic 3D Reconstruction", *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 9, no. 1, pp. 28 – 39, 2020.

[63] Z. R. Mahayuddin and N. Mamat, "Implementing augmented reality (AR) on phonics-based literacy among children with autism," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 6, pp. 2176-2181, 2019.

[64] Z. R. Mahayuddin, N. A. Suwadi, R. Jenal, and H. Arshad, "T. Implementing smart mobile application to achieve a sustainable campus," *International Journal of Supply Chain Management*, vol. 7, no. 3, pp. 154-159, 2018.

[65] H. Rahman, H. Arshad, R. Mahmud, Z. R. Mahayuddin, and W. K. Obeidy, "A Framework to Visualize 3D Breast Tumor Using X-Ray Vision Technique in Mobile Augmented Reality," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 2-11, pp. 145-149, 2017.

[66] Z. R. Mahayuddin, H. M. Jais, and H. Arshad, "Comparison of human pilot (remote) control systems in multirotor unmanned aerial vehicle navigation," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 1, pp. 132-138, 2017.

[67] Z. R. Mahayuddin and N. A. Khairuddin, "Rapid Simulation Model Building in Cellular Manufacturing using Cladistics Technique," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 2, pp. 489-495.

[68] H. Arshad, Z. R. Mahayuddin, C. H. C. Haron, and R. Hassan, "Flank wear simulation of a virtual end milling process," *European Journal of Scientific Research*, vol. 24, no. 1, pp. 148-156, 2008.

A Framework for Secure Healthcare Data Management using Blockchain Technology

Ahmed I. Taloba, Alanazi Rayan, Ahmed Elhadad, Amr Abozeid, Osama R. Shahin, Rasha M. Abd El-Aziz
Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia

***Abstract*—In the current era of smart cities and smart homes, the patient's data like name, personal details and disease description are highly insecure and violated most often. These details are stored digitally in a network called Electronic Health Record (EHR). The EHR can be useful for future medical researches to enhance patients' healthcare and the performance of clinical practices. These data cannot be accessible for the patients and their caretakers, but they are readily available for unauthorized external agencies and are easily breached by hackers. This creates an imbalance in data accessibility and security. This can be resolved by using blockchain technology. The blockchain creates an immutable ledger and makes the transaction to be decentralized. The blockchain has three key features namely Security, Transparency, and Decentralization. These key features make the system to be highly secured, prevent data manipulation, and can only be accessible by authorized persons. In this paper, a blockchain-based security framework has been proposed to secure the EHR and provide a safe way of accessing the clinical data of the patients for the patients and their caretakers, doctors, and insurance agents using cryptography and decentralization. The proposed system also maintains the balance between data accessibility and security. This paper also establishes how the proposed framework helps doctors, patients, caretakers, and external authorities to securely store and access patients' medical data in EHR.**

Keywords*—Blockchain; electronic health record (EHR); storage; security; accessibility; cryptography; decentralization*

I. Introduction

In the modern world, medical data sharing leads to the discovery of new techniques and treatments for curing several diseases. This can be done by storing the medical data digitally and by facilitating remote accessibility. The data stored in the electronic record is from the patients after visiting the hospital and making them the only owner of such records. The number of data stored in the electronic records is going on the increase and forms big data which can be used for several purposes in the healthcare domain. The vitality of data storage and sharing gives rise to several business entities for collecting, processing, analyzing, and storing the data to share them with other authorized sectors. This process increases several business organizations to focus on cloud storage and processing, data analytics, and provenance that renders existing organizations depending on the availability of data to operate and for their existence. To achieve the high demand in big data storage, several stakeholders invested in cloud computing and storage. This storage attracted the interests of several users including the patients, healthcare sectors, and research sectors for data storage in cloud repositories and provides controlled, cross-domain and flexible sharing of data to the beneficiaries. The major challenge in cloud data storage and sharing is the risk of the data being exposed to unauthorized third parties [1].

With the fundamental development of information and telecommunication technology, health-related services have been brought to the patient's doorstep with the help of the Telecare Medicine Information System (TMIS). The TMIS can help doctors to provide medical support from any remote location by discussing with patients about their illness and also by sharing critical information with other medical experts. In this way, the TMIS can reduce the treatment cost drastically. This system facilitates accurate decision-making in disease diagnosis by accessing up-to-date medical history. But the limitation here is the decision-making for new patients whose medical history and other related data are not available in health records. This can be overcome by using the EHR which holds all the data such as patient's details, scan reports, clinical notes, sensor data, billing details, medications, medical history, insurance details, and other related information. This type of record would suffer privacy and security issues in data-sharing [2]. Recently, wearable device technologies and the Internet of Things (IoT) have been evolving in the healthcare sector. Data from each wearable device were stored in the cloud which can provide big health data and valuable visions. This data is linked with the EHR to improve monitoring the heath, diagnosing the disease, and in the treatment of diseases [3].

The Electronic Medical Record (EMR) is a systematized digital record that holds the healthcare details of patients and populations. The initial perspective of EMR is to replace the traditional paper-based medical records and to enhance hospital data management in healthcare sectors. After that, the increase in the self-health concern, the general population also needed to access health records of their own. Hence, a novel personalized data management of healthcare information has been introduced, which is named Electronic Health Record (EHR) [4].

The EHR and EMR are offering improved security and user experience along with other healthcare-related aspects. Still, some security concerns have been believed to be resolved using Blockchain technology. The blockchain in the healthcare sector provides a secure and temper-proof system for recording medical data. This technology can also prevent inefficiency, insecurity, non-temper-proof, unorganized nature, duplication, and redundancy of data that occurs from the paper-based medical record [5]. State of transactions must not be easily detectable back to the relevant patient populations, according to advocates for transaction privacy on the

blockchain. To do so suggests using tokenization, which is a method of making only a representation to the sensitive material public while keeping the raw, confidential data private. Furthermore, it designates the need for health records to be securely stored off the blockchain. Since this blockchain is frequently used to store references to records stored in an access database, the database should be protected in and of itself. Our architecture protects the information both on blockchain as well as in the database by encoding information in the database.

An EHR is an electronic version of the patient's medical history which includes the patient's clinical data obtained from demographics, progress reports, problems, symptoms, immunization reports, medications, radiology reports, laboratory reports, and immunization reports. Recording the patient details in a paper-based report leads to an extensive paper trail in many healthcare organizations and hence, they are moved to EHR. The EHR should satisfy the requirements like accomplishing whole data, flexibility to failure, being available at any time, and beinging reliable to security guidelines [6]. In the current decade, various technologies have been used to secure patients' private data from healthcare sectors. The healthcare data of a patient includes the patient's details, their height, weight, symptoms of disease-affected, and previous medical history. This medical data grows with time. The data recorded in electronic health records are simple data points but difficult to manage. This record has been generated, stored, and manipulated by several stakeholders for proper patient care and the effective use of such medical records. These stakeholders get authorization whenever they need the data. The EHR is made of the parameters as shown in Fig. 1 [7].

Blockchain consists of a continuous sequence of blocks that stores all the records like a conventional public ledger. A block consists of only one parent block with a block header that holds the previous block hash value. The Ethereum blockchain also stores the uncle block hash values. The first block is called the genesis block. The genesis block has no parent blocks. The block header is given in Fig. 2.

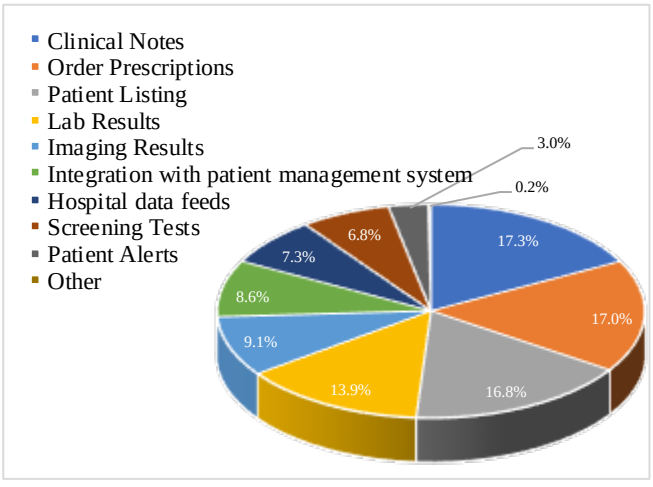


Fig. 1. Parameters of HER.

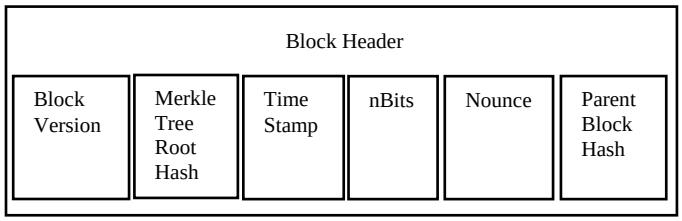


Fig. 2. Structure of Block Header.

Block version - Set of rules for block validation.

Merkle tree root hash - hash value of all records in a block.

Timestamp - Current time in seconds. Universal time since 1-Jan. 1970.

nBits - Target threshold value.

Nonce - 4-byte field starting from 0 and incremented while calculating each hash.

Parent block hash - Hash value with 256-bit pointing to the previous block.

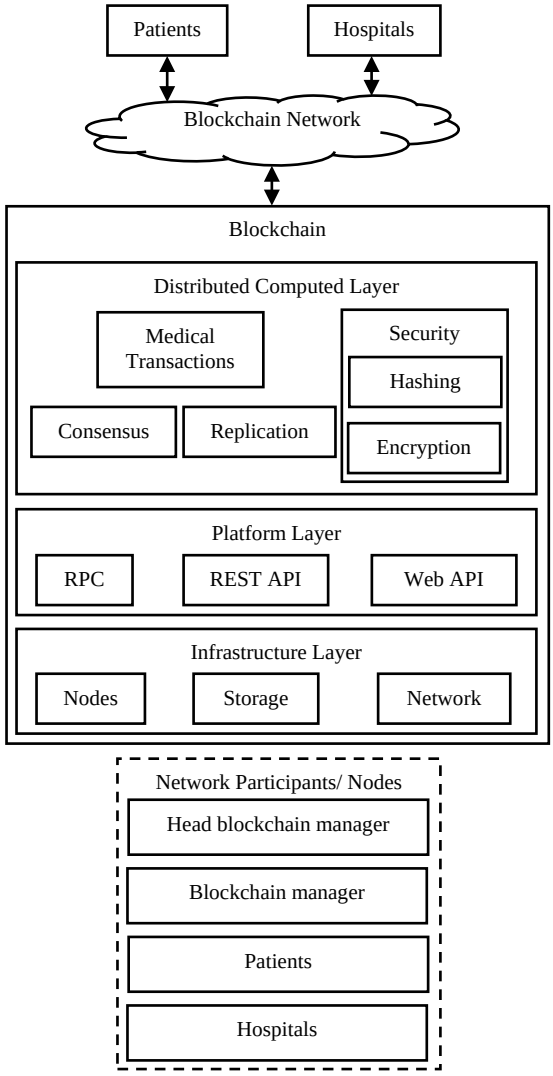


Fig. 3. Architecture of Blockchain Network in Healthcare Data Management.

While blockchain technology has many benefits for an EHR management system, it also has some drawbacks. The primary limitation of blockchain technology is that if % of the system's processing elements collide, the chain structure could be rewritten. To realize the benefits of a decentralized system, group members would have some confidence that at least mining nodes would not like to compromise the blockchain's immutability. Second, while using a permissioned blockchain undermines the incentive of external forces to connect PHI, it cannot hide transaction records. This gives nodes the ability to perform unfavorable network analysis. An adversary may indeed be able to identify the frequency through which a particular node attends a physician or the providers or third parties with whom a focus of previous associates by analyzing blocks of transactions. Finally, because cryptographic algorithms are distributed systems, their operation has a high memory usage. As a result, large amounts of data cannot be retained efficiently on the blockchain. As a result, while blockchains could be used for access management and data integrity, the information itself will be stored somewhere else and may be open to attack unrelated to the blockchain.

In a trustless environment, an asymmetric cryptography-based digital signature has been used throughout the network. Each user has a pair of private and public keys. The digital signature includes the signing and verification phase. The signing phase involves sharing the encrypted data with the private key and the original data. The verification phase involves validation of the data with a public key, whether it has been tampered with or not. The key characteristics of the blockchain are decentralization, persistency, anonymity, and suitability [8]. Fig. 3 shows the architecture of a blockchain network in healthcare data management.

Here, API – Application Programming Interface

RPC – Remote Procedure Calls

REST – REpresentational State Transfer

The traditional method is a Client-Server architecture or Singleton approach. The client is the end-user. The server gets the requests from the client which are then processed and the result will be forwarded to the client. A single authority (Server) will control the whole process. Whenever there is an attack on the server, the whole system will be collapsed. The modern Blockchain technology consists of data split over several systems. Each system is called a node. All the data are stored in a block that is connected via links formed by hash values. To calculate the hash value, the transaction in a block and the hash value of the previous block have been used [9].

The currently using healthcare data management involves centralized servers which seek permission to access multiple entities of medical data in a network which leads to delayed services and can be suspected to leakage of such information. Most of the patients are unaware of which entity stores and uses their medical data in such healthcare systems. The major challenge in this system is the security while accessing the data with various entities within the network. In such cases, Blockchain technology can be used to secure the accessibility and integrity of healthcare-related information [10].

The remaining sections in this paper cover the Related Works which describes the existing approaches that use blockchain in the management of healthcare data in Section 2, followed by the Proposed Methodology in Section 3, then Result and Discussion part in Section 4, and finally, the paper concludes in Section 5.

II. Related Work

A secure cloud-based EHR system has been implemented to accomplish confidentiality, authenticity, and integrity of healthcare data and to facilitate data sharing with the help of the C-AB/IB-ES scheme and blockchain. This system uses 5 entities namely key generation center, hospital, patient, cloud, and users who access the data. At first, the patients sign the health-related data and authorize the hospitals to access their data. This authorization letter will be submitted to the blockchain data pool and wait for consensus node processing. The hospital then encrypts the data and submits it to the data pool with the hospital's signature. The consensus node monitors the data pool and captures the matched authorization letter and the encrypted data. The signature is verified to make sure that the data is completed and with the patient's authorization. Then a consensus protocol would be performed to select a bookkeeping node that submits the encrypted data to the cloud along and the data description and its address were also written to the blockchain [11]. A prototype has been developed and implemented in a mobile platform for data sharing using Amazon cloud computing. This application uses the combination of blockchain and the decentralized Interplanetary File System (IPFS). The Ethereum blockchain has been used to demonstrate the performance of the developed Android mobile application. An Ethereum blockchain has been employed to build the e-healthcare system. Ethereum is a new distributed blockchain network like Bitcoin. The most significant merit of Ethereum is its adaptability and flexible nature which can be used to build an application using blockchain [12].

A permission blockchain network has been implemented for healthcare data management to overcome the issues associated with the permission-less blockchain network. This is because the permissioned blockchain network can overcome the problems like unauthorized network participation which causes impersonation of members, clear transaction data which includes the sensitive and confidential data of the patients that can be accessed by all the members in that network, network throughput is slow which hinders the treatment for patients, limited usability due to the payment for transaction and mining rewards. Also, the permissioned network prevents the demerits of permission-less networks such as high energy consumption, limited scalability, and low transaction throughput [13]. A survey on the investigation of the privacy and security issues while using wearable devices in the healthcare domain is given in [14]. For this survey, wearable healthcare devices have been designed and developed to collect the health-related data of the patients. By analyzing this data, the health status of a patient can be retrieved. The approach followed in this survey is the cross-sectional approach. This survey collected data from 106 respondents. Among them, 50% of respondents don't know the privacy concern in the healthcare data. The respondents

are also unaware of the security issues in the data collected using wearable devices. This survey finally suggests that the patients who are using the wearable devices should be educated about the privacy and security concerns in using them.

A new personal healthcare record sharing system with blockchain-based data integrity verifiable has been implemented in [15]. This scheme aims in resolving the issues that persisted in sharing the healthcare records like privacy disclosure, ability to search using limited keywords, loss in access control rights to share the personal health record. These issues have been overcome by using the techniques like searchable symmetric encryption and encryption based on attributes. This scheme varies from the existing methods in the way, that it uses an attribute private key to be distributed by the patients which avoids several problems that cause security issues in the existing systems. Also, this scheme uses blockchain to manage the keys which prevent the single-point failure issue in the management of centralized key. The efficiency of the data integrity verification has been improved by storing the hash value of the encrypted health data in the blockchain and storing the index set in the smart contract.

A blockchain-based access control manager for managing healthcare-related data has been described in [16]. This system is believed to overcome the challenges faced in interoperability by the industries stated by National Coordinator for Health Information Technology's (ONC) Shared Nationwide Interoperability Roadmap. Interoperability is one of the vital constituents for any structures that support Precision Medicine Initiative (PMI) and Patient-Centered Outcomes Research (PCOR). For access control management, this system uses a public blockchain for the data stored off-blockchain. Published research from the Massachusetts Institute of Technology has been borrowed to analyze the management and access control of personal data in a public blockchain. An implementation of a framework called Decentralized Application (DApp) in a private blockchain network platform using backend Distributed File System (DFS) has been given in [17]. The DApp is now using Proof-of-Work (PoW) consensus algorithm and is later suggested to use Delegated Proof-of-Stack (DPoS) consensus algorithm or Practical Byzantine Fault Tolerance (PBFT) consensus algorithm. This system uses Ethereum for implementing the smart contract-based healthcare blockchain. This application can easily detect anomalies, missing data, and the insertion of unauthorized data. The major elements used in the smart contracts are the events, functions, modifiers, and state variables which were inscribed via a high-level programming language called Solidity. To deploy the smart contract in test-net and test-net ethers, Remix and Kovan test network was applied to pay the fee for the transaction. There are three stages in creating a smart contract using Solidity, namely the writing, compiling, and then announcing. The real-time solidity compiler generates the bytecode and Ethereum Wallet is used in announcing the smart contracts to the blockchain.

The framework proposes prioritizes secure communication and contains several contributions aimed at improving privacy and interoperability. To begin, unlike other blockchain EHR systems that have been proposed, the blockchain stores hashing algorithms of data references while sending the real request network information in a private exchange over the framework. Our framework uses proxy re-encryption to simplify the secure transmission of EHRs, but it lacks techniques like private transactions for privacy. Furthermore, we can store keys and small encoded records straight on the blockchain using proxy re-encryption, making it easier to transfer records like prescribing to dispensaries or even other third parties. This eliminates the need for users to store keys locally, allowing patients to eliminate access permissions if preferred.

A review of Healthcare Information Management Systems (HIMS) based socio-technical issues has been performed and found the problems such as low privacy and security, lack of data transparency, data integrity and accessibility, errors in the prescription of medication and supply chain, and lack of knowledge interpretation. This review also provides the possible solutions in identity and risk management, auditing functions, and solutions for privacy and security issues using blockchain technology. Also, it provides some recommendations for future research and development of HIMS [18]. Model Chain, a healthcare predictive modeling framework for decentralized privacy-preserving in private blockchain network has been demonstrated which uses privacy-preserving online machine learning algorithm adopting blockchain technology, in which the transaction metadata has been applied for distributing the partial model and also designed a proof-of-information system for finding the order of online learning process. This approach aims to improve the interoperability among the organizations to support Nationwide Interoperability Roadmap and national healthcare delivery priorities such as Patient-Centered Outcomes Research (PCOR) and to find the solution for privacy-preserving healthcare predictive modeling by using a peer-to-peer network like blockchain [19].

The health information that is monitored and transmitted by the remote health monitoring devices through IoT has been protected by applying the smart contract mechanism based on blockchain technology to enable secure managing and analyzing the data obtained from medical sensors. Ethereum protocol is applied to a private blockchain network to make the medical sensors communicate with smart devices which call the smart contract and record the events completely into the blockchain network. The use of such a smart contract system facilitates real-time monitoring of patients and their medical data by transmitting them as notifications to the hospitals and patients in a highly secured platform. This approach can solve several security issues occurring in remote health monitoring of patients and automatically notifies the parties involved in the process in a HIPAA-compliant manner [20].

III. Proposed Methodology

The EHR has been managed by the healthcare institutions instead of the respective patients. This leads to difficulty for other health centers to access the patient details to provide perfect medical advice to the patients. Hence, the patients need to retain their health information for future access. The blockchain allows to store the healthcare data and provides

free access to the EHR via corresponding data providers and websites [21]. The proposed system develops a security framework for EHR which provides access to multiple authorities in a shared system using blockchain. Fig. 4 demonstrates the process flow of the proposed framework for healthcare data storage and access.

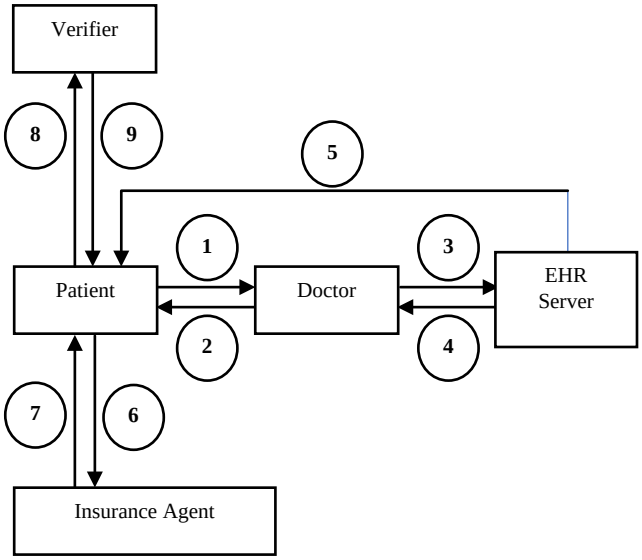


Fig. 4. Proposed Blockchain Framework for Healthcare Data Storage and Access.

The proposed blockchain framework enables the patients to directly access their data from EHR and can able to download and share them. There are 5 parties involved in the multi-user system. They are the Patient, Doctor, EHR Server, Insurance Agent, and the Verifier. The steps involved in the storage and sharing of patient's medical and personal data are given as follows:

1) The patient consults the doctor.

2) Treatment has been given to the patient by the doctor.

3) The EHR Server is a node present in the blockchain network which serves as a miner who collect the transaction data and store them as blocks. After creating the EHR, they are verified by all the nodes in the blockchain network whether it is a valid one or not. These transactions are stored in a memory pool which acts as a waiting area for all the transactions performed on each node and maintains those details within their node. The miner node collects this transaction information and forms it into blocks. The verification of such data can be done by using the hash, which is a 256-bit number that represents unique data. Once the verification got completed, the miner picks it from the memory pool and inserts them into a new block which will then be submitted to the blockchain.

4) Once the block is created, it will be distributed by the miner to all the nodes available in that blockchain network.

5) Access control has been provided to each of the nodes within the network. The proposed framework works in a way that the information of each patient will be secured by them. It means, the patient's details can be accessed by patients and also the doctor who gives treatment to them. This can be done only with the permission of that patient.

6) This data when claimed for insurance purposes, then it will be shared by the doctor who treated that patient to the insurance agent.

7) The insurance agent can refer the EHR of the patients, only who claimed from them and can approve the insurance payment to the patient.

8) The patient then sends the request for the verification of their data to the data verifier.

9) The verifier finally verifies and approves whether the data provided are safe and secured or not and delivers a verification result to the patients.

This framework allows only the patients to view their data in EHR. When a patient permits the doctor who treats them, then that doctor can also view the data. Likewise, the insurance agent can view only the data of the patient who claims insurance amount from them.

Algorithm-1: Formation and addition of Patient Blockchain

Input: Details of Patient for EHR

Output: Forming Patient blockchain and adding blocks to it

- Provide the medical data of the patient to the EHR
- Generate private and public keys using RSA cryptography technique
- The public key is used by the patients for encryption and the private key is for decrypting the encrypted data, by the doctor and insurance agent
- Generate the Hash for Encrypted EHR based on HMAC-SHA1 Algorithm
- With the help of the patient's ID, generate a Bilinear Map for the Encrypted EHR
- With the help of the Patient's name, ID and password, create a genesis block for the Patient
- Add the Encrypted EHR and Hash with Bilinear Map to the Block
- Add this block to the patient blockchain.

Algorithm-1 gives the formation and addition of Patient Blockchain. A Bilinear Map is used to enhance the security of the proposed framework. A function that combines 2 vectors to get a new vector is called the bilinear map which can be mathematically represented as:

$\overrightarrow{v_1} \times \overrightarrow{v_2} \rightarrow y$

Here, $\overrightarrow{v_1}$ = Encrypted EHR

$\overrightarrow{v_2}$ = Patient ID

y = Bilinear Map

This bilinear map is generated using the concept of identity-based encryption. Fig. 5 shows the Patient blockchain which consists of a genesis block with the patient's name, ID, and Password. The treatment taken by the patients has been added as a new block one by one. Only the patient can view the data stored in the patient blockchain, other than that no one can view it.

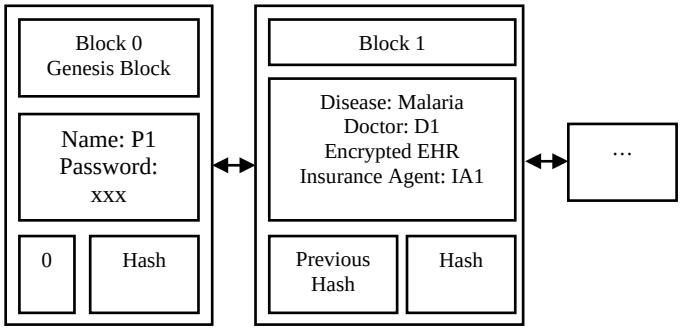


Fig. 5. Patient Blockchain.

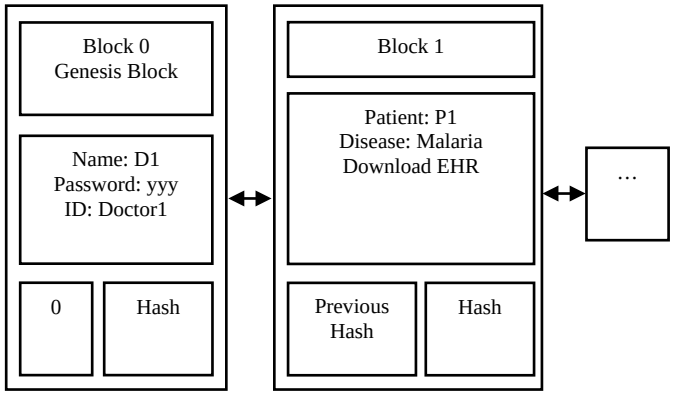


Fig. 6. Doctor Blockchain.

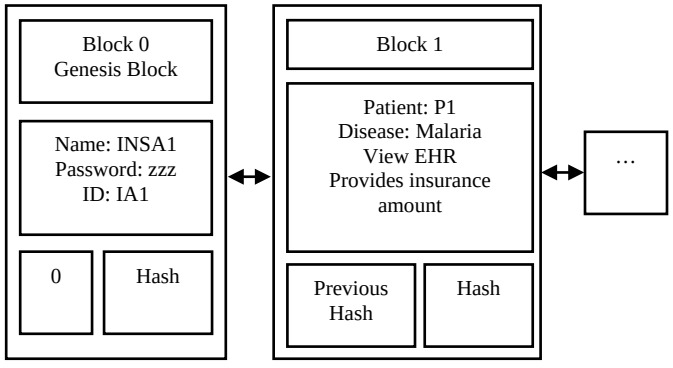


Fig. 7. Insurance Agent Blockchain.

Algorithm-2: Formation and addition of Doctor and Insurance Agent Blockchain

Input: Patient block which is referred from the patient blockchain. Forming Doctor blockchain and adding blocks to it

Output: Forming Insurance Agent blockchain and adding blocks to it

- The doctor and the Insurance Agent downloads the data of the referred patient block to their block with the help of a private key
- Encrypted EHR and Hash with Bilinear Map is retrieved from the block
- Using a private key, decrypt the Encrypted EHR
- The EHR is accessed by the Doctor and the Insurance Agent
- With the help of the Doctor's name, ID and password, create a genesis block for the Doctor similarly create for the Insurance Agent also
- Add the Encrypted EHR and Hash with Bilinear Map to the block
- Add this block to the Doctor blockchain and Insurance Agent blockchain
- The insurance amount for the treatment is transferred to the Patient block.

Algorithm-2 gives the formation and addition of Doctor and Insurance Agent Blockchain. Fig. 6 shows the Doctor Blockchain which consists of a genesis block with the Doctor's name, ID, and password. The information related to the treatment of the diseases was added as a new block one by one. Only the blocks with the patient's permission can be viewed by the doctors. Fig. 7 shows the Insurance Agent Blockchain which consists of a genesis block with the Insurance Agent's name, ID, and password. The information related to the treatment of the diseases was added as a new block one by one. Only the blocks with the patient's permission can be viewed by the insurance agents.

Algorithm-3: Blockchain Validation

Input: Patient Blockchain

Output: Validation Result (Safe or Not Safe)

- Download Patient blockchain
- Status is Safe
- **for** each Block from Blockchain
  - From Block, Encrypted EHR and Hash with Bilinear Map were retrieved
  - Generate new Hash for Encrypted EHR based on HMAC-SHA1 Algorithm
  - Generate new Bilinear Map for Encrypted EHR
- **if** ((Hash == new Hash) & (Bilinear Map == new Bilinear Map)
  - Block = Safe
- **else**
  - Block = Not safe
- **break**
- **end for**

All the data in the EHR were encrypted within the blockchain and cannot be accessed by anyone. The Data Verifier finally verifies the Patient Blockchain, whether it is safe or not. Algorithm 3 gives the steps involved in Blockchain Validation. The block which has to be verified by the data verifier will be searched in the blockchain. Then the Encrypted EHR and the Hash with Bilinear Map will be retrieved based on the HMAC-SHA1 Algorithm, a new Hash for the Encrypted EHR will be generated. Then, if the Hash value is equal to the new Hash value along with the Bilinear Map equal to the new Bilinear Map, the corresponding Block will be considered as Safe Block and if not, then the corresponding Block will be considered as not Safe Block.

IV. Result and Discussion

The access control of the proposed system has been designed in such a way, that the parties like insurance agents and doctors, who were granted permission from the patient can only have the access to the EHR of the patient blockchain so that preventing the access of unauthorized parties to the EHR.

The time consumed for the access of EHR in a blockchain using the proposed approach is compared with existing centralized storage systems and the result has been graphically represented in Fig. 8. Data has been requested to the EHR for accessing and the time is taken to receive the requested data has been noted.

In centralized storage, the EHR will be stored in a centralized server. The patient who needs access to those records should raise an EHR request to the centralized server and this time is noted as T_1. The centralized server, after receiving the EHR request, search for the availability of the particular data and then transmit them to the patient and this time will be noted as T_2. Hence, the time consumed for data retrieval will be calculated using the following formula:

$$Time\ consumtion\ (in\ secs) = T_2 - T_1$$

Moreover, the time consumed for searching and accessing the data would be directly dependent on the size of the EHR. It means the time consumption varies with the size of the EHR, i.e. if the EHR size is small, the time consumed will be less and if the EHR size is large, the time consumed will also be high. The comparison result proves that in centralized storage, the time consumption is higher than that of the proposed blockchain method which has been shown in Fig. 8.

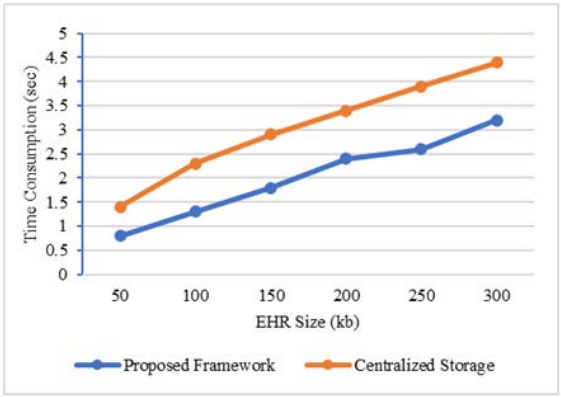


Fig. 8. Comparison of Time Consumption between Centralized Storage and Proposed Framework.

Table I gives the comparison of the features associated with the proposed approach with some existing approaches. The result of the comparison clearly shows that the proposed framework outperforms the existing works and can provide a safe and secure storing and sharing of patient details in EHR using blockchain technology.

TABLE I. Comparison of Proposed Framework with Some Existing Approaches

Feature	[3]	[8]	[9]	[10]	Proposed
Authentication	✓		✓	✓	✓
Identity Management	✓			✓	✓
Decentralized Access		✓	✓	✓	✓
Privacy	✓	✓	✓	✓	✓
Integrity	✓	✓	✓	✓	✓
Availability		✓		✓	✓
Flexibility		✓			✓

V. Conclusion

The proposed framework described the various features of blockchain in the field of healthcare for data storage in EHR and sharing them between the users. This framework overcame the limitations of current data models and supply chains. The access control, time consumption for requesting and searching data in EHR in a blockchain, and the feature comparison were also discussed in this paper and the results show that the proposed system outperforms the existing approaches in all possible ways. From the proposed approach, it is clear that the use of blockchain in healthcare data management prevents data breaches and fraudulent billing and enhances privacy, security, and transparency. Also, data sharing via blockchain facilitates safe and secure sharing among authorized third parties. This approach guarantees secure healthcare management among all the levels which include patients, doctors, hospitals, insurance companies, Pharmaceuticals, etc.

Acknowledgment

The authors would like to thank the Deanship of Scientific Research at Jouf University for supporting this work by Grant Code: (DSR-2021-02-0375).

Funding Statement: This work was funded by the Deanship of Scientific Research at Jouf University under grant No (DSR-2021-02-0375).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Shamshad, S.; Mahmood, K.; Kumari, S.; Chen, C.M. A secure blockchain-based e-health records storage and sharing scheme. J. Inf. Secur. Appl. 2020, 55, 102590.

[2] Shahnaz, A.; Qamar, U.; Khalid, A. Using blockchain for electronic health records. IEEE Access 2019, 7, 147782–147795.

[3] Ying, Z.;Wei, L.; Li, Q.; Liu, X.; Cui, J. A lightweight policy preserving EHR sharing scheme in the cloud. IEEE Access 2018, 6, 53698–53708.

[4] Zibin Zheng; Shaoan Xie et al. “An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends”, 2017 IEEE International Congress on Big Data (BigData Congress), 25-30 June 2017.

[5] E. Bertino, R. Deng, X. Huang and J. Zhou, “Security and privacy of electronic health information systems”, International Journal of Information Security, vol. 14, no. 6, pp. 485-486, 2015.

[6] J. Vora et al., "BHEEM: A Blockchain-Based Framework for Securing Electronic Health Records", in 2018 IEEE Globecom Workshops (GC Wkshps), 2019.

[7] V. V., K. Sabarivelan, J. Tamizhselvan, B. Ranjith and V. B., “Utilization of Blockchain in Medical Healthcare Record using Hyperledger Fabric”, International Journal of Research in Advent Technology, Vol.7, No.4, April 2019 E-ISSN: 2321-9637.

[8] Liang, X.; Zhao, J.; Shetty, S.; Liu, J.; Li, D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In Proceedings of the 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), pp. 1–5.

[9] Ramani, V.; Kumar, T.; Bracken, A.; Liyanage, M.; Ylianttila, M. Secure and efficient data accessibility in blockchain-based healthcare systems. In Proceedings of the IEEE Global Communications Conference (Globecom), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 206–212.

[10] Xia, Q.I.; Sifah, E.B.; Asamoah, K.O.; Gao, J.; Du, X.; Guizani, M. MeDShare: Trust-less medical data sharing among cloud service providers via blockchain. IEEE Access 2017, 5, 14757–14767.

[11] Wang, H.; Song, Y. Secure cloud-based EHR system using attribute-based cryptosystem and blockchain. J. Med. Syst. 2018, 42, 152.

[12] Nguyen, D.C.; Pathirana, P.N.; Ding, M.; Seneviratne, A. Blockchain for secure EHR sharing of mobile cloud-based e-health systems. IEEE Access 2019, 7, 66792–66806.

[13] Ismail, L.; Materwala, H.; Zeadally, S. Lightweight blockchain for healthcare. IEEE Access 2019, 7, 149935–149951.

[14] Cilliers, L. Wearable devices in healthcare: Privacy and information security issues. Health Inf. Manag. J. 2020, 49, 150–156.

[15] Wang, S.; Zhang, D.; Zhang, Y. Blockchain-based personal health records sharing scheme with data integrity verifiable. IEEE Access 2019, 7, 102887–102901.

[16] Linn, L.A.; Martha, B.K. Blockchain for Health Data and Its Potential Use in Health It and Health Care Related Research. In Use of Blockchain for Healthcare

[17] and Research Workshop; ONC/NIST: Gaithersburg, MD, USA, 2016.

[18] Asma Khatoon, A Blockchain-Based Smart Contract System for Healthcare Management. Electronics 2020, 9, 94.

[19] Litchfield, A.T.; Khan, A. A Review of Issues in Healthcare Information Management Systems and Blockchain Solutions; CONF-IRM, 2019.

[20] Kuo, T.T.; Ohno-Machado, L. Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks. arXiv 2018, arXiv:1802.01746.

[21] K. N. Griggs, O. Ossipova, C. P. Kohlios, A. N. Baccarini, E. A. Howson, and T. Hayajneh, “Healthcare blockchain system using smart contracts for secure automated remote patient monitoring,” Journal of Medical Systems., vol. 42, no. 7, pp. 130–138, 2018.

Usability Evaluation of Web Search User Interfaces from the Elderly Perspective

Khalid Krayz Allah, Nor Azman Ismail, Layla Hasan, Wong Yee Leng
School of Computing, Faculty of Engineering
Universiti Teknologi Malaysia Johor, Malaysia

***Abstract*—The elderly population is increasing in many countries, often with health and incapacity challenges, largely disengaged them from the world of digital tools like Internet usage. They browse the Internet daily for obtaining needed information through various search engines through the search UI. Earlier technologies were fabricated for improving daily life, but the specific needs of the elderly are neglected often. Currently, available online search UIs are well-developed, but they did not consider usability in their design specifically for the elderly. This research aims to evaluate web search UIs based on the elderly perspectives to identify existing search UIs usability issues and recommend improvements to web search UI designs. The observation technique evaluated two web search UIs (Google interface and Bing interface) with fifteen participants aged 60 years and above. System Usability Scale (SUS) questionnaire was applied to measure the user satisfaction of the current two interfaces. The data collected from the observations were analyzed using content analysis, while the data acquired from the questionnaires were analyzed using the t-test. The results revealed a statistically significant difference in SUS ratings, with Google scoring 73.5 and Bing scoring 66.5, indicating that users prefer the Google interface over the Bing interface. Besides that, the usability issues were identified, and recommendations to improve the design of the search UI were suggested. These findings contribute to a better understanding of the issues that prevent elderly users from using web search UI and valuable feedback to designers on improving the UI to suit the elderly better.**

***Keywords*—Usability; Google interface; Bing interface; SUS questionnaire; web search user interfaces; observation method**

I. Introduction

The number of elderly people is increasing rapidly in most of the countries in the world as well as their use of the Internet is also increasing [1, 2]. They use the Internet daily for searching various information, mostly via search engines. Due to the growth and impact of information technology in our day-to-day life, one should need to gather more information from the websites. So, searching on the Internet is an important cognitive process to find out the needs of different kinds of resources to achieve their aim [3]. The search UI is the way to communicate and interact the users with search engines to acquire the desired information [4]. Although research on estimation models based on web search logs aims to improve our daily lives, the needs of the elderly are frequently overlooked [5]. Those designs may be challenging to learn and use for elderly people. Researchers are looking at the web search activity for the usage of elders in professional and business domains as a result of this trend.

Furthermore, the user interface (UI) is considered to be an important component of any interactive software system from the users' perspective because it is the most visible front-end component through which the users could see and work with and perform primary evaluation while utilizing the system [6, 7]. As a result, the needs of users in system development would lead to effective user interfaces and useable collaborative systems. Alternatively, the bad user interface design would cause a greater challenge for the users. Yet, there is no clear evidence that poor user interface design is very much challenging.

The exposure to web search for the elderly user is less because of the cognitive functions that reduce due to age [8]. Both elderly and younger ones have some search strategies to use their knowledge and skills. The elderly person searches very little but attains more appropriate information and performs well by acquiring better results [8, 9]. However, elderly users cannot adapt to a newer searching strategy and change their search pattern or style even for highly challenging search tasks [9, 10]. Also, elderly users have physical problems like visual impairment, colour identification etc., cognitive impairment, knowledge about computers and technologies [10].

Usability is an essential key factor for the software developers or users, as it ensures the successiveness of the system and its further development by focusing on the needs and requirements of the users [11, 12]. Usability, according to ISO 9241 [13], is defined as the absence of usability difficulties or the measuring of efficiency, effectiveness and satisfaction. Usability is contingent on the absence of usability issues. The main concern of users is the usability of the software and the consequences of utilizing it without the knowledge of the systems core components, it is working, or its production [14]. While developing software, the needs of the users must be considered and given careful consideration.

Various usability evaluation methods were developed, which comprises a set of techniques for evaluating the usability of the systems user interface and identifying specific issues [15, 12]. In general, usability evaluation methods can be categorized into expert-based and user-based [8, 9].

Expert-based techniques, often called inspection methods, involve experts evaluating the user interface and identifying potential issues that users might encounter while interacting [16]. Such studies can result in a formal report highlighting problems or making suggestions for improvements [17]. The heuristic evaluation method and cognitive walkthrough

method are the two most commonly utilized expert-based usability approaches in the area of human interaction [18].

User-based techniques, as well known as test methods, are useful evaluation methods for obtaining information about real users' behaviours as well as finding and identifying usability issues by noticing people that represent users be using on the system interface [9, 10]. User-based techniques are used to see how well a system assists the end-user with their tasks [8]. Empirical or experimental techniques, methods used in this study, query methods, and physiological monitoring methods, such as gaze and heart rate and skin conductance measurements, are all extensively used in the field of human-computer interaction among the various approaches that are based on the users [18].

Several earlier studies evaluated the usability of search UIs based on user-based methods for elderly people [19, 20, 21, 22]. Furthermore, there seems to be a scarcity of research that assesses the usability of search engine UIs and offers particular changes to their design to make them easier to use by the elderly.

This research aims to evaluate the usability of web search UIs from elderly users' perspectives to uncover the usability issues on these UIs, and based on the results to suggest specific recommendations for usable web search UIs for elderly people.

The specific objectives of this research are:

1) To use observational techniques to identify usability issues on two common web search user interfaces; Google and Bing.

2) To use the System Usability Scale (SUS) questionnaire to measure the user satisfaction of the two web search user interfaces (Google interface and Bing interface).

3) To identify additional user requirements that could aid in creating a proposed design for the newly designed UI for the elderly group.

4) To recommend specific improvements to design usable web search user interfaces for elderly users.

The results of this research will uncover challenges on the search UIs from the perspectives of elderly users and would reflect the requirements to conduct the improvements process to meet users' needs and requirements. This will enhance the ease of use of web searching by the elderly community and provide information to web search interface designers about possible improvements for designing a user-friendly search user interface for elderly users.

This research is divided into six sections. Section II presents earlier studies that evaluated the usability of web search user interfaces for elderly users. The methodology is presented in Section III. Section IV presents the results, while Section V presents the discussion. Finally, in Section VI, the conclusion is outlined.

II. Usability Evaluation for the Elderly

The literature showed that research had been done to evaluate the usability of desktop or laptop user interfaces for elderly people. Many studies have been published that employed usability testing to evaluate the usability of websites and smartphone user interfaces for elderly users using representative user groups [19 - 23]. Specifically, Patsoule and Koutsabasis conducted a comparative usability evaluation of two websites where the participants with 12 older adults aged 65 years and over [24]. Controlled usability testing, as well as post hoc interviews and questionnaires, were employed to assess their performance on six standardized activities. The rebuilt website was far more functional and acceptable than the previous version [24].

In addition, Haesner et al. (2018) conducted a usability test with older users to analyze the usability and acceptance of Google Glass [25]. The participants were 30 elders aged 65 and up who were requested to complete a set of standardized tasks and evaluate usability using a system usability scale questionnaire in order to acquire valuable information into specific usability difficulties. The final findings revealed that usability should be considered while developing mobile applications for elderly people by using Google Glass [25].

Alternatively, there have been a number of user evaluation studies of web search UIs based on user-based methods [19, 20.21,22]. Sanchiz et al. (2019) used eye-tracking metrics to test 9 search issues using standard web browsers and/or empirical search interfaces. The final empirical findings showed that older adults spend significantly more time on search engines than younger adults [4].

More specifically, Aula et al. [19, 26] have conducted two usability tests to compare the usability of Etsin, a friendly search engine for the elderly, and Google by giving search tasks to elderly users. The observation method was used to monitor the elderly's interacting behavior during the search tasks, followed by an interview to find all usability problems of web user interface engines. The results identified age-related issues in the search UIs that should be taken into account, which are beneficial to elderly users. For example, the search engine interface should be a simple one that is easy to use and understand.

III. Research Methodology

This study is conducted in three phases: usability testing, evaluation, and analyzing and interpreting data, as shown in Fig. 1. To achieve the objectives of this research, two usability testing methods were employed: observation and System Usability Scale (SUS) questionnaire. This section consists of five sub-sections which describes: the methods employed in this research, participants, apparatus, the procedure of the testing, and the analysis of the collected data.

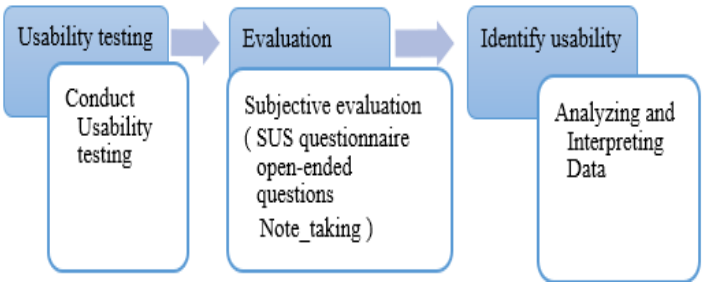


Fig. 1. Three Main Phases of the Methodology.

A. Methods

1) Observation and note-taking: Observing how users interact with a system is a common technique to learn more about its use. This technique tries to monitor participants while they utilize the web search user interface to conduct their searches. Taking notes during the usability test, these participants' reactions to the challenges they encounter while engaging with the system are also recorded [19, 22, 26]. Conducting a usability test is the most effective way to obtain high-quality qualitative data by observing the interface and users' reactions to the task, which allows the practitioner to quickly identify system design flaws [27, 28]. As a result, the practitioner observed the participants throughout the sessions and took notes. As a result, users are encouraged to "think aloud" about their actions [29].

In this research, observers were employed to take notes and observe the participation behaviors while interacting with the interfaces during the usability test.

2) System usability scale (SUS): The system usability scale (SUS) is a popular tool among HCI researchers for assessing perceived usability in both usability and survey research [30, 31, 32]. According to reports, the SUS was used in 43% of usability studies [33, 34]. Furthermore, a study that evaluated SUS discovered a valid and reliable tool for evaluating usability [34, 35]. The SUS is a ten items questionnaire with alternating positive and negative statements and a five-point Likert scale spanning from strongly disagree to strongly agree in ascending order designed to avoid answer biases. These ten items will be assessed, and a final score (0–100) will be determined by grading them from A to F or using adjective ratings. Bangor Different SUS grade rankings are presented by Kortum and Miller [36]. Before 2009, the SUS study was only focused on perceived usability (unidimensional). However, Lewis and Sauro [34, 37] suggested that the SUS is a bi-dimensional measure (usability and learnability sub-scales) in 2017. If the research situation is convenient and the participants are experienced, the SUS was advised in employing bi-dimensional analyses [37, 38]. Furthermore, according to Tullis and Stetson [34, 39], SUS allows researchers to test perceived usability with a small sample size of 8–12.

In this study, bi-dimensional and unidimensional analyses of the SUS scale were undertaken due to the diverse backgrounds of the participants.

B. Participates

The participants were recruited from elderly homes and through personal contacts. The participants in the study were given a feedback form with demographic information and questions regarding their computer and Internet search experience.

C. Apparatus

Windows operating system ten was utilized by all of the participants in the search sessions. Only nine users used Chrome to access the Google interface, while six used Mozilla Firefox to access the Bing interface. The monitors were identical in size and resolution (1024 x 768 on a 15-inch panel).

D. Procedure

At the start of the usability evaluation, all of the participants were told that the study was all about finding facts on the Internet and gathering data on how useful the interface is. They were also told that throughout the test session, they might run into some difficulties with a specific test task in which they don't have to worry, and if they feel difficult to finish it, then they can just say that it is difficult to continue and would like to stop, and can move on to the next task. They were also promised that if they typed text into the text field, the search engine would return any publications that contained the text or all the text they had entered. Before the test began, the participants signed a document indicating their consent to participate in the testing as well as the recording and reporting of their responses for the study.

All of the participants were given a sequence of seven search tasks containing a variety of interests. They were also instructed to conduct relevant searches on those topics using the Google and Bing search engines. They were then invited to ask the first question for the job of their choice and then proceed with the activity normally. The participants utilized the computer on their own during the search phase. They were, however, able to ask questions and receive assistance if needed during the search.

The search session was for 20-30 minutes. During the session, observers were present to take note of the participants' reactions while using the search engine. After the participants searched in the search engine, all the participants were given a questionnaire having a set of questions about the interface they have worked on to detect usability issues and additional users' requirements. The questions were intended in such a way that it uses the SUS questionnaire for UI satisfaction [18, 40]. The satisfaction of the participants is measured on a scale of 5-points. The questions include the overall reaction to the web search interface. Two open-ended questions allowed the participants to write down the positive and the negative feedback about the interface in their point of view. These questions provide valuable information about the interface design in addition to the statistical data. Finally, to make it easier to relate the interfaces, they were shown both of them simultaneously with the different questions with the same meaning.

E. Analyzing and Interpreting Data

Following the sessions, the observers went through their notes and the questionnaires to flesh out the usability issues, mostly experienced by more than eight of the fifteen participants for each search UI. During the search, the participants' verbalizations and behaviours were recorded. Following that, a list of usability issues for each interface was developed. Finally, the common usability issues and participant behaviour observations found in both the Google engine interface and the Bing engine interface were assembled.

IV. Results

A. Participants' Characteristics

A total of 15 people volunteered in this study, eight male (53%) and seven female (47%). All of the participants were above 60 years old, with an average age of 62.5 years for males and 61.5 years for females. The age ranges from 60 to 67 years for males and from 60 to 65 years for females. The demographic information of the participants is summarized in Table I. All of the participants used computers in their daily lives when it came to searching the Internet, with an average experience of 9 years (3 to 10 years). The majority of male participants utilized the Internet to search on a regular basis, while female participants had an average experience of 4 years (2 to 7 years). Most of the participants had previous experience with search engines like Bing, Google, and MSN Search, where the others are rookie searchers who are mostly directed to facts from well-known URLs, if at all.

TABLE I. The Participants' Information

Participates sample characterization		
Gender	Male, 8 (53%)	Female, 7 (47%)
Mean Age	62.5	61.57
Mean Experience	6 years	4 years

B. Qualitative Results

This sub-section presents the qualitative results obtained from the content analysis of the observation and note-taking methods and the questionnaire's open-ended questions. It shows the major problems that the elderly users faced while interacting with both interfaces, which provide important information about their perspectives and experiences. The collected data were transcribed as clear data, as shown in Table II.

TABLE II. Usability Problems Identified from the Qualitative Results

Usability Problems	**Google Interface**	**Bing Interface**
The default font size is small. Also, the font size varies from one browser to another.	✓	✓
The voice search button, the image search button and the text search button are small and close to each other.	✓	✓
Difficult to go back to bookmark and history in different browsers	✓	✓
A large number of results		✓
Misunderstanding suggested results such as People Also Search For and Related Searches	✓	✓
The home page is long		✓
Scrolling a page is a difficult task	✓	✓
Unclear to go back to the home page	✓	✓
Lots of different colours in the default page background		✓
The main and result page view is slightly unstructured	✓	✓
The setting menu is complex.	✓	✓

Despite some participants being less experienced Internet users, all participants could successfully complete at least five out of seven search tasks during the search sessions. Also, it was noticed that the number of tasks performed by the more experienced participants differed significantly from those done by less experienced in this field.

The majority of issues stemmed from a lack of understanding of the basic web structure (using the back button to go to the main page) and using standard interaction styles (typing a query into the search box without concentrating and expecting the request to be completed without clicking on the search button). These issues appeared to perplex new users; they were more annoying. The expectation that these issues will arise due to a lack of practice with standard interaction elements and the expectation that the problems will be less with the experience. These issues can also be alleviated by interface design solutions, which will be discussed later.

The results of the usability tests were analyzed to create a list of the issues discovered through the observational technique and open-ended questions. Then, a new solution for the prototype was proposed. The following go over the problems that were discovered:

1) Button size: For a user-friendly interface, bigger button sizes have greater significance. The button size of the main navigation links and the home page is intended to be an important standard because the on-screen button size has a greater effect on interacting speed and communication accuracy. The clicking accuracy with a target diameter of 64-pixels was significantly better than on 32-pixel targets for elders.

2) Font size: Also, font size is another major problem needed to be considered. Fancy font types with font sizes below 14 cause loss of clarity for the elderly people in accessing, and these findings are in line with previous research [41]. Therefore, the utility will be more effective if the font size chosen is at the size of 18 pts.

3) Searching ways: There are common ways of searching in use: searching by text, searching by voice, and searching by image. There is a misunderstanding between searching by voice, text, and image by elderly users. They should be separated from each other with enough space and recognized with the button. Also, the button should be having a recognized name or logo to show the differences between them.

4) Home page length: Page scrolling is a complex task for aged people as they require using drag-and-drop or mouse scrolling. Also, they have reduced memory capability than the younger. So, while scrolling the lengthy home page, the contents on the first screen couldn't be recalled, and they got confused. Hence, for older people, if the page length is within one screen, it would be better than having 2 to 3 screens. For aged ones, it would be easier if the average quantity of associated results exposed on the screen was around 5 to 7.

5) Backtracking support: The use of the back button to return to the home page or previous results page after seeing

the result document generates a lot of confusion among elderly users since they don't grasp the web's structure and can't tell the difference between the back and forward and home buttons. Elderly users frequently misunderstand the back button on a browser or a webpage, resulting in returning to the same page. These conditions cause elderly users to be unable to distinguish between the functions of each button. Every time an elderly user clicked on the forward, back, or home buttons, they needed some assistance from their moderators to confidently click on them and avoid being lost. These findings are consistent with Cioara et al.'s study [42]. For elderly users, it would be preferable if the search result were displayed in a separate tab or window, allowing them to go to the webpage they desired quickly. It is also quite beneficial if the pop-up window is smaller than the existing window since it allows them to return to the original window by simply clicking on it.

6) Colour contrast of foreground and background objects: Because of aging, elderly users cannot precisely distinguish colour variations like pink, magenta, and purple. Suppose the scenario rises for clicking the particular colour button to move on to the next. In that case, the aged people could not perform well, as their retina could not clearly support the vision to differentiate the color variations. This may lead to mistakes and errors while browsing, and this reduces the search accuracy. Also, the color contrasting of the foreground and the background objects is a major problem, i.e., if the texts in the light background are dim, then it would be hard for the older user to read and proceed further because their vision would not support with age. The web page layout with an off-white background and a high contrast text over a pale background is easy for the older to read.

7) Menu: The observation results found that elderly users are confused by the Hamburger menu and Dropdown menu. Pull-right menus are also challenging for elderly people to navigate, and they frequently have to make multiple attempts before even being able to choose their preferred option. Thus, simpler menus are healthier choices, such as the Classic navigation menu and Sidebar menu.

8) Search results: Search results refer to the list created by search engines in response to a query. Today's search results also include sponsored search results. In addition, the search results also show Related Searches, People Also Search For, People Also Ask, and Top Stories etc. Because of that, elderly people misunderstand these suggested results. For the elders, the view of less is more, so natural search results returned by the search engine's internal evaluation algorithm based on relevancy are better suited and do not include any other results.

9) Result page view: Because users do not often process search results in a logical order, today's search-results pages could have a wide range of layouts. They use the pin-ball pattern to distribute their attention more evenly across the page than in the past. Images, video, embedded text content, and even interactive features are frequently included in today's search engine results pages. Any given search can yield a wide variety of visual elements. The variety of information and presentation is crucial in shifting user attention throughout the SERP [43]. To make it easier for the elderly to find the information, they need a simple list of less than ten blue links, each neatly packaged with a URL, blue link, and text snippet.

C. Results of System Usability Scale (SUS)

In a test between two web search interfaces, randomly 15 users worked with two web search user interfaces (Google and Bing). They conducted seven different tasks on both interfaces before completing the ten-item SUS questionnaire [44], with the findings provided in Table III. (The difference score is calculated by subtracting the Bing interfaces score from the Google interfaces score) [45].

TABLE III. PAIRS OF SYSTEM USABILITY SCALE SCORES AND THEIR DIFFERENCES

Participant	Google interface	Bing Interface	Differences
1	77.5	72.5	5
2	67.5	65	2.5
3	70	65	5
4	65	57.5	7.5
5	65	32.5	32.5
6	80	72.5	7.5
7	70	67.5	2.5
8	77.5	65	12.5
9	70	70	0
10	77.5	67.5	10
11	80	70	10
12	77.5	75	2.5
13	75	75	0
14	70	67.5	2.5
15	80	75	5
Mean	73.5	66.5	7

To measure aspects related to interface usability, by using the SUS questionnaire [46], created by John Brooke in 1996 [47] with the minor modifications by Finstad [46] in 2006. It is a versatile and quick method that is extensively used to assess the system's usability.

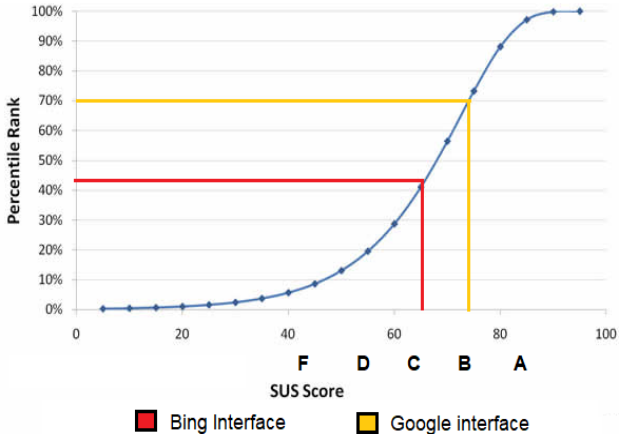


Fig. 2. Percentile Rankings of SUS Scores from between Bing Interface and Google Interface.

The SUS mark in Google Interface was estimated as 73.5, which is more than the average of 68, and the SUS mark in Bing Interface was 66.5, which is less than the average of 68. The results of the SUS are presented in Table III. In addition, the percentile rank recommended by Sauro in 2012 [40] was mapped to the measured SUS score, and the percentage was determined, as shown in Fig. 2. The SUS score of Bing is equivalent to 45% or grade C+, and the SUS score of Google Interface is equivalent to 69% or grade B+.

Following the score analysis recommended by Bangor et al. in 2009 [48], the usability of the Google Interface is valued as "Good", and the usability of the Bing Interface is valued as "Ok" (see Fig. 3). As a result, the usability level of the UI is considered to be difficult. However, it may not be said that the usability of the interface is poor.

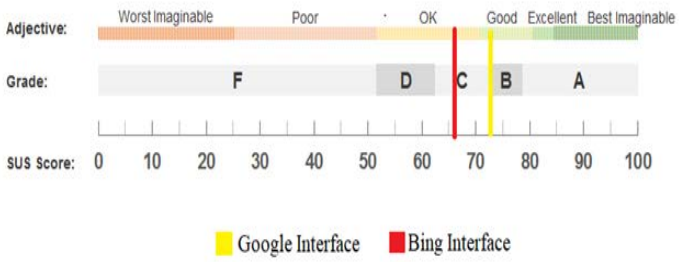


Fig. 3. Mapping the SUS Score on the Bangor & AL (2009) Interpretation Diagram.

According to Lewis and Sauro in 2009 [49], the learnability score of a system can be calculated individually from usability as the second aspect of SUS, as presented in Fig. 4.

Learnability = (item 10 + item 4) * 12.5

Usability = sum of Item (1,2,3,5,6,7,8,9) *3.125

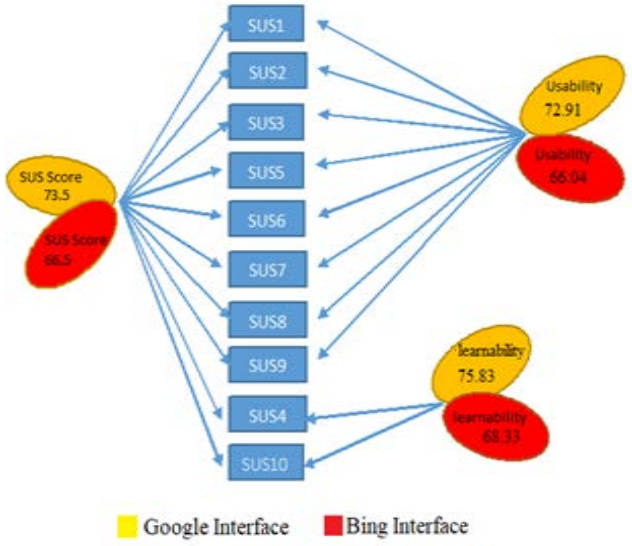


Fig. 4. Illustrate Usability, SUS Score and Learnability.

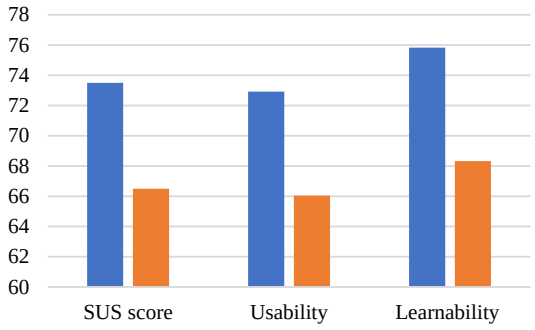


Fig. 5. Google Interface and Bing Interface and their Usability and Learnability Differences.

According to the collected data, the Google UI has a learnability score of 75.83, whereas the Bing UI has 68.33. They are remarkably low than the usability score in both interfaces, which was 72.91 and 66.04, as shown in Fig. 4. As a result, the usability level for the Bing interface is lower than the learnability level. It is rated as OK, whereas the usability level for the Google interface is virtually Good. Clearly, greater thought should be given to the factors that influence usability during the design process.

From Fig. 5, it is observed that the usability score of the Google interface is greater than that of the Bing interface, with the mean value of 72.91 and 66.04, respectively. This shows that the Google interface is more user friendly for elderly people compared with the Bing interface.

The paired t-test (often called the paired-samples t-test) compares the average of two related groups to see if they differ statistically significantly [45]. The results of the paired-samples t-test obtained using SPSS software are shown in Tables IV, Table V, and Table VI.

TABLE IV. GOOGLE AND BING SUS PAIRED SAMPLES STATISTIC

Pair 1	**Mean**	**N**	**Std. Deviation**	**Std. Error Mean**
Google_SUS	73.5000	15	5.49350	1.41842
Bing_SUS	66.5000	15	10.55597	2.72554

TABLE V. GOOGLE AND BING SUS PAIRED SAMPLES CORRELATION

Pair 1	**N**	**Correlation**	**Sig.**
Google_SUS & Bing_SUS	15	.696	.004

TABLE VI. GOOGLE AND BING SUS PAIRED SAMPLES TEST

Pair 1	***Mean***	**Paired Differences**				**t**	***df***	***Sig. (2-tailed)***
		Std. Deviation	***Std. Error Mean***	**95% Confidence Interval of the Difference**				
				Lower	***Upper***			
SUS of Google & Bing	7.0	7.80339	2.01483	2.67863	11.32137	3.474	14	.004

A test statistic (t) is equal to 3.474. The p-value is used to determine whether this is significant or not. Because this is a two-sided test, the result is significant at $p < .05$.

In this study, the p-value is = 0.004 and the p-value is <0 .05. This value is too small, which mean that there is less than one in a billion chances that the means SUS scores are equal to each other. We can also conclude that it can be over 99.6% sure that the Google and Bing interfaces have different SUS scores. Google interface's SUS score of 73.5 is statistically higher than Bing interfaces of 66.50. Also, the Google interface shows a reduced standard mean error rate of 1.41842 compared to the Bing interface (2.72554). So, it can be concluded that elderly users perceived the Google interface as easier to use than the Bing interface.

V. Discussion

In both Google and Bing, the least experienced users had significantly more usability issues when using the search engines. This shows that practice with search engines is advantageous, if not essential, for efficient search (the same is true for young; see Pollock and Hockley) [50].

In Google, the most frequent ease of use issue was with the usage of smaller fonts for buttons and explanation as the participants were not able to focus and understand the text written to bring the mouse over and click them properly. The user can change the size of the text on well-designed websites on their own. However, without resizing the entire page, which can be done using the browser's zoom, the text boxes and the font used inside the boxes cannot be resized well. As a result, to make searching easier for people with low vision, text boxes and fonts should be larger or, at the very least, resized as the text size is increased. It would be easier for elderly users to modify their queries if the font was larger and the gaps between letters were wider.

Another reason for the smaller font size concerns in Google and Bing is unquestionably the participants' lack of knowledge with the keyboard (e.g., both the Backspace button and Delete button confused participants) and the mouse (e.g., keeping the mouse still while clicking). Nonetheless, age-related psychomotor issues additionally added to this issue by making it harder to perform fine motor movements, particularly when the target text or button is small (e.g., when attempting to situate the cursor between two letters). Bigger textual style, bigger letter spacing and "clear content-box" buttons may help the elderly participants better as these would help focus them on the target with the cursor easily. Besides that, adding a clear button to clear the text box query with one click may help them to reduce using Backspace and Delete buttons.

Similar to Aula and kaki, many participants often had problems due to a lack of knowledge of the web's structure [19, 26]. Even though the participants understood the basic working of the back button, they could not get back to the first search page after following the result pages one after another. This outcome demonstrates that the Internet browser interfaces may not be as natural as Kubeck et al. [51] introduced. Some pre-planning may be necessary to make the initial browsing experience less perplexing. Opening the results in a new web browser or in browsers that allow tabbed browsers to open in new tabs might remedy the issues stated above. This technique keeps the main results page of the web search, allowing users to view back the results from there.

Apart from this, the many advanced functionalities of Google and Bing, like advanced search, search tools, language setting etc., confused the participants. Thus, aiming to provide older adults with a simple, easy–to–use interface would be more beneficial. Advanced searching can be thought of as an option for experienced users. Some participants accidentally clicked it and were confused with its working and couldn't get back to the basic search page without assistance. Thus, it's better to altogether avoid the advanced search option. However, the language option seems more important for those users who are not well acquainted with English, and so it's recommended to have the language option.

Participants in both Google and Bing frequently neglect to focus and click into the search text box before typing in the search text box, which is a widespread problem. The only visible difference between having the cursor on top of the search text box and really having the focus on the text box is whether or not the cursor blinks. You can type text into the search text box if it blinks. This indication is clearly insufficient for people with reduced vision as the elderly are more concentrated on the keyboard while typing than on the screen. To alleviate this problem when the search text box is off–focus, Aula et al. made the focused feedback increasingly via greying the search text box and the text and provided a thick border and black highlighted text whenever the focus was in the search text box [26]. When the user types without focusing on the search text box, an attention sound is recommended. Furthermore, some volunteers placed the mouse cursor on top of the search text box before attempting to edit the search query but then failed to press the click to move the focus to the search text box. Thus, the suggestion is similar to Aula et al. [26], which to grey the search text box and the button when the cursor is not in the text box and also to have a tooltip with the text "Click the mouse button to insert text" when the mouse pointer is above the text box.

In many cases, it was seen that the users started searching their queries within the resultant websites without even realizing that they had left the search engine. As an outcome, they came up with better new queries inside the site search but received no results. After being indicated to attempt the inquiry again utilizing a search engine, they could find the results. Thus, with better training regarding the web search engine, the participants would probably be more at ease in using the search engine.

In Google and Bing, information related to the previously visited search results is saved at the history menu and also bookmarked results are saved in the browser. Thus, participants faced difficulties going back to previously visited results and bookmarked results in different browsers. Because each browser has its way to bookmark results and save the history browsing, it would be better to design Bookmark and History menus in the main page interface as part of interface elements to solve this confusion. So whatever browser was used, all UI elements is still the same, and the bookmark and history menus will be loaded with when opened by any browser.

Occasionally, information from a previous search was mixed with information from a recent Google and Bing engine search. Thus, after typing in a new query and noticing a wrong result list, participants wanted to get back to the original search page, but they did not know how to. It would be better to provide users with a clear starting point for new searches to alleviate this confusion. Thus, a separate "Begin new search" button is recommended to clear all the information about the previous searches [19, 26] and provide a natural starting point for new search tasks.

VI. Suggested Solution

This study discovered that a simple design tends to make the search experience for elderly people less troublesome and more manageable. Many participants also expressed that search interfaces with all their different colours seemed messy and complex. It's recommended that the improved search engine for elders use colours sparingly.

The goal of the improved search engine interface is not to displace more complicated systems, but to provide elderly people with the option of using a basic search engine, personalized and adaptable interface and easy–to–use interface for using the web for their daily Internet searching to overcoming their weak intellectual and physical abilities, as well as their cognitive abilities.

Table VII presents suggested improvements to design a usable web search user interface for elderly users based on the elderly users' major problems during interacting with both interfaces.

The suggested elderly web search interface for the elderly community is based on the finding from the experimental results of Both Google and Being interface from the elderly perspective and previous studies. In addition to the original features of existing search interfaces, the new web search user interface for elderly users allows customization towards the users' wishes. According to Peter Brusilovsky and Maybury [52, 53], the adaptation of user interface for web applications could be in the areas given below:

Content selection: The presentation of contents in the user interface to users.

1) Information presentation: Visual presentation of every piece of information is in the interface.

2) Concepts of navigation: Navigation through the user interface by the users for gathering their desired information given in the webpage.

In order to achieve a coherent design between different variations, we suggest fixing the positions of different SUI parts. Fig. 6 depicts the general structure of the suggested search UI for the elderly. It consists of eight groups of elements: a setting section, a help section, login and logout, profile image, a menu for saved bookmarked results with different categories, a history section menu, a theme mode button, the main search results section. The search input consists of elements: text, voice, image with a larger button size and label. Besides adding a clear button to clean text search "query" and begin a new search. A bookmarked menu and history menu provide direct links for the elderly to access certain saved results for ease of use and navigation. A menus support the elderly to save time and effort instead of researching again for the same results and reduce the memory load of the elderly.

TABLE VII. Suggested Design Solutions based on the Identified Usability Problems

Usability Problems	**Design Solutions**
The default font size is still small.	The default font size is large, and it is constant in the interface without any change, even if using a different browser with ease of modification.
The voice search button, the image search button and the text search button are small and close to each other.	They should be separated from each other with enough space and recognized with a button. Also, the button should be having a recognized name or logo to show the differences between them.
The home page is long	Reduce the length of the home page
Scrolling a page is a difficult task	Reduce the length of the home page at most one screen and half
Unclear to go back to the home page	Make the results open in a new window and that can be clear to go back to the main home page.
Lots of different colours are on the page background.	The background is unchangeable and to be white or with clear color with font dark.
The main and result page view is slightly unstructured	Restructure it to show all UI elements on the main page
The setting menu is complex.	Reduce the elements of the setting that do not affect the main job of the interface. Thus, the simpler menus are healthier choices, such as the Classic navigation menu and Sidebar menu.
Difficult to go back to previous open results and bookmarked results in different browsers. That is because each browser has its way to bookmark results and save the history browsing.	Bookmark and result History should be put in the main home interface. So whatever browser was used, all UI elements is still the same, and the bookmark and history will be loaded with when opened by any browser.
A large number of results	Reduce the view search results on the screen to be 5 to 7 on each screen.
Misunderstanding suggested results such as People Also Search For and Related Searches.	Remove it

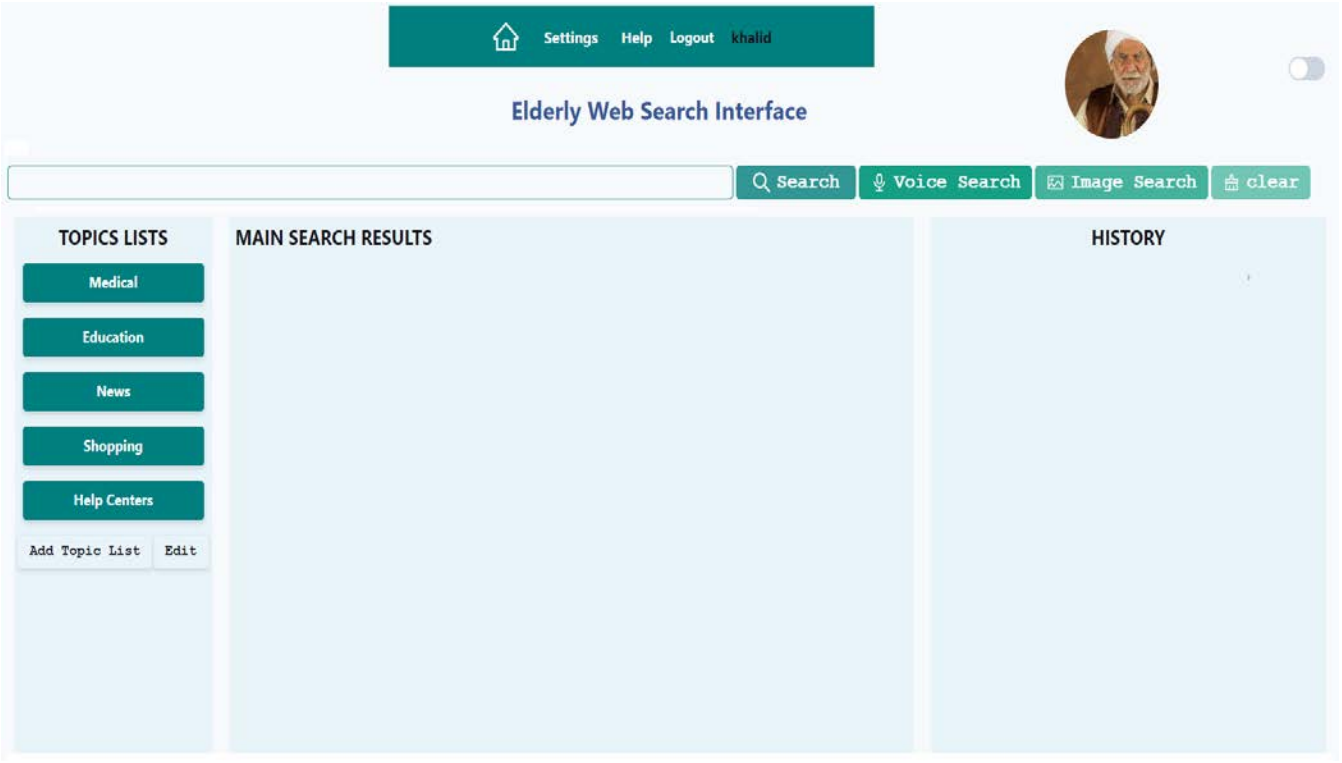


Fig. 6. The General Structure of the Suggested Search user Interface for Elderly.

VII. Conclusion

This research conducted usability evaluations of Google and Bing interfaces from the perspectives of elderly users and suggested how developers can better improve search engine interfaces to make them more usable to elderly users. Applying an observational evaluation technique and SUS questionnaire in this research was shown to be effective in identifying a large proportion of usability problems elderly users faced while interacting with web search UIs. The results show that the Google interface is much better than the Bing interface because it is more user-friendly and simplified to use. It has a larger button size and font size, reduced home page length, larger letter spacing, and visible content-box buttons that would assist the elderly better since it helps them easily focus on the target through the cursor. Also, the Google interface is more suitable for the intellectual ability of the elderly that includes their memory and learning capacity, technical knowledge and experience in using the computers. Moreover, it shows good usability and learnability with a higher SUS score than the Bing user interface. Therefore, the usability level is lower than the learnability level for the Google Interface and the Bing interface, and it is obvious that it should be paid more attention to the elements influencing usability during the design process.

Simple interfaces and simple result pages benefit elderly users by reducing issues produced by not understanding what is happening, reducing the total number of features to be learned, and, most importantly, making the users feel in control of the situation. Aside from that, it is recommended that a "clear text box query" button be added to avoid misunderstandings in using the Backspace and Delete buttons. Furthermore, the needs of senior users varied. Therefore interfaces should be adaptive and personalized for each user to match their individual requirements.

In light of the findings of this study, specifically the design solutions, a prototype of a web search user interface will be designed as future work. The prototype will be put through its paces with elderly volunteers. The elderly volunteers' interactions with the prototype while executing various tasks and their feedback from a think-aloud helped develop the prototype. The improved prototype can then be used to create a commercially viable web search user interfaces launcher.

Acknowledgment

Malaysia's Ministry of Higher Education (MOHE), Libya's Ministry of Higher Education, and the University of Gharyan all contributed to this work. We would like to thank the Universiti Teknologi Malaysia (MOHE) and the UTM VicubeLab research group.

References

[1] Allah, Khalid Krayz, Nor Azman Ismail, and Mohamad Almgerbi. "Designing web search UI for the elderly community: a systematic literature review." Journal of Ambient Intelligence and Humanized Computing, 1-25, 2021.

[2] Sanchiz, Mylene, Jessie Chin, Aline Chevalier, Wai-Tat Fu, Franck Amadieu, and Jibo He. "Searching for information on the Web: Impact of cognitive aging, prior domain knowledge and complexity of the search problems." Information Processing & Management 53, no. 1, 281-294, 2017.

[3] Wagner, N., Hassanein, K., & Head, M. The impact of age on website usability. Computers in Human Behavior, 37, 270–282. https://doi.org/10.1016/j.chb.2014.05.003, 2014.

[4] Sanchiz, M., et al. "User-friendly search interface for older adults: supporting search goal refreshing in working memory to improve

information search strategies." Behaviour & Information Technology, 1-16, 2019.

[5] Miyake, Asuka, Yuji Morinishi, and Masahiro Watanabe. "Estimation Models of User Skills Based on Web Search Logs." International Conference on Human-Computer Interaction. Springer, Cham, 2016.

[6] Salman, Hasanin Mohammed, Wan Fatimah Wan Ahmad, and Suziah Sulaiman. "Usability Evaluation of the Smartphone User Interface in Supporting Elderly Users from Experts' Perspective." Ieee Access 6: 22578–91, 2018.

[7] Fisk, Arthur D., Sara J. Czaja, Wendy A. Rogers, Neil Charness, and Joseph Sharit. Designing for older adults: Principles and creative human factors approaches. CRC press, 2020.

[8] Sanchiz, Mylene, et al. "Searching for information on the web: Impact of cognitive aging, prior domain knowledge and complexity of the search problems." Information Processing & Management 53.1 281-294, 2017.

[9] Chevalier, Aline, Aurélie Dommes, and Jean-Claude Marquié. "Strategy and accuracy during information search on the Web: Effects of age and complexity of the search questions." Computers in Human Behavior 53 (2015): 305-315, 2015.

[10] Chin, Jessie, and Wai-Tat Fu. "Interactive effects of age and interface differences on search strategies and performance." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 403-412. 2010.

[11] Nivala, Annu-Maaria, Stephen Brewster, and L. Tiina Sarjakoski. "Usability Evaluation of Web Mapping Sites." In Landmarks in Mapping, pp. 239-256. Routledge, 2014.

[12] Barnum, Carol M. Usability testing essentials: ready, set... test!. Morgan Kaufmann, 2020.

[13] Chiew, Thiam Kian. "A systematic literature review of the design approach and usability evaluation of the pain management mobile applications." Symmetry 11, no. 3, 400, 2019.

[14] Bačíková, Michaela, Jaroslav Porubän, Matúš Sulír, Sergej Chodarev, William Steingartner, and Matej Madeja. "Domain Usability Evaluation." Electronics 10, no. 16 (2021): 1963, 2021.

[15] Ivory, Melody Y., and Marti A. Hearst. "The state of the art in automating usability evaluation of user interfaces." ACM Computing Surveys (CSUR) 33, no. 4, 470-516, 2001.

[16] Jessie Chin and Wai-Tat Fu, "Interactive Effects of Age and Interface Differences on Search Strategies and Performance." ACM, 2010.

[17] Boot, W., Charness, N., Czaja, S.J. and Rogers, W.A. Designing for older adults: Case studies, methods, and tools. CRC Press, 2020.

[18] Sauro, Jeff, and James R. Lewis. 2011. "When Designing Usability Questionnaires, Does It Hurt to Be Positive?" Conference on Human Factors in Computing Systems - Proceedings, 2215–23, 2011.

[19] Aula, Anne. "User study on older adults' use of the Web and search engines." Universal Access in the Information Society 4, no. 1, 67-81, 2005.

[20] Kobayashi, M., Hiyama, A., Miura, T., Asakawa, C., Hirose, M. and Ifukube, T. September. "Elderly user evaluation of mobile touchscreen interactions". In IFIP conference on human-computer interaction (pp. 83-99). Springer, Berlin, Heidelberg, 2011.

[21] Dolničar, V., Šetinc, M. and Petrovčič, A. "Toward an age-friendly design of smartphone interfaces: The usability test of a launcher for older adults ". Uporabna Informatika, XXIV, 24, pp.4-15, 2016.

[22] Salman, H.M., Ahmad, W.F.W. and Sulaiman, S., "Usability evaluation of the smartphone user interface in supporting elderly users from experts' perspective " Ieee Access, 6, pp.22578-22591, 2018.

[23] Di Nuovo, A., Broz, F., Belpaeme, T., Cangelosi, A., Cavallo, F., Esposito, R. and Dario, P., October. VA web based multi-modal interface for elderly users of the robot-era multi-robot service ". In 2014 IEEE international conference on Systems, Man, and Cybernetics (SMC) (pp. 2186-2191). IEEE, 2014.

[24] Patsoule, E. and Koutsabasis, P., "Redesigning websites for older adults ": a case study. Behaviour & Information Technology, 33(6), pp.561-573, 2014.

[25] Haesner, M., Wolf, S., Steinert, A. and Steinhagen-Thiessen, E. "Touch interaction with Google Glass–Is it suitable for older adults? ". International Journal of Human-Computer Studies, 110, pp.12-20, 2018.

[26] Aula, A. and Käki, M. "Less is more in Web search interfaces for older adults ". First Monday, 2005.

[27] Cornet, V.P., Daley, C.N., Srinivas, P. and Holden, R.J., September. "User-centered evaluations with older adults: testing the usability of a mobile health system for heart failure self-management ". In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 61, No. 1, pp. 6-10). Sage CA: Los Angeles, CA: SAGE Publications, 2017.

[28] PREECE, J., ROGERS, Y., & PREECE, J. "Interaction design: beyond human-computer interaction ". Chichester, Wiley, 2019.

[29] Joe, J., Chaudhuri, S., Le, T., Thompson, H. and Demiris, G. "The use of think-aloud and instant data analysis in evaluation research: Exemplar and lessons learned ". Journal of biomedical informatics, 56, pp.284-291, 2015.

[30] J. R. Lewis, "Usability: Lessons Learned. and Yet to Be Learned," Int. J. Hum. Comput. Interact., vol. 30, no. 9, pp. 663–684, 2014, doi: 10.1080/10447318.2014.930311.

[31] P. Corti, B. Lewis, and A. T. Kralidis, "Hypermap registry: an open source, standards-based geospatial registry and search platform," Open Geospatial Data, Softw. Stand., vol. 3, no. 1, p. 8, Dec. 2018, doi: 10.1186/s40965-018-0051-x.

[32] D. Pal and V. Vanijja, "Perceived usability evaluation of Microsoft Teams as an online learning platform during COVID-19 using system usability scale and technology acceptance model in India," Child. Youth Serv. Rev., vol. 119, p. 105535, 2020, doi: 10.1016/j.childyouth.2020.105535.

[33] J. Sauro and J. R. Lewis, "Correlations among prototypical usability metrics: Evidence for the construct of usability," Conf. Hum. Factors Comput. Syst. - Proc., no. August, pp. 1609–1618, 2009, doi: 10.1145/1518701.1518947.

[34] M. K. Othman, A. Nogoibaeva, L. S. Leong, and M. H. Barawi, "Usability evaluation of a virtual reality smartphone app for a living museum," Univers. Access Inf. Soc., no. 0123456789, pp. 9–13, 2021, doi: 10.1007/s10209-021-00820-4.

[35] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," Int. J. Hum. Comput. Interact., vol. 24, no. 6, pp. 574–594, 2008, doi: 10.1080/10447310802205776.

[36] A. Bangor, T. Staff, P. Kortum, J. Miller, and T. Staff, "Determining what individual SUS scores mean: adding an adjective rating scale," J. usability Stud., vol. 4, no. 3, pp. 114–123, 2009.

[37] J. R. Lewis Senior HF Engineer and J. Sauro, "Revisiting the Factor Structure of the System Usability Scale," J. Usability Stud., vol. 12, no. 4, pp. 183–192, 2017.

[38] S. Borsci, S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci, "Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience," Int. J. Hum. Comput. Interact., vol. 31, no. 8, pp. 484–495, 2015, doi: 10.1080/10447318.2015.1064648.

[39] T. S. Tullis and J. N. Stetson, "A Comparison of Questionnaires for Assessing Website Usability ABSTRACT : Introduction," Usability Prof. Assoc. Conf., pp. 1–12, 2004, [Online]. Available: http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf.

[40] Jeff Sauro and James R. Lewis. "Standardized Usability Questionnaires. Quantifying the User Experience" , 2012. https://doi.org/10.1016/b978-0-12-384968-7.00008-4.

[41] Rot, A., Kutera, R., Gryncewicz, W. Design and assessment of user interface optimized for elderly people. A case study of actgo-gate platform, in: ICT4AWE 2017 - Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health. SciTePress, pp. 157–163, 2017. doi:10.5220/0006320001570163.

[42] Cioara, T., Anghel, I., Valea, D., Salomie, I., Martin, V.S., Marchena, A.G., Jimeno, E. and Vastenburg, M. Adaptive workspace interface for facilitating the knowledge transfer from retired elders to start-up companies. In Ambient Assisted Living and Enhanced Living Environments (pp. 287-309). Butterworth-Heinemann, 2017.

[43] Moran, K. and Goray, C. "Complex Search-Results Pages Change Search Behavior: The Pinball Pattern ", 2019.

[44] Lewis Senior HF Engineer, James R, and Jeff Sauro. "Revisiting the Factor Structure of the System Usability Scale." Journal of Usability Studies 12 (November): 183–92, 2017.

[45] auro, Jeff, and James R Lewis. Quantifying the User Experience, Second Edition: Practical Statistics for User Research. ACM SIGSOFT Software Engineering Notes, 2016.

[46] Finstad, Kraig. "The system usability scale and non-native English speakers." Journal of usability studies 1, no. 4: 185-188, 2006.

[47] Brooke, John. 1996. "SUS: A 'Quick and Dirty' Usability Scale." Usability Evaluation In Industry, no. November 1995: 207–12, 2006. https://doi.org/10.1201/9781498710411-35.

[48] Bangor, Aaron, Technical Staff, Philip Kortum, James Miller, and Technical Staff. "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale." Journal of Usability Studies 4 (3): 114–23, 2009.

[49] Lewis, James R., and Jeff Sauro. "The Factor Structure of the System Usability Scale." Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5619 LNCS (August): 94–103, 2009, https://doi.org/10.1007/978-3-642-02806-9_12.

[50] Pollock, A. and Hockley, A. What's wrong with Internet searching. D-lib magazine, 3(3), pp.1-5, 1997.

[51] Kubeck, J.E., FINDING INFORMATION ON THE WORLD WIDE WEB: EXPLORING OLDER ADULTS'EXPLORATION. Educational Gerontology, 25(2), pp.167-183, 1999.

[52] Brusilovsky, Peter, and Mark T. Maybury. "From adaptive hypermedia to the adaptive web." Communications of the ACM 45, no. 5: 30-33, 2002.

[53] Heumader, Peter, Klaus Miesenberger, and Tomas Murillo-Morales. "Adaptive User Interfaces for People with Cognitive Disabilities within the Easy Reading Framework." In International Conference on Computers Helping People with Special Needs, pp. 53-60. Springer, Cham, 2020.

A Novel Framework for Cloud based Virtual Machine Security by Change Management using Machine

S.Radharani, V.B.Narasimha
Department of CSE, UCE, Osmania University
Hyderabad, India

***Abstract*—The increased growth in the cloud-based application development and hosting, the demand for higher application and data security is also increasing. The cloud-based applications are hosted on virtual machines and the data generated or used by these applications are also hosted inside the virtual machines. Hence, the security of the applications and the data can be achieved only by securing the virtual machines. There are number of challenges to achieve the security of the virtual machines. Firstly, the size of the virtual machines is large, and the generic cryptographic methods are primarily designed to handle smaller size of the data. Thus, the applicability of these methods for virtual machine are subjected to analysis. Secondly, the additional time required for applying the cryptographic algorithms on the virtual machines impact the response time of the applications, which again impacts the service level agreements. Finally, the virtual machines during the migration are highly vulnerable as the virtual machines are migrated inside the data center networks as simple text data. A good number of research attempts have tried to solve these challenges. Nonetheless, most of the parallel research works have either compromised on the strength of the security protocols or have compromised on the time taken to apply the cryptographic methods. However, the need of the research is to identify the attacks based on the characteristics of connection requests and reduce the time for the encryption and decryption of the virtual machines. This work proposes a novel framework for detection of the attacks based on a machine learning driven algorithm by analyzing the connection properties and prevent the attacks by selective encryption of the virtual machines using another machine learning driven algorithm. This work demonstrates nearly 98% accuracy in detection of the newer and existing attack types.**

***Keywords*—DevOps; deep clustering; VM security; cloud security; VM versioning; progression cryptography**

I. Introduction

The security for the cloud infrastructure has always been a persistent issue as most of the consumers and practitioners do not have clear understanding about the security factors and implementation details. To some extent, the service providers have made a closed loop about the knowledge of cloud security inside the organizations and sometimes only to the selective groups. This makes the deployment of the cloud-based security protocols even harder for the researchers. Nonetheless, the recent research outcomes by various research attempts are opening the closed loops of the knowledge and exploring the possibilities of the deployment of novel and higher performing security protocols. One such work presented by P. Mishra et al. [1]. Nevertheless, the challenges of cloud securities are not only restricted to the data stored on the cloud. Rather, the security challenges can be observed in all the layers of cloud implementations as on the infrastructure layer, platform layer and the services layer. Another work by P. Mishra et al. [2] have confirms this claim. Thus, deploying the security protocol for all the layers of cloud implementation is highly complex due to various aspects such as model complexity or compatibility or interoperability between the layers.

Henceforth, the implementation of the cloud security protocols can be best implemented using the virtual machine architectures. As the virtual machines holds the applications core and the data, generated or consumed by these applications, hence protecting the virtual machines must be the primary concern, which is implemented in this work. This work identifies the challenges of cloud security, internally which is virtual machine security and proposes a machine learning driven framework to protect the VMs.

II. Parallel Research Outcomes

The security of the cloud-based applications is critical as mentioned earlier. The applications and the data on the cloud are visible to authenticated and unauthenticated parties at the same time. Though, the access and identity management aspects of the online access can restrict the privileges on the applications and the data. Nonetheless, the visibility of data cannot be restricted. Hence, the possibilities of the attacks also increase on such data. The work by M. R. Watson et al. [3] have clearly listed the vulnerabilities on the cloud systems and also produced a clear guideline for managing the security. Considering the similar directions, to produce a framework for detection of the attacks based on characteristics, yet another work by V. Varadharajan et al. [4] can be highlighted. These parallel research outcomes are primarily focused on an old framework called Rekall [5] and the produced recent outcomes are the attempts to reduce the complexity and at the same time increasing the responsiveness of the same outlined characteristics. These outcomes have mainly concentrated on the prevention of the attacks.

In the other hand, the domains for attack detections are also very popular among researchers. The work by T. K. Lengyel et al. [6] have clearly listed the possibilities of the attack analysis frameworks to detect the attacks. Nonetheless, these detection processes can be highly complex for the distributed architectures such as cloud or fog or edge-based computing. The application, the data and the userbases are always distributed and most of the times, the execution is parallel. Hence, the protective framework must also comply with the

distributed nature of the architecture. The work by S. Gupta et al. [7] have confirmed to this believe.

The attacks are not only restricted to platform and the service layers. Multiple attacks are also reported on the physical hardware devices. The immediate but costly solution is to provide the hardware security modules or the HSM devices. Nonetheless, as mentioned these solutions are costly and for a cloud-based architecture, the applicability of the HSMs is very limited due to the limited physical access to the infrastructure. The work by D. Kirat et al. [8] have spoken in favor of this statement and confirms the claim. Although, the analysis of the intrusion or attack detections must take place at all the layers of cloud computing and infrastructure layer is not an exception. The work by C. Spensky et al. [9] have elaborated on the possibilities and feasibilities of monitoring for the attack detections on the physical infrastructure layer. This work has been criticized for not considering the possibilities for remote monitoring, which can be achieved using the access to the virtual machines. In the recent times, a good number of virtual machine managers have incorporated the monitoring layers in the VMM structure.

Reciting back to the monitoring of the virtual machines for attack detection and prevention methods have improved a lot using the virtual machine monitoring possibilities. The survey done by F. Cai et al. [10] confirms few claims directly and indirectly as firstly, the deployment of the security features can be best adopted on the virtual machines. Secondly, the existing cryptographic methods can easily be outperformed in the recent higher demand for best response times and finally, the newer types of the attacks are increasing day by day and a method for detecting the attacks based on the behavior must be adopted. Thus, the demand for the automated framework with these features is the demand of the current research as also demonstrated in the work by A. Almtrf et al. [11].

The primary features of the expected framework must comply with few additional characteristics. The first characteristics is the close association with the software and the hardware modules to track the flow of the application processing characteristics as rightly stated in the work by A. Khurshid et al. [12]. The second characteristics of the proposed framework is to track the changing nature of the data as mentioned in the work by N. E. Moussaid et al. [13]. The final characteristics must comply with the deployed virtual machine-based applications hosted on the cloud platforms as suggested by X. Lu et al. [14]. Thus, this work considers all the recommendations from the parallel research outcomes and further produces the proposed framework for detection and prevention of the attacks on the cloud application, in tern the virtual machine security.

Further, this work realizes the characteristic based detection of the attacks. This not only identifies the known attack types, but also identifies the newer attack types. The work by B. Sudhakar et al. [15] has clearly listed the attack types and the mapping to the connection properties. The conclusive mapping from this work is furnished here [Table I].

TABLE I. ATTACK TYPES AND CONNECTION PROPERTIES MAPPING [15]

Attack Type	**Connection Properties**
Browser Based Attacks	1. Count of the connection requests 2. Access Type Requests
Brute Force Based Attacks	1. Count of the connection requests 2. Ratio between the request and responses
DoS Based Attacks	1. Access Type Requests 2. Service Request Types 3. The rate of change in the service request types
SSL Based Attacks	1. Service Request Types 2. The rate of change in the service request types
Scan Based Attacks	1. Ratio between the request and responses
DNS Based Attacks	1. Service Request Types

It is worth the mention, that these all characteristics or connection properties are available in the KDD dataset [16].

Thus, in the next section of this work, the problem identified in this section in the parallel research outcomes is formulated using mathematical models.

III. PROBLEM FORMULATION & PROPOSED SOLUTIONS

After the fundamental understanding of the research problems in the previous section of this work, this section focuses and elaborates the core problems and proposes solutions to these problems using the mathematical modeling techniques.

The first problem elaborates on the responsiveness of the cloud-based applications due to the adaptation of the attack detection methods. The parallel research outcomes, as seen in the previous sections, shows higher time complexity. The increased time complexity is due to the nature of analysis deployed by these algorithms, which primarily focus on large number of characteristics or the connection properties. Hence, this must be resolved.

Lemma – 1: The reduction of the connection characteristics using the correlation method can reduce the time complexity of the detection method.

Proof: The connections characteristics or the properties extracted from the connection requests can be a very large dataset because of multiple monitoring system. Many of the times, these large datasets provide limited and redundant information, which is again at the cost of higher time complexity. Thus, a machine learning driven process to reduce the number of characteristics can certainly reduce the time complexity.

Assuming that, the set of connection properties, C[], is a collection of multiple characteristics and each characteristics can be identified as C_i. Thus, for n number of total characteristics, the relationship can be formulated as:

$$C[] = \langle C_1, C_2, C_3 ...., C_n \rangle \quad (1)$$

Also, assuming that, C_X is the class variable, which defines the nature of the connection in terms of attacks or normal from the historical information sets.

Hence, the characteristics analysis for detection of the attacks using the standard algorithms can be formulated as,

$$TH_i = \Phi(\exists C_i) : \prod_{RowID=C_i} C[] \tag{2}$$

Here, Φ is the function for extracting the threshold and further, the threshold for attribute C_i is stored in TH_i. Clearly, the threshold must be calculated relatively as with the consideration of the other parameters.

Further, the combined information from the thresholds from all the characteristics can decide the nature of the connection in terms of the class variable as,

$$C_X = \sum_{i=1}^{n} TH_i \tag{3}$$

It is natural to realize that due to Eq. 2 and Eq. 3, the time complexity, T_1, can be formulated as,

$$T_1 = n(n-1) \tag{4}$$

Or,

$$T_1 = n * n = O(n^2) \tag{5}$$

For a large dataset with 100s of parameters or characteristics, this time complexity for detection of the attacks can be very high. Thus, this problem must be solved using parameter reduction process.

Thus, based on the Eq. 1, the correlation formulation can be formulated as,

$$\rho(C_x, C_i) = \frac{(\eta(C[x]) - C_x).(\eta(C[i]) - C_i)}{\sigma C_x . \sigma C_i} \tag{6}$$

Here, ρ defines the correlation value or correlation coefficient, η defines the mean value and σ defines the standard deviation.

The standard deviation calculation can be formulated as,

$$\sigma C_i = \sqrt{\frac{\sum \{C_i - \eta(C[i])\}^2}{n}} \tag{7}$$

Further, the total correlation sets can be stored in Corr[] and the highest values can be taken to identify m number of characteristics for final analysis as,

$$Corr[] =< \exists \rho(C_x, C_i) > \tag{8}$$

And,

$$m :\rightarrow Corr[] \tag{9}$$

Thus, in the light of Eq. 4, the new time complexity, T_2, can be formulated as,

$$T_2 = m(m-1) \tag{10}$$

Or,

$$T_2 = m * m = O(m^2) \tag{11}$$

As, $m << n$, thus it is conclusive to state that

$$T_2 << T_1 \tag{12}$$

Thus, reduction of the time complexity using the attribute reduction method is highly feasible.

The second problem elaborates on the detection of the attack types. The attack types can be identified using a cluster analysis on the connection characteristics or the properties. As seen in the previous section of the work, the parallel research outcomes mostly fail to detect the newer attacks, though the types of the attacks are not very new and have a strong similarity with the existing and known types of attacks.

This problem can be solved using deep cluster technique. The clustering method for identification of the attacks is significant as the identification of attacks direct towards anomalies in the connection, which is easily identifiable as outliers using the clustering method.

Lemma – 2: The deep clustering method can identify the newer types of attacks using the outlier identification method.

Proof: The outliers as a result of clustering process identifies the anomalies using various characteristics and similarities of the characteristics domain values. Based on the nature of the data used in the clustering process, the outliers can define various meanings. As in this research the data used are the connection characteristics, hence the outliers will denote the abnormal connections or the attacks.

Continuing and revising the Eq. 1, for all the characteristics, there must be domains for each characteristic as,

$$C[][] = \langle C_1[], C_2[], C_3[]...., C_n[] \rangle \tag{13}$$

Further, the clustering process must be performed initially for each and every characteristic or attribute domains as

$$CL_i[] \leftarrow \Phi(\exists C_i[]) \tag{14}$$

Here, the set of clusters for the i^{th} attribute will be stored in $CL_i[]$ and Φ denotes the clustering process.

Henceforth, the number of members in each cluster must be validated and the cluster with the lowest number of members are the potential clusters, inside which the outliers will reside.

Thus, the iterative clustering must be performed until the outliers, in this case the attacks, is not identified as,

$$\omega \leftarrow \Phi(\exists \left| CL_i[] \right|_{low}) \quad (15)$$

The terminating condition for Eq. 15 iteration is $\omega \to 1$.

Henceforth, it can be stated conclusively, the minute deviations can be identified using this proposed method and further any new attack can also be detected, which has very little similarity to the existing attack types.

The final problem, which this research aims to solve is the reduction of the cryptographic algorithm implementation time. As seen in the previous section of this work, the cryptographic algorithms are not designed to handle the large data, which is case of virtual machine files are very large in volume. Also, due to the higher adaptability of the DevOps processes across all organizations for application development, the changes made to the application and indirectly to the virtual machines are very high. This makes the process of applying cryptographic algorithms further difficult.

Henceforth, the solution is to track the changes made to the virtual machines in terms of application code and data and apply incremental encryption process to reduce the time.

The proposed solutions are converted to algorithms, which are furnished in the next section of this work.

IV. Proposed Algorithms and Framework

After the formulation of the concepts of solutions in the previous section, in this section of the work, the proposed algorithms and the proposed frameworks are furnished.

Firstly, the Connection Characteristics Reduction using Correlation Analysis algorithm is furnished.

Algorithm - I: Connection Characteristics Reduction using Correlation Analysis (**CCR-CA**) Algorithm

Input:
Connection Characteristics set as CS[]

Output:
Reduced Characteristics set as RCS[]

Process:
Step - 1. Load the CS[] set
Step - 2. Mark the class characteristics as CX from CS[x]
Step - 3. For each attrbiute in CS[] as CS[i]
- a. Calculate the standard deviation, as SD[] using Eq. 7
- b. Calculate the correlation of CS[i] with CX as Corr[i] using Eq. 6

Step - 4. For each element in Corr[] as Corr[j]
- a. If Corr[j] Not Equal Corr[j+1] & Corr[j] is Max
  - i. Store RCS[j] = CS[i]
- b. Else,
  - i. Continue
- c. Corr[j] = Null
- d. Stop if Count(RCS[]) >= Count(CS[])/2

Step - 5. Return RCS[]

The above algorithm is framed to solve the first problem discussed and based on the proposed Lemma – 1.

Secondly, the Deep Clustering Based Attack Detection algorithm is furnished.

Algorithm - II: Deep Clustering Based Attack Detection (**DC-AD**) Algorithm

Input:
Reduced Characteristics set as RCS[][]
Output:
Detected Attacks as DS[]

Process:
Step - 1. Load the RCS[][] set
Step - 2. For each element in RCS[][] as RCS[i][]
- a. Apply K-Means Clustering on RCS[i][] and store the result in CL[i][] using Eq. 14

Step - 3. For each element in CL[][] as CL[j][]
- a. If count(CL[j][]) -> min(count(CL[][])
  - i. Apply K-Means Clustering on CL[j][] and store the result in CL1[i][] using Eq. 15
  - ii. Repeat the process untill Count(CL1[i][]) -> 1
  - iii. Identify the attack characteristics as DS[k] = RCS[i]
- b. Else,
  - i. Continue

Step - 4. Return DS[]

The above algorithm is framed to solve the second problem discussed and based on the proposed Lemma – 2.

Thirdly, the Random Crypto Key Generation algorithm is furnished.

Algorithm - III: Random Crypto Key Generation (**RCKG**) Algorithm

Input:
Large Random numbers as P & Q

Output:
I. Public Key as PK
II. Private Key as PKK

Process:
Step - 1. Calculate the modulus, M as M = P * Q
Step - 2. Select the derived encryption factor, DE as DE>1 and DE<(P-1).(Q-1)
Step - 3. Generate public key, PK as PK = (M,DE)
Step - 4. Generate private key, PKK as PKK = {1 MOD (P-1).(Q-1)}/DE
Step - 5. Return PK and PKK

Fourthly, the Progressive Virtual Machine Encryption using Change Detection algorithm is furnished.

Algorithm - IV: Progressive Virtual Machine Encryption using Change Detection (**PVME-CD**) Algorithm

Input:

I. Version Management of VM as VMS[]
II. Public Key as PK (M, DE)

Output:

Encrypted Virtual Machine as VME

Process:

Step - 1. Load the virtual machine versions as VMS[]
Step - 2. For each element in VMS[] as VMS[i]
  a. Configuration Management:
    i. Identify the import and include statements
    ii. Store the configuration management as CM[i]
  b. Data Management:
    i. Identify the variable in the code
    ii. Store the data management as DM[i]
  c. Life Cycle Management:
    i. Identify the loops and conditional statements
    ii. Store the life cycle management as LCM[i]
Step - 3. For each element in VMS[] as VMS[i]
  a. If CM[i] Not Equals to CM[i+1]
  b. Then, Store the changes CMC[j] = CM[i]-CM[i+1]
  c. If DM[i] Not Equals to DM[i+1]
  d. Then, Store the changes DMC[j] = DM[i]-DM[i+1]
  e. If LCM[i] Not Equals to LCM[i+1]
  f. Then, Store the changes LCMC[j] = LCM[i]-LCM[i+1]
  g. Merge the changed components as CC[i] = CMC[j] U DMC[j] U LCMC[j]
  h. Build the encrypted VMS[i] as VME = pow(CC[i],DE) mod M
Step - 4. Return VME

Fifthly & finally, the Progressive Virtual Machine Decryption using Change Detection algorithm is furnished.

Algorithm - V: Progressive Virtual Machine Decryption using Change Detection (**PVMD-CD**) Algorithm

Input:

I. Encrypted Virtual Machine as VME
II. Private Key as PKK (DE, M)

Output:

Decrypted Virtual Machine as VM

Process:

Step - 1. Load the encrypted virtual machine as VME
Step - 2. Build the decrypted virtual machine, VM as VM = pow(VME,DE) mod M
Step - 3. Return VM

The above algorithms are framed to solve the third problem discussed in the previous section of this work.

Further, the final framework is furnished here [Fig. 1]:

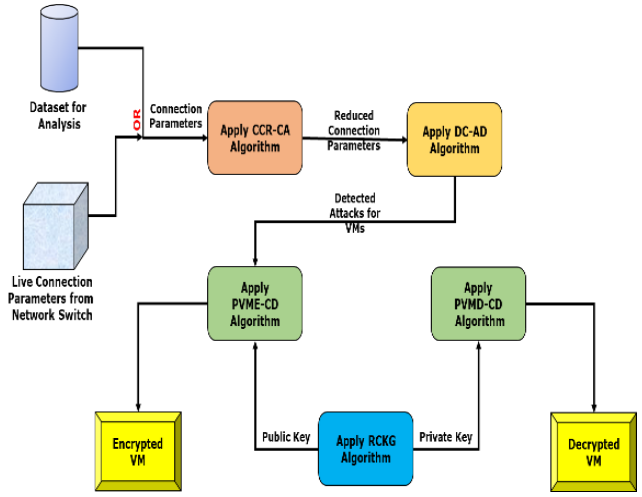


Fig. 1. A Framework for Cloud based Virtual Machine Security by Change Management using Machine Learning.

Further, in the next section of this work, the obtained results from these proposed algorithms are discussed.

V. Result and Discussion

After the detailed understanding on the proposed algorithms, here the obtained results are furnished.

Firstly, the used dataset [16] is analyzed here [Table II].

Further, the data is visualized graphically here [Fig. 2].

Here this is important to observe that, the many attributes have higher unique distributions and further demonstrates unique characteristics to detect large number of attacks.

Secondly, the impact or the correlation analysis results are furnished here [Table III].

The obtained results are again visualized graphically [Fig. 3].

Here it is natural to realize that the many of the attributes have demonstrated higher correlation than the other attributes. As per the proposed algorithm, the threshold of the correlation is calculated as 0.223 and based on the correlation theory, the positive impacted and meaning full attributes correlation must be above 0.50. Thus, again based on the proposed algorithm, the median value of the correlation is considered as 0.135.

Henceforth, based on the new correlation threshold, the following attributes are identified in the reduced set [Table IV].

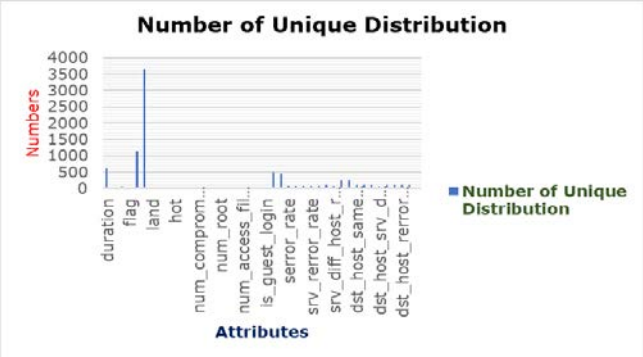
Further, the reduced set is also analyzed graphically here [Fig. 4].

Here, it is worth noting that, due to this process the information loss is minimum as the diversified nature of the dataset with high distribution is kept intact.

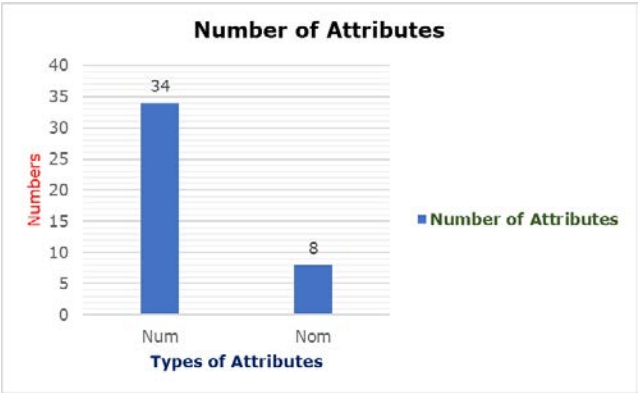
Further, the results from the deep clustering process to detect the attacks are furnished here [Table V].

TABLE II. DATASET ANALYSIS

SNO	Attribute Name	Attribute Type	Missing Value (%)	Number of Unique Distribution
1	“duration”	Num	0%	624
2	“protocol_type”	Nom	0%	3
3	“service”	Nom	0%	64
4	“flag”	Nom	0%	11
5	“src_bytes”	Num	0%	1149
6	“dst_bytes”	Num	0%	3650
7	“land”	Nom	0%	2
8	“wrong_fragment”	Num	0%	3
9	“urgent”	Num	0%	4
10	“hot”	Num	0%	16
11	“num_failed_logins”	Num	0%	5
12	“logged_in”	Nom	0%	2
13	“num_compromised”	Num	0%	23
14	“root_shell”	Num	0%	2
15	“su_attempted”	Num	0%	3
16	“num_root”	Num	0%	20
17	“num_file_creations”	Num	0%	9
18	“num_shells”	Num	0%	4
19	“num_access_files”	Num	0%	5
20	“num_outbound_cmds”	Num	0%	1
21	“is_host_login”	Nom	0%	2
22	“is_guest_login”	Nom	0%	2
23	“count”	Num	0%	495
24	“srv_count”	Num	0%	457
25	“serror_rate”	Num	0%	88
26	“srv_serror_rate”	Num	0%	82
27	“rerror_rate”	Num	0%	90
28	“srv_rerror_rate”	Num	0%	93
29	“same_srv_rate”	Num	0%	75
30	“diff_srv_rate”	Num	0%	99
31	“srv_diff_host_rate”	Num	0%	84
32	“dst_host_count”	Num	0%	256
33	“dst_host_srv_count”	Num	0%	256
34	“dst_host_same_srv_rate”	Num	0%	101
35	“dst_host_diff_srv_rate”	Num	0%	101
36	“dst_host_same_src_port_ra ”	Num	0%	101
37	“dst_host_srv_diff_host_ra”	Num	0%	58
38	“dst_host_serror_rate”	Num	0%	99
39	“dst_host_srv_serror_rate”	Num	0%	101
40	“dst_host_rerror_rate”	Num	0%	101
41	“dst_host_srv_rerror_rate”	Num	0%	100
42	“class”	Nom	0%	2



(a)



(b)

Fig. 2. (a) and (b) Analysis of the Dataset.

TABLE III. CORRELATION ANALYSIS

SNO	Correlation with “Class” Variable
1	0.150
2	0.112
3	0.368
4	0.525
5	0.016
6	0.097
7	0.008
8	0.039
9	0.009
10	0.057
11	0.135
12	0.618
13	0.021
14	0.018
15	0.022
16	0.021
17	0.016
18	0.052
19	0.070
20	0.000
21	0.010
22	0.116
23	0.353
24	0.092
25	0.282
26	0.280
27	0.517
28	0.513
29	0.550
30	0.261
31	0.192
32	0.399
33	0.645
34	0.636
35	0.276
36	0.030
37	0.022
38	0.312
39	0.308
40	0.528
41	0.506

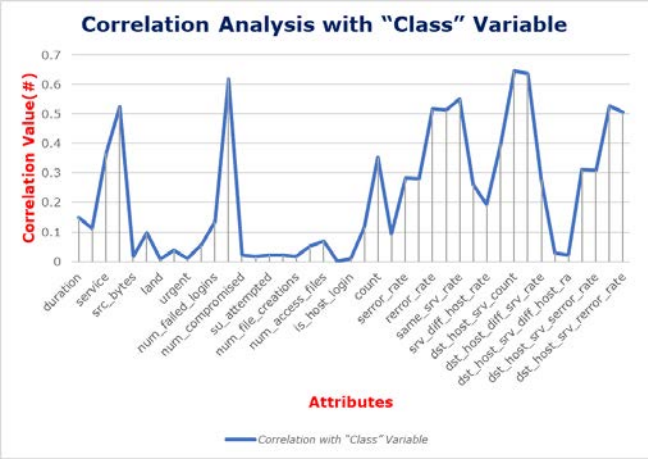


Fig. 3. Correlation Analysis.

TABLE IV. REDUCED ATTRIBUTE SET WITH CORRELATION

SNO	Attribute Name	Correlation with "Class" Variable
1	"dst_host_srv_count"	0.645
2	"dst_host_same_srv_rate"	0.636
3	"logged_in"	0.618
4	"same_srv_rate"	0.550
5	"dst_host_rerror_rate"	0.528
6	"flag"	0.525
7	"rerror_rate"	0.517
8	"srv_rerror_rate"	0.513
9	"dst_host_srv_rerror_rate"	0.506
10	"dst_host_count"	0.399
11	"service"	0.368
12	"count"	0.353
13	"dst_host_serror_rate"	0.312
14	"dst_host_srv_serror_rate"	0.308
15	"serror_rate"	0.282
16	"srv_serror_rate"	0.28
17	"dst_host_diff_srv_rate"	0.276
18	"diff_srv_rate"	0.261
19	"srv_diff_host_rate"	0.192
20	"duration"	0.150
21	"num_failed_logins"	0.135

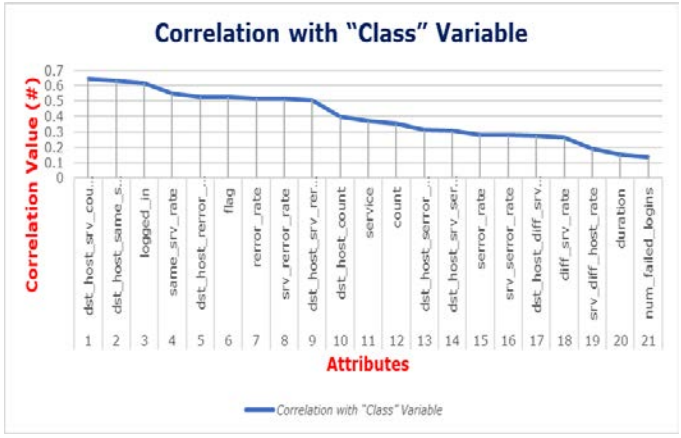


Fig. 4. Reduced Attribute Set Correlation Analysis.

TABLE V. ATTACK DETECTION ACCURACY ANALYSIS

Analysis Metric	Number of Values	Percentage (%)
"Correctly Classified Instances"	84248	98.2335
"Incorrectly Classified Instances"	1515	1.7665
"Kappa statistic"	0.9638	-
"Mean absolute error"	0.032	-
"Root mean squared error"	0.12	-
"Relative absolute error"	-	6.5494
"Root relative squared error"	-	24.2856
"Total Number of Instances"	85763	-

The results are observed visually here [Fig. 5].

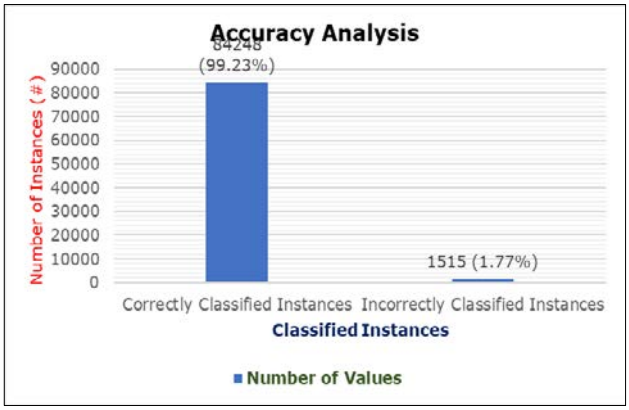


Fig. 5. Detection Accuracy Analysis.

Clearly from the results it is worth noting that the accuracy of the proposed deep clustering algorithm for attack detection is 99.23% with the newer types of attacks.

Further, the change detection algorithm for the virtual machines produces a log of tracked changes. The analysis is performed over 1000 virtual machines, however, for the visualization on 10 virtual machine logs are presented here.

Sample Change Detection Log File
Change Tracking for VM #1 Tracking for version #1 VM Size reduced by 126 GB
Change Tracking for VM #2 Tracking for version #1 VM Size reduced by 95 GB
Change Tracking for VM #3 Tracking for version #1 VM Size reduced by 94 GB Tracking for version #2 VM Size reduced by 42 GB
Change Tracking for VM #4 Tracking for version #1 VM Size increased by 138 GB
Change Tracking for VM #5 Tracking for version #1 VM Size reduced by 183 GB Tracking for version #2 VM Size increased by 28 GB
Change Tracking for VM #6 Tracking for version #1 VM Size increased by 236 GB Tracking for version #2 VM Size reduced by 28 GB
Change Tracking for VM #7 Tracking for version #1 VM Size increased by 208 GB
Change Tracking for VM #8 Tracking for version #1 VM Size increased by 275 GB
Change Tracking for VM #9 Tracking for version #1 VM Size reduced by 32 GB Tracking for version #2 VM Size reduced by 42 GB Tracking for version #3 VM Size increased by 79 GB
Change Tracking for VM #10 No Changes Detected

From the above sample log file, the following aspects are conclusive regarding the virtual machine change detection algorithm:

1) The changes for any virtual machine can be detected over multiple versions of the same VM.

2) The changes are reflected in terms of size; however, the actual change management is tracked based on characteristics of the virtual machines.

3) The detection algorithm also ensures no changes if the version of the same virtual machine is not updated.

Henceforth, it is conclusive that, the change management algorithm is perfectly justifying the claims made is this work.

Further, the key generation algorithm outputs are analysed here [Table VI]. During the testing phase, the algorithm is tested for more than 1000 instances. However, for representation purposes only 10 examples from the total outcomes are furnished.

TABLE VI. KEY GENERATION TIME ANALYSIS

Test Sequence #	**Key Generation time (ns)**
Seq #1	7
Seq #2	9
Seq #3	14
Seq #4	7
Seq #5	10
Seq #6	20
Seq #7	15
Seq #8	25
Seq #9	10
Seq #10	14

The results are visualized graphically here [Fig. 6].

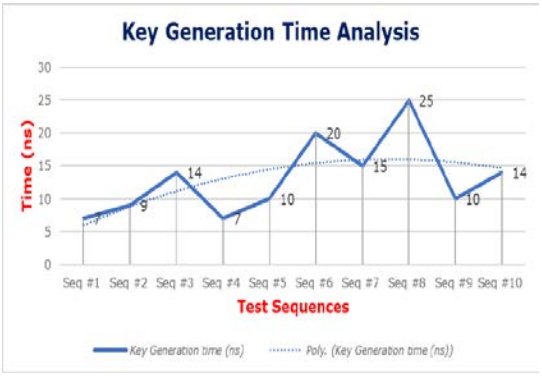


Fig. 6. Key Generation Time Analysis.

It is evident from the above results, that the time taken for the key generation demonstrates fairly liner characteristics, which is always expected for any best key generation algorithms.

TABLE VII. CRYPTOGRAPHIC ALGORITHMS TIME ANALYSIS

Test Sequence #	**Encryption Time (ns)**	**Decryption Time (ns)**
Seq #1	19	22
Seq #2	6	16
Seq #3	4	19
Seq #4	10	7
Seq #5	17	1
Seq #6	11	20
Seq #7	18	14
Seq #8	14	20
Seq #9	8	15
Seq #10	12	16

Further, the encryption and decryption time analysis is furnished here [Table VII]. During the testing phase, the algorithm is tested for more than 1000 instances. However, for representation purposes only 10 examples from the total outcomes are furnished.

The results are also visualized graphically here [Fig. 7].

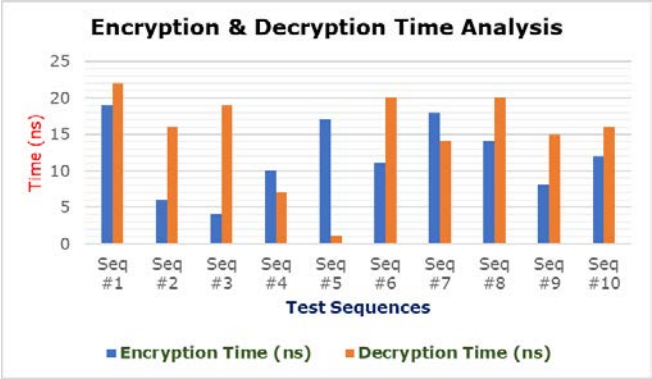


Fig. 7. Encryption and Decryption Time Analysis.

The results obtained in terms of time taken to perform the encryption and decryption operations on the detected changes on the virtual machines clearly showcase a trend of reduced time. This reduction is achieved due to the change management algorithm deployed for the virtual machine versions.

Further, in the next section of this work, the obtained results are compared with the other parallel research outcomes.

VI. Comparative Analysis

The obtained result from the proposed framework is highly satisfactory. Nonetheless, without a comparative analysis, no work can be concluded as benchmarked outcome. Thus, in this section of the work, the proposed framework using various parameters is compared with the parallel popular research outcomes [Table VIII].

Henceforth, it is conclusive to state that, the proposed framework has outperformed the parallel popular research works in terms of capabilities and as well as in terms of model complexity.

Finally, in the next section of the work, the research conclusion is presented.

TABLE VIII. Comparative Analysis

Author, Year	Methodology	Capabilities	Model Complexity
X. Lu et al. [14], 2020	Machine Learning	Reactive Security	$O(n^2)$
N. E. Moussaid et al. [13], 2020	Machine Learning	Reactive Security	$O(n^2)$
F. Cai et al. [10], 2019	Machine Learning	Reactive Security	$O(n^2)$
B. Sudhakar et al. [15], 2019	Machine Learning	Reactive Security	O(n*m)
Proposed Framework	Machine Learning	Reactive & Proactive Security	O(n)

VII. Conclusion

This research establishes benchmark in many aspects. In any of the parallel research outcomes, the reduction of time for applying the cryptographic aspects is ignored, which as per this work is most evident to increase the responsiveness of the cloud security. Also, this work elaborates the possibilities of detection of the attacks with the simplest model with least complexity. The proposed mathematical models and algorithms are strong evidence of the claim that, this framework is not only capable of detection of existing or known attacks, rather, this framework can also detect newer or unknown types of attacks based on the connection characteristics analysis. The detection rate on the benchmarked dataset is over 98%, which is again a benchmark for these types of framework.

References

[1] P. Mishra et al., "Intrusion detection techniques in cloud environment: A survey", J. Netw. Comput. Appl., vol. 77, pp. 18-47, 2017.

[2] P. Mishra et al., "VAED: VMI-assisted evasion detection approach for infrastructure as a service cloud", Concurrency Comput.: Practice Experience, vol. 29, 2017.

[3] M. R. Watson et al., "Malware detection in cloud computing infrastructures", IEEE Trans. Depend. Sec. Comput., vol. 13, no. 2, pp. 192-205, Mar./Apr. 2016.

[4] V. Varadharajan and U. Tupakula, "On the design and implementation of an integrated security architecture for cloud with improved resilience", IEEE Trans. Cloud Comput., vol. 5, no. 3, pp. 1-14, Jul.–Sep. 2017.

[5] Rekall: Memory FOrensics and Analysis Framework, May 2014, [online] Available: http://www.rekall-forensic.com/.

[6] T. K. Lengyel, Stealthy Monitoring with Xen Altp2m, 2016, [online] Available: https://blog.xenproject.org/2016/04/13/stealthy-monitoring-with-xen-altp2m/# comments.

[7] S. Gupta and P. Kumar, "System cum program-wide lightweight malicious program execution detection scheme for cloud", Inf. Secur. J.: A Global Perspective, vol. 23, no. 3, pp. 86-99, 2014.

[8] D. Kirat et al., "BareCloud: Bare-metal analysis-based evasive malware detection", Proc. 23rd USENIX Secur. Symp., pp. 287-301, 2014.

[9] C. Spensky, H. Hu and K. Leach, "LO-PHI: Low-observable physical host instrumentation for malware analysis", Proc. Netw. Distrib. Syst. Secur. Symp., pp. 1-15, 2016.

[10] F. Cai, N. Zhu, J. He, P. Mu, W. Li and Y. Yu, "Survey of access control models and technologies for cloud computing", Cluster Comput., vol. 22, no. S3, pp. 6111-6122, May 2019.

[11] A. Almtrf, Y. Alagrash and M. Zohdy, "Framework modeling for user privacy in cloud computing", Proc. IEEE 9th Annu. Comput. Commun. Workshop Conf. (CCWC), pp. 0819-0826, Jan. 2019.

[12] A. Khurshid, A. N. Khan, F. G. Khan, M. Ali, J. Shuja and A. U. R. Khan, "Secure-CamFlow: A device-oriented security model to assist information flow control systems in cloud environments for IoTs", Concurrency Comput. Pract. Exper., vol. 31, no. 8, Apr. 2019.

[13] N. E. Moussaid and M. E. Azhari, "Enhance the security properties and information flow control", Int. J. Electron. Bus., vol. 15, no. 3, pp. 249-274, 2020.

[14] X. Lu, L. Cao and X. Du, "Dynamic control method for tenants' sensitive information flow based on virtual boundary recognition", IEEE Access, vol. 8, pp. 162548-162568, 2020.

[15] B. Sudhakar, V. B. Narsimha, G. Narsimaha, Detection of Intrusion using Hybrid Feature Selection and Flexible Rule Based Machine Learning, International Journal of Engineering and Advanced Technology (IJEAT), 2019.

[16] 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.

Comparison of Convolutional Neural Network Architectures for Face Mask Detection

Siti Nadia Yahya[1]
Postgraduate Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Muhammad Noor Nordin[3]
Medical Engineering Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Aizat Faiz Ramli[2*]
Electronics Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Hafiz Basarudin[4]
Communication Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Mohd Azlan Abu[5]
Malaysia-Japan International Institute of Technology
Universiti Teknologi Malaysia
Jalan Sultan Yahya Petra, 54100, Kuala Lumpur, Malaysia

***Abstract*—In 2020 World Health Organization (WHO) has declared that the Coronaviruses (COVID-19) pandemic is causing a worldwide health disaster. One of the most effective protections for reducing the spread of COVID-19 is by wearing a face mask in densely and close populated areas. In various countries, it has become mandatory to wear a face mask in public areas. The process of monitoring large numbers of individuals to comply with the new rule can be a challenging task. A cost-effective method to monitor a large number of individuals to comply with this new law is through computer vision and Convolution Neural Network (CNN). This paper demonstrates the application of transfer learning on pre-trained CNN architectures namely; AlexNet, GoogleNet ResNet-18, ResNet-50, ResNet-101, to classify whether or not a person in the image is wearing a facemask. The number of training images are varied in order to compare the performance of these networks. It is found that AlexNet performed the worst and requires 400 training images to achieve Specificity, Accuracy, Precision, and F-score of more than 95%. Whereas, GoogleNet and Resnet can achieve the same level of performance with 10 times fewer number of training images.**

***Keywords*—Convolution neural network; deep learning; transfer learning; computer vision; facemask detection; COVID-19**

I. Introduction

Wearing face masks in public area is becoming more common due to the prevalence of COVID-19 outbreak all over the world [1]. Before the pandemic, small minority of the population especially in east Asian countries have been wearing face masks as a prevention against common flu. COVID-19 is the most recent pandemic virus to make a huge impact on human health in the past century. The exponential rate of COVID-19 transmission has forced the World Health Organization (WHO) to declare COVID-19 a worldwide pandemic in 2020. 150,047,341 have been infected by COVID-19 as of April 2021 across 188 countries. The virus is spreading through close contact, as well as in overcrowded publica areas. In multiple countries, individuals are constrained by the law to wear face mask in public areas. The rule was implemented as a reaction to a sudden spike in cases and fatalities in a various country. To enforce the public to comply with this rules, governmental agencies such police and health agency have to allocate significant number of their workforce to continuously public areas.

This paper demonstrates the application of Convolution Neural Network CNN to automate the classification of images of an individual wearing facemask and those without facemask. The ability to automate facemask detection can significantly reduce man power requirement and governmental expenditure. This paper also presents the performance comparison of popular CNN architectures for images classifications namely AlexNet, GoogleNet, ResNet-18, ResNet-50, and ResNet-101. The results presented in this research can be used by other researchers and machine learning engineers to identify suitable CNN architectures given the number of training data set, hardware capabilities and required accuracy to automate the images classification of a person wearing face mask.

This paper is organized as follows. Section II, provides discussion on the findings by other researchers on facemask detection and classification. Section III, describes the methodology on how the five different CNN architectures; AlexNet, GoogleNet, ResNet-18, ResNet-50 and ResNet-101 are being trained. Section IV, discussed the performance evaluation metrics that were used to compare the 5 different CNN architectures. The results of the study are presented and discussed in Section V. Finally, conclusions are drawn.

*Corresponding Author

A. Deep Learning

Deep Learning is a subset of Machine Learning, which in turn is a subset of Artificial Intelligence. The term "Artificial Intelligence" (AI) refers to techniques that enable computers to mimic human behavior. Machine Learning is an algorithm that has been trained using data to emulate human like decision making. Deep Learning and Artificial Neural Network are a subset of Machine Learning that are inspired by human brain structure. [2] Deep learning methods take the opportunity to achieve the same findings as humans by consistently assessing data using a specified structured methodology. Deep learning does this through the use of a multi-layered structure of algorithms known as neural networks.

As shown in Fig. 1, the design of the Artificial Neural Network is based on the anatomy of the human brain. Artificial Neural Network can be trained to detect objects and categorize various sorts of data in the same way that humans do. Singular layers of neural organizations might be considered as a type of filter that capacities from coarse to fine, expanding the likelihood of recognizing and creating the right outcome. The human brain works in a similar way. While going up against with new data, the brain endeavors to contrast it with recently known objects. Deep neural networks employ the same principle. It may be use neural networks to accomplish a variety of tasks such as grouping, classification, and regression. Can be use neural networks to categorize or classify unlabeled data based on similarities between samples. In the classification process, it might prepare the network on a labeled dataset to characterize the examples in the dataset into discrete classifications.

B. Convolution Neural Networks (CNN)

As illustrated in Fig. 2 and Fig. 3, a CNN contains three layers as a convolutional layer, a pooling layer, and a fully connected layer [3]. The 'input layer' of each CNN utilized receives pictures and recompress them before passing data on to following layers for extracting features. The next layers are referred to as 'Convolution layers,' and they serve as image filters, extracting features from pictures and generating match local features during testing. Activation function 'Rectified Linear Unit' (ReLU) is employed to replace every negative integer in the pooling layer with zero. ReLU also helps the CNN maintain mathematical stability by avoiding learnt values from being stuck at zero or blowing up toward infinity. Then the data is transferred to the 'pooling layer' [4]. This layer decreases the size of large pictures yet retaining the most important information. It maintains the most value from each frame by retaining the best fits of every feature within the frame. Flatten is the way toward changing over information into a one-dimensional cluster for contribution to the next layer. The yield of the convolutional layers is flattened to make a solitary extensive component vector. It is also connected to the final classification algorithm, forming a fully connected layer. The next-to-last layer is a fully connected layer that turns the greater filtered pictures into labeled with probability for every class of every picture being categorized. To give classification output, the last layer of the CNN architecture employs a classification layer such as softmax [5].

C. AlexNet

AlexNet was created by Alex Krizhevsky and is a convolutional neural network model that has made huge contributions to Artificial Intelligence (AI), particularly the use of deep learning to machine vision [6]. The CNN model won the ImageNet Large Scale Visual Detection (ILSVRC) competition in 2012, which evaluates methods for huge object recognizing and picture classifications. AlexNet consists of 60 million parameters, three fully connected layers, 650,000 neurons and five convolutional layers [7]. Convolutions of 11x11, 5x5, and 3x3 dimensions were used, as well as max pooling, dropout, data augmentation, ReLU activations, and SGD with momentum. The initial two convolutional layers are normalization and a maximum pooling layer. Plus, the third and fourth are directly connected while the fifth is joined by a maximum pooling layer. [6] The input is given it into softmax classifier, the second of which feeds into a softmax classifier. The authors used a regularisation approach termed "dropout" with a ratio of 0.5 to avoid overfitting in the fully-connected layers. The use of Rectified Linear Unit (ReLU) on each of the first seven layers is another AlexNet model feature.

D. GoogleNet

Szegedy proposed GoogleNet, which was the champion of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014 [6]. For the major auxiliary classifiers in the network, GoogleNet features four convolutional layers, three softmax layers, seven million parameters, five fully connected layers, nine inception modules, three average pooling layers and four max-pooling layers. In the fully connected layer, dropout regularization is used, and ReLU activation is used in all of the convolutional layers. GoogleNet has 22 total layers and is significantly deeper and wider than AlexNet, but it has a much smaller number of network parameters. The DistBelief distributed machine learning system was used to train GoogLeNet architecture with a small amount of model and data parallelism. In the RGB color space, the size of the receptive field in this network is 224x224 with a zero mean. GoogleNet was developed with the intent of being able to function on a variety of devices, including those with limited computational resources, such as those with a low memory footprint [6]. GoogleNet is 22 layers deep if just layers with parameters are counted, or 27 levels if pooling is counted, and has 7 million parameters. This network has a 27MB file size and a 224-by-224 image input size. The GoogLeNet was intended to be a computational force and reckoned with higher computational proficiency than a portion of its archetypes or equivalent organizations created at that point. The main convolution layer utilizes a filter patch that is altogether huge in contrast with other patch sizes in the network. The significant objective of this layer is to quickly limit the input image while holding spatial data by utilizing enormous filter sizes. The size of the input image is diminished by a factor of four at the subsequent convolution layer and another factor of eight preceding arriving at the primary initiation module, however a more noteworthy number of highlight maps are produced. The GoogLeNet architecture is comprised of nine inception modules. Furthermore, some inception modules include two max-pooling layers.

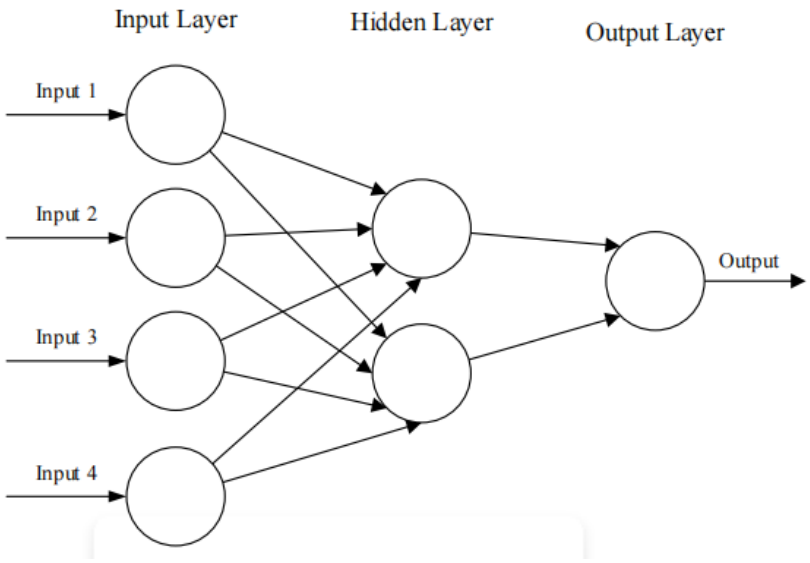


Fig. 1. Artificial Neural Network Connection [8].

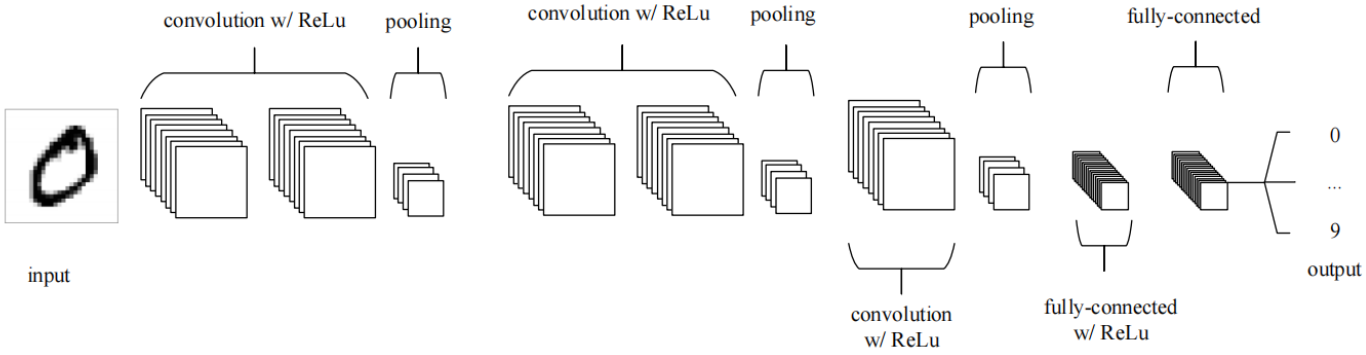


Fig. 2. Convolutional Neural Networks [8].

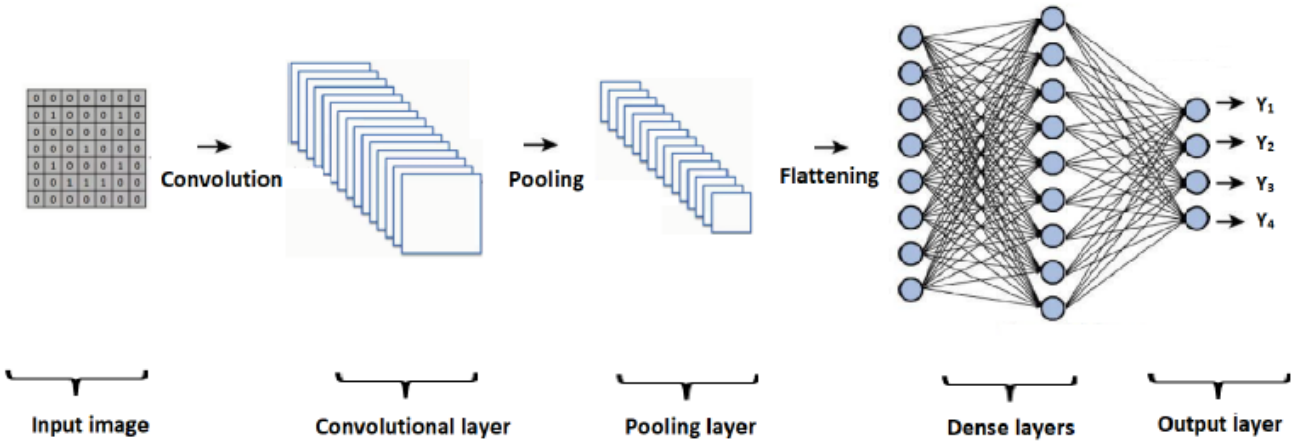


Fig. 3. Representation of a Convolutional Neural Network's Architecture [6].

Just before the linear layer, a dropout layer is being utilized. The dropout layer is a regularization procedure utilized during preparing to keep the network from overfitting. The linear layer is comprised of 1000 hidden units that compare to the 1000 classes in the Imagenet dataset. The last layer is the softmax layer, which utilizes the softmax function, an actuation function used to assess the likelihood circulation of an assortment of the number contained inside an information vector. A softmax actuation function is a vector wherein the assortment of qualities addresses the likelihood of a class or occasion event. The vector's qualities all amount to one.

E. ResNet (Residual Network)

The ResNet feature is based on deep architectures that have demonstrated good convergence and accuracy, were created by He et al. [6] They won the ImageNet Large Scale Visual Recognition Challenge classification competitions in 2015. A residual neural network (ResNet) is a form of Artificial Neural Network (ANN) that is based on pyramidal cell frameworks in the cerebral cortex. Skip connections, or shortcuts, are used by residual neural networks to skip over some layers. ResNet was created using numerous stacked residual units and a variety of layer counts such as 18, 34, 50, 101, 152, and 1202. The number of operations, on the other hand, might vary depending on the architecture. The residual units for all of the preceding are made up of convolutional, pooling, and layering operations. The ResNet 18 network provides an excellent balance of depth and performance, and it is made up of a fully connected layer with a softmax, five convolutional layers, and one average pooling layer. ResNet-18 comprises 11.7 million parameters and an 18 depth layer with a size of 44MB.

Resnet-50 has 49 convolutional layers, 25.6 million parameters, 50 depth layer with a size of 96MB and a fully connected layer at the end of the network. [9] ResNet-101 is a deep convolutional neural network with 101 layers. This architecture has 44.6 million parameters and 101 depth layers, and it is 167MB in size. It will load a pre-trained rendition of the network that has been trained on over 1,000,000 images from the ImageNet information base. The network was pre-trained to recognize images into 1000 distinctive item classes. ResNet is a reliable architecture for identifying a wide range of classes, having won the 2016 ImageNet competition.

A deep residual network (ResNet) is made up of modules, which are entities with identical loops layered on top of each other. Each module is made up of multiple convolutional layers that are used to become familiar with the features of the input space. After the second convolutional layer, a dropout layer was added. Each module delivers more generalized output with greater regularization with the inclusion of the Dropout layer. Many architectures in the literature use dropout, and it is often used on layers with a large number of parameters to minimize feature adaptation and overfitting. Dropouts outperform in generalization. As a default, Softmax is utilized after fully connected layers.

II. Related Work

This section discusses similar research that has been conducted relating to face mask detection.

A. Facial Mask Detection using Semantic Segmentation

The objective of the paper presented by Meenpal et al., 2019 [10] was to develop a binary face classifier that can identify each face in the frame regardless of alignments, including a strategy for generating accurate face segmentation masks of any arbitrary size input image. The approach begins with an RGB image of any size and utilizes Predefined Training Weights of VGG – 16 Architecture for feature extraction. For segmented face masks, experiments on the Multi Parsing Human Dataset revealed a mean pixel-level efficiency of 93.884%.

B. Real-Time Face Mask Identification using Facemasknet Deep Learning Network

Inamdar & Mehendale, 2020 develop a deep learning architecture called Facemasknet [11] COVID-19 face mask classification. The proposed architecture provides three characterizations which are people wearing a mask, erroneously worn masks, and no mask detected. Utilizing a deep learning technique called Facemasknet, they got a precision of 98.6%.

C. Covid-19 Facemask Detection with Deep Learning and Computer Vision

Vinitha & Velantina, 2020 developed a real-time face detection from a live feed via their webcam [12]. The research was conducted using OpenCV framework and A.I framework such as Python, Tensor Flow and Keras. Their aim is to employ deep learning and computer vision to determine if the person in the picture or video feed is wearing a mask.

D. Deep Learning based Safe Social Distancing and Face Mask Detection in Public Areas for COVID-19 Safety Guidelines Adherence

Yadav, 2020 presents a technique for forestalling the transmission of the virus by observing individuals progressively to check whether they are utilizing safe social distance and wearing face mask in public area [13]. The method used for this research includes Raspberry pi4, OpenCV framework, MobileNetV2 and TensorFlow. The detection of face mask wearing achieves an accuracy of 91.2%.

E. A Hybrid Deep Transfer Learning Model with Machine Learning Methods for Face Mask Detection in the Era of the COVID-19 Pandemic

Loey et al., 2021 presents a hybrid model for face mask identification based on deep and conventional machine learning. There are two elements to the technique that follows [1]. The first element will use Resnet-50 to extract features. The ensemble method, decision trees, and Support Vector Machine (SVM) are used in the second element to categorise face masks.

F. Validating the Correct Wearing of Protection Mask by Taking a Selfie: Design of a Mobile Application 'CheckYourMask' to Limit the Spread of COVID-19

The COVID-19 contagiousness is considered to be high in comparison to the flu. Hammoudi et al., 2020 presents a mobile application design that allows anybody with a smartphone the ability to snap a picture to verify that his or her protective mask is properly positioned on his or her face [14]. The technique used in this research included Android, OpenCV, and Haar-like. True Detection (TD) accuracy for the face is 99.92% and the nose is 100%.

G. Identifying Facemask-Wearing Condition using Image Super-Resolution with Classification Network to Prevent COVID-19

Qin & Li, 2020 proposed another facemask wearing condition identification technique as a cooperation with picture super-resolution with classification algorithm (SRCNet) to quantify three different classification issues using

unconstrained 2D facial image photos [15]. The suggested technique included four major steps which are pictures pre-processing, image recognition and cropping, picture super-resolution and recognition of wearing a face mask circumstances. The proposed technique reported a 98.7% accuracy rate.

H. An Application of Mask Detector for Prevent Covid-19 in Public Services Area

Henderi et al., 2020 created systems for real-time monitoring of those who do not wear a face masks in public areas [16]. The authors utilize images and video input from a camera and connects it to a Speed Maix Bit CPU to process data and show it onto the LCD display. The materials used to develop the system are MicroPython, Sipeed Maix Bit, MaixPy and Python 3.

I. An Automated System to Limit COVID-19 using Facial Mask Detection in Smart City Network

Rahman et al., 2020 propose a framework that limits COVID-19 spread in an active city network where all open spots are monitored by Closed-Circuit Television (CCTV) cameras by recognizing people who are not wearing any facial masks [4]. When an individual without a face mask is detected, the city network alerts the necessary authorities. The materials used for this system is CCTV and GPS. The system achieved an accuracy rate of 98.7% of facemask detection.

J. Retinamask: A Face Mask Detector

Jiang & Fan, 2020 presented RetinaFaceMask, a high-accuracy and effective face mask detector [17]. The presented RetinaFaceMask detector is a one-stage detector comprised of a feature pyramid network that combines high-level semantic information with numerous feature vectors and a novel context attentive modules focused on identifying face masks. Furthermore, they provide a cross-class object removal approach for rejecting predictions with low confidence and a high intersection of a union. The framework used in this research were MobileNet, ResNet, and PyTorch.

K. Performance Evaluation of Intelligent Face Mask Detection System with Various Deep Learning Classifiers Keywords

The research by [18] focuses on the use of deep learning algorithms to identify persons who do and do not wear a face mask. The framework has been trained to decide if an individual is wearing a face mask or not. At the point when the algorithm perceives an individual without a mask, an alarm will be set off to caution the people around or the relevant authorities, so that appropriate action may be taken against such offenders. Like most establishments, associations, businesses, shopping centers, and hospital must resume normal operations before the epidemic is removed, to incorporate a face mask recognition strategy with the current information system at the passage and leave entryways is emphatically encouraged.

III. Methodology

The main objectives of this research are to demonstrates the application of using existing CNN architectures; ResNet-101, ResNet-50, ResNet-18, GoogleNet and AlexNet in classifying images of an individual wearing and not wearing facemask. The first stage is to conduct transfer learning on the networks using image datasets. The most essential criterion for evaluating the performance is to see if the prediction accuracy varies among all CNN architectures used in this research. This study is divided into two phases which are training and testing of the face mask detector.

In the training phase, the dataset is loaded for the model to be trained, and the model is serialized. The training datasets consist of images of faces with and without facemask. Fig. 4, 5, 6, 7 and 8 shows example of training images of faces at various angles without facemask. To gauge and compare the performance of ResNet-101, ResNet-50, ResNet-18, GoogleNet and AlexNet, the dataset were varied as follows:

- 20 datasets (10 images with facemask, 10 without facemask)
- 40 datasets (20 images with facemask, 20 without facemask)
- 60 datasets (30 images with facemask, 30 without facemask)
- 80 datasets (40 images with facemask, 40 without facemask)
- 100 datasets (50 images with facemask, 50 without facemask)
- 200 datasets (100 images with facemask, 100 without facemask)
- 300 datasets (150 images with facemask, 150 without facemask)
- 400 datasets (200 images with facemask, 200 without facemask)

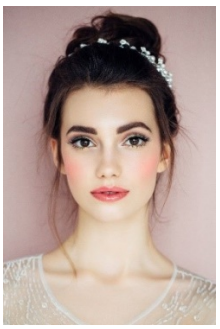


Fig. 4. Example on Front Facing Image [19].

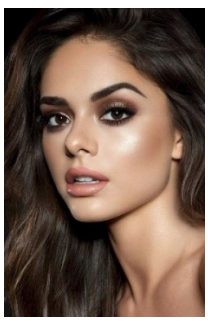


Fig. 5. Example on Left Facing Image [19].

Fig. 6. Example on Right Facing Image [19].

Fig. 7. Example on Bottom Facing Image [19].

Fig. 8. Example on Top Facing Image [19].

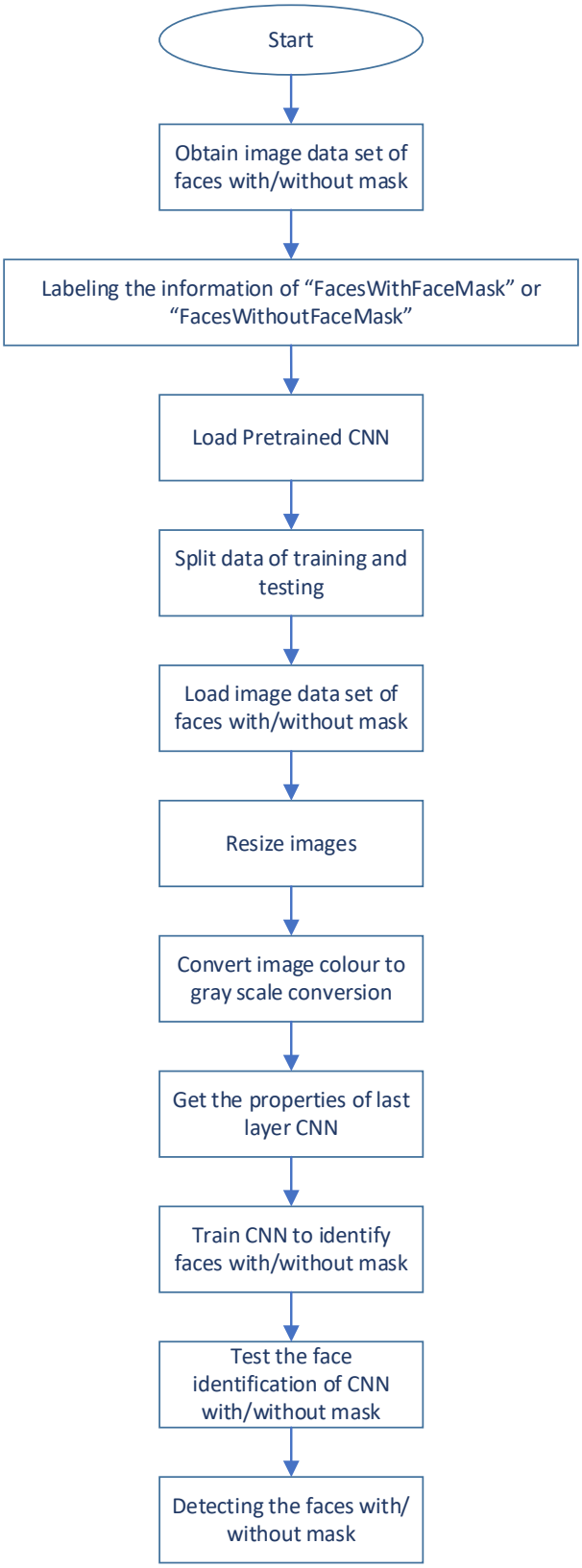


Fig. 9. Process Flow for the Transfer Learning.

The limitation of this research each of the training images used consists of only one face and Matlab programming was used to perform the transfer learning. For Alexnet, all the training images were resized to 227 x 227 x 3 pixels. Whilst for the ResNet-18, ResNet-50, ResNet-101 and GoogleNet the training images were resized to 224x224x3 pixels. After the model was trained, the images were stacked as input to recognize whether an individual was wearing a face mask or not.

In the testing phase, the dataset consists of a total of 200 images which were not part of the training data set, 100 images of faces without face masks and 100 images of faces with face masks. The testing datasets consist of faces at various angles and were fed into a trained CNN architectures to classify if the face detected in the image is wearing a face mask or not. After that detection of face mask takes place, the result will state either the face on the images uploaded wearing a face mask or not wearing a face mask and appeared on the screen display. The screen will mention 'FacesWithFaceMask' for the result of images of people wearing a face mask and 'FacesWithoutFaceMask' for the images of people not wearing a face mask.

Fig. 9 summarizes the flowchart for the transfer learning of CNN for face mask classification. First, the data set of human faces wearing and not wearing a face mask were compiled. The data sets were labelled with their respective categories, "FacesWithFaceMask" and "FacesWithoutFaceMask." A specific CNN architecture is then loaded. The training datasets are then randomly divided into training and testing with a ratio of 70:30. All the images are resized according to the requirements of specific CNN architectures and converted into a grayscale. The features of the final layer CNN are obtained as every CNN network has a different final layer. The CNN network is then train to recognize faces with and without a face mask using the training dataset. Using an image which are not part of the training dataset, the performance of the various trained CNN networks in classifying face with and without facemask are evaluated.

IV. Performance Metrics

Several performance measures are used in this research to compare the performance of the various CNN networks in classifying images of a person with and without facemask.

The sensitivity performance, also known as recall, refers to the accuracy of true positives and how many positive class samples were appropriately labelled [20]. Sensitivity can be calculated using (1), where True Positive (TP) is the number of events that are both positive and accurately identified, which means the number of images with a person are wearing a face mask and are correctly identified as such. While, False Negative FN is defined as the number of positive events that are incorrectly classified as negative. In this research FN, is the number of images of a person wearing facemask but has been incorrectly classified by CNN as not wearing a facemask.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

Specificity is defined as the probability distribution of true negatives with a secondary class, which generally corresponds to the possibility that the negative label was correct, and is presented in (2). True Negatives (TN), also known as negative cases that are classified as negative, showing that individuals are not wearing a face mask but are labeled as wearing a face mask. False Positive (FP) is the individuals who are not wearing a face mask but are erroneously categorized as wearing a face mask.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

Accuracy is the most often used parameter for assessing classification performance. This measure computes the proportion of properly identified samples and is represented by (3) [20].

$$\text{Accuracy} = \frac{TP + TN}{T\,P + TN + FP + FN} \tag{3}$$

Precision is calculated by using (4), which divides the total number of true positives plus false positives by the number of true positives. This statistic evaluates the algorithm's predicting capabilities and is concerned with accuracy. Precision refers to the model's "accuracy" in terms of both the number of positive predictions and the number of positive predictions that occur. [20].

$$\text{Precision} = \frac{TP}{T\,P + FP} \tag{4}$$

The harmonic mean accuracy and recall are used to generate the F-score, as shown in (5). It is related with the examination of positive classes. A perfect score for this parameter means that the model leads the positive class [20].

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

V. Result and Discussion

Fig. 10 and 11 illustrates the correct classification of trained CNN networks in identifying 'FacesWithFaceMask' and 'FacesWithoutFaceMask'. Results shown to Fig. 12, 13, 14, 15 and 16, all of the networks performed differently in terms of its statistical significance.

The performance results of AlexNet shows that the Sensitivity is highest at 20 and 40 training images, as shown in Fig. 12. However, for Specificity, Accuracy, Precision, and F-score, the results are its lowest at 20 training images. The performance of AlexNet to classify images of a person wearing facemask can be improved by providing it with more training data set. AlexNet has the lowest accuracy and performed the worst when compare to other CNN architectures. Similar findings was also reported by Neha Sharma et al. [21]. To achieve an acceptable level of performances, AlexNet requires 200 training images to achieve average performance (sensitivity, accuracy, specificity, precession and F-score) of more than 95%. More training data can result in longer training processing time which can undesirable for low powered machines. AlexNet has relatively poorest performance compared to other CNN architectures is because the network has far fewer layers (AlexNet consists of only 8 layers).

Fig. 10. Result will show "FacesWithoutFaceMask' for the Image of People Not Wearing a Face Mask.

Fig. 11. Result will Show "FacesWithFaceMask' for the Image of People Wearing a Face Mask.

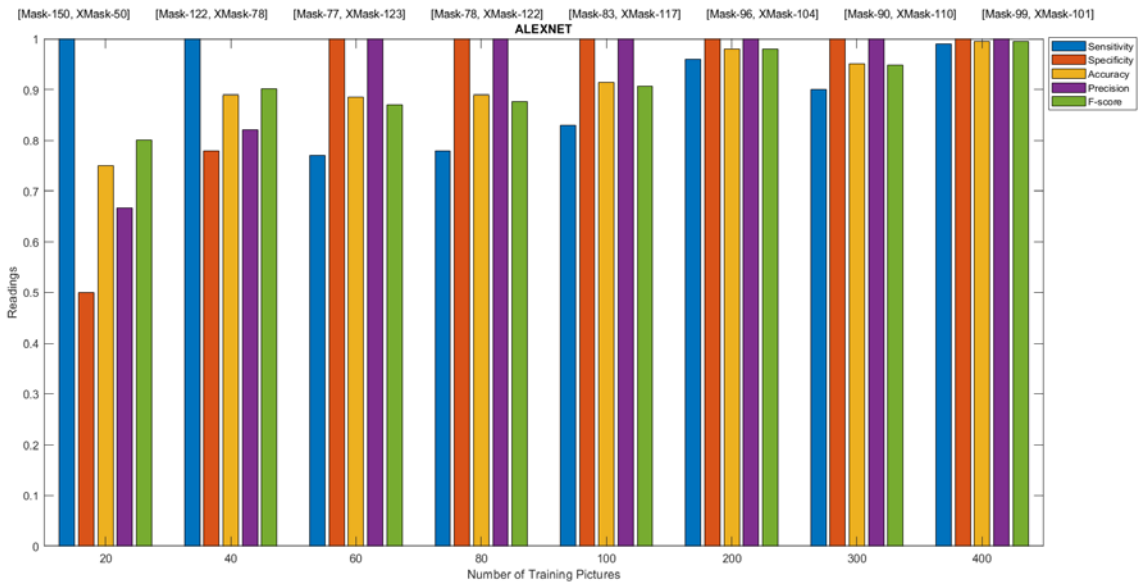


Fig. 12. Performance Results (%) for AlexNet.

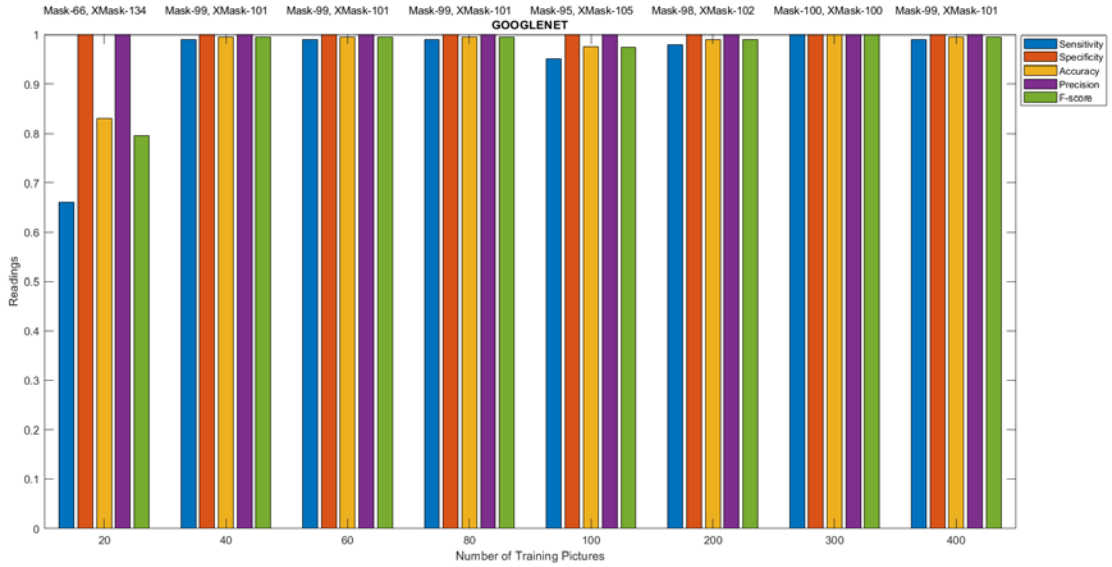


Fig. 13. Performance Results (%) for GoogleNet.

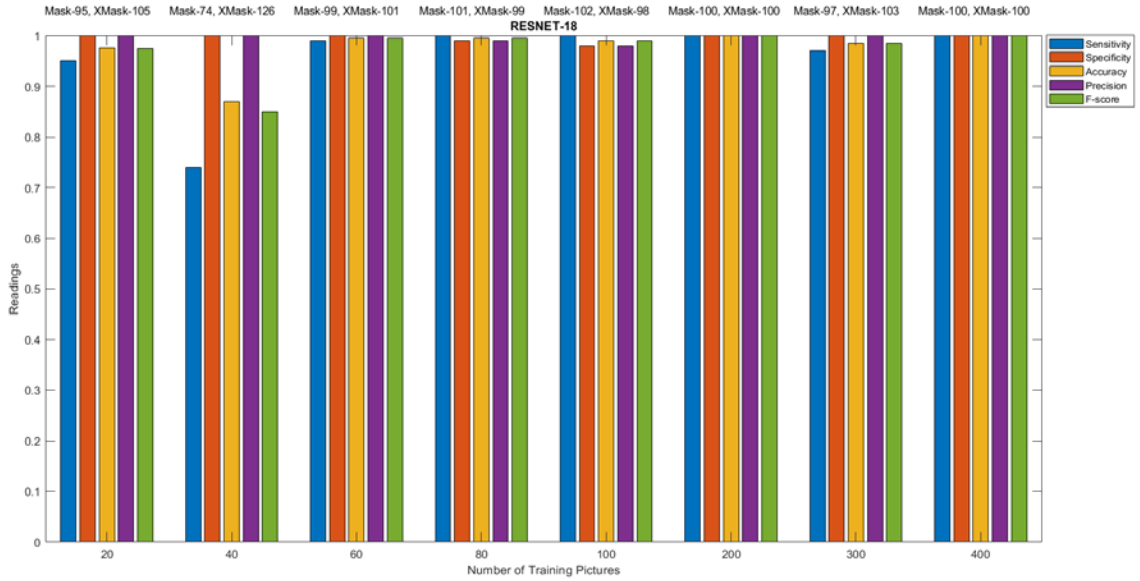


Fig. 14. Performance Results (%) for ResNet-18.

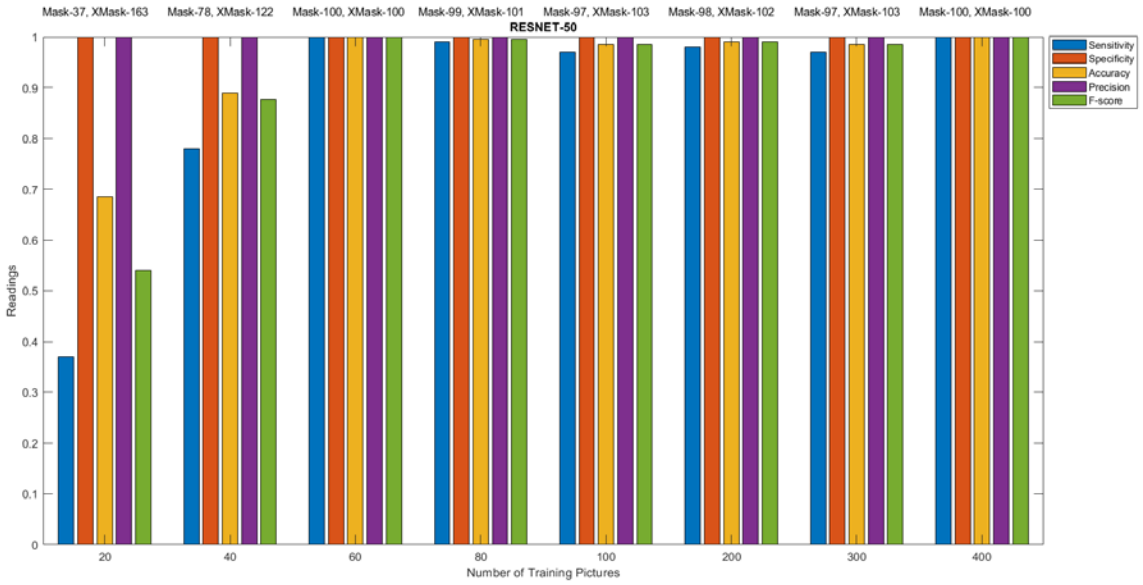


Fig. 15. Performance Results (%) for ResNet-50.

In comparison, GoogleNet which consists of 22 layers only needs 40 images of training data to achieve Accuracy, Sensitivity, Specificity, Precision and F-score of more than 95%.

Fig. 14 and 15 shows that the percentage of performance results for ResNet-18 and ResNet-50 for Sensitivity, Specificity, Accuracy, Precision, and F-score are unstable at 20 to 40 training images. However, as the number of training images increases to 60, the results stabilize with an average performance evaluation of 97%. According to Fig. 16, the best performed network based on the performance evaluation metrics is ResNet-101. ResNet-101 only needs 40 images of training data to achieve Accuracy, Sensitivity, Specificity, Precision and F-score of more than 98%. ResNet-101 achieved the highest average performance as the network has far more layers compared to the CNN architectures presented in this paper. As the name implies, ResNet-101 consists of 101 layers. However, the complexity of the network means it requires more processing power.

The Specificity and Precision performance of GoogleNet, ResNet-50, and ResNet-101 are consistently more than 98% from the start. However this is not the case for AlexNet and ResNet-18 which indicates that these architectures struggle to classify faces without facemask with low number of training data. Low Precision value means that the networks made several incorrectly classification on faces without facemask as wearing facemask. Overall, as the number of training images increases for AlexNet, GoogleNet, ResNet-18, ResNet-50 and ResNet-101, the performance in terms Specificity, Accuracy, Precision, and F-score also improve.

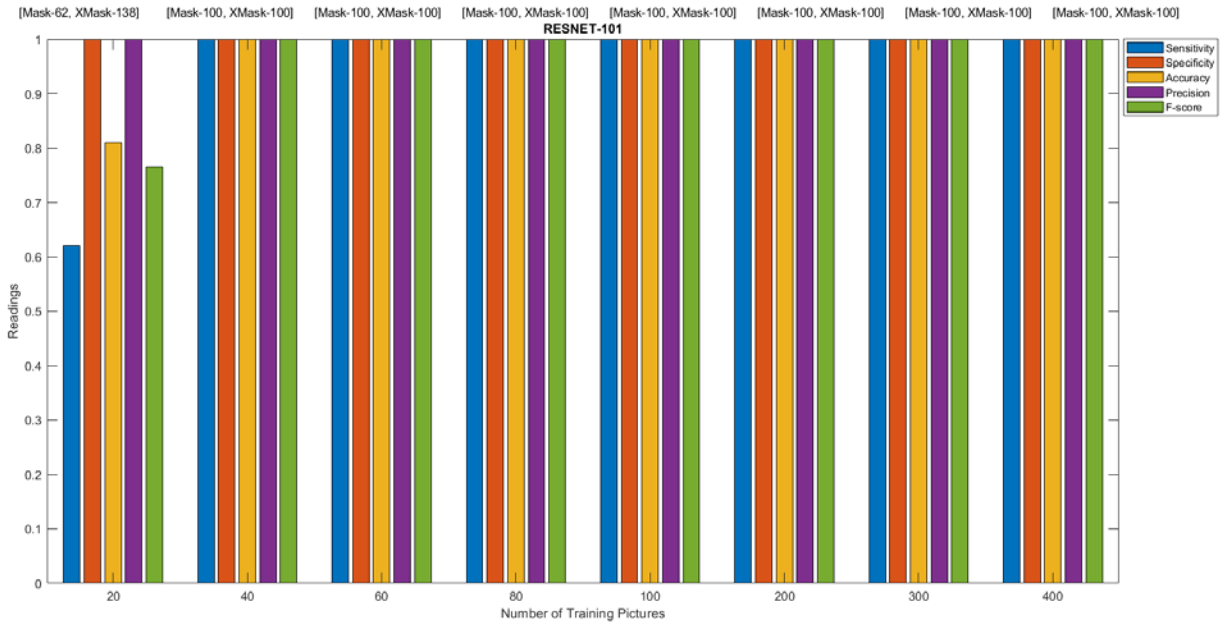


Fig. 16. Performance Results (%) for ResNet-101.

VI. Conclusion

This papers successfully demonstrates application of transfer learning in popular CNN architectures to classify images of a person wearing or not wearing a facemask. In this study, five network architectures were compared using performance metrics; AlexNet, ResNet-18, ResNet-50, GoogleNet, and ResNet-101. It is found that on average AlexNet performed the worst and requires far greater training data to achieve accuracy of more than 95%. ResNet-101 achieved the highest accuracy and requires smallest number of training data to far more 95% accuracy. However the complexity of ResNet-101 means that it requires far greater processing power. GoogleNet strikes the balance of not being overly complex whilst achieving an acceptable level of performance with relatively small number of training data images. Further research can be conducted to gauge the performance and computational resources requirement of various CNN architectures to determine whether a person is wearing a face mask correctly or incorrectly.

References

[1] Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," Meas. J. Int. Meas. Confed., vol. 167, no. May 2020, p. 108288, 2021, doi: 10.1016/j.measurement.2020.108288.

[2] H. Kim, "5 . 1 Artificial Intelligence , Machine Learning , and Deep," pp. 151–193, 2020.

[3] G. Wu et al., "Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks," Remote Sens., vol. 10, no. 3, pp. 1–18, 2018, doi: 10.3390/rs10030407.

[4] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud, and J. H. Kim, "An automated system to limit COVID-19 using facial mask detection in smart city network," IEMTRONICS 2020 - Int. IOT, Electron. Mechatronics Conf. Proc., 2020, doi: 10.1109/IEMTRONICS51293.2020.9216386.

[5] S. Datta, "A review on convolutional neural networks," Lect. Notes Electr. Eng., vol. 662, no. 03, pp. 445–452, 2020, doi: 10.1007/978-981-15-4932-8_50.

[6] V. Maeda-Gutiérrez et al., "Comparison of convolutional neural network architectures for classification of tomato plant diseases," Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041245.

[7] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan, "A comparative study for chest radiograph image retrieval using binary texture and deep learning classification," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2015-Novem, pp. 2940–2943, 2015, doi: 10.1109/EMBC.2015.7319008.

[8] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," no. December, 2015, [Online]. Available: http://arxiv.org/abs/1511.08458.

[9] M. D. Putro, D. L. Nguyen, and K. H. Jo, "Real-Time Multi-view Face Mask Detector on Edge Device for Supporting Service Robots in the COVID-19 Pandemic," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12672 LNAI. pp. 507–517, 2021, doi: 10.1007/978-3-030-73280-6_40.

[10] T. Meenpal, A. Balakrishnan, and A. Verma, "Facial Mask Detection using Semantic Segmentation," 2019 4th Int. Conf. Comput. Commun. Secur. ICCCS 2019, pp. 1–5, 2019, doi: 10.1109/CCCS.2019.8888092.

[11] M. Inamdar and N. Mehendale, "Real-Time Face Mask Identification Using Facemasknet Deep Learning Network," SSRN Electron. J., 2020, doi: 10.2139/ssrn.3663305.

[12] V. Vinitha and V. Velantina, "Covid-19 Facemask Detection With Deep Learning and Computer Vision," Int. Res. J. Eng. Technol., vol. 07, no. 08, pp. 3127–3132, 2020.

[13] S. Yadav, "Deep Learning based Safe Social Distancing and Face Mask Detection in Public Areas for COVID-19 Safety Guidelines Adherence," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 7, pp. 1368–1375, 2020, doi: 10.22214/ijraset.2020.30560.

[14] K. Hammoudi, A. Cabani, H. Benhabiles, and M. Melkemi, "Validating the correct wearing of protection mask by taking a selfie: Design of a mobile application 'CheckYourMask' to limit the spread of COVID-19," C. - Comput. Model. Eng. Sci., vol. 124, no. 3, pp. 1049–1059, 2020, doi: 10.32604/cmes.2020.011663.

[15] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19,"

Sensors (Switzerland), vol. 20, no. 18, pp. 1–23, 2020, doi: 10.3390/s20185236.

[16] Henderi, A. S. Rafika, H. L. H. Spits Warnar, and M. A. Saputra, "An Application of Mask Detector for Prevent Covid-19 in Public Services Area," J. Phys. Conf. Ser., vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012063.

[17] M. Jiang and X. Fan, "Retinamask: A Face Mask Detector," arXiv, 2020.

[18] C. Jagadeeswari and M. U. Theja, "Performance Evaluation of Intelligent Face Mask Detection System with various Deep Learning Classifiers Keywords :," Int. J. Adv. Sci. Technol., vol. 29, no. 11, pp. 3074–3082, 2020.

[19] iStock, "Stock Images, Royalty-Free Pictures, Illustrations & Videos - iStock." 2020, [Online]. Available: https://www.istockphoto.com/.

[20] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," Informatics Med. Unlocked, vol. 19, p. 100360, 2020, doi: 10.1016/j.imu.2020.100360.

[21] N. Sharma, V. Jain, and A. Mishra, "An Analysis of Convolutional Neural Networks for Image Classification," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 377–384, 2018, doi: 10.1016/j.procs.2018.05.198.

Encoding LED for Unique Markers on Object Recognition System

Wildan Pandji Tresna[1], Umar Ali Ahmad[2], Isnaeni[3], Reza Rendian Septiawan[4], Iyon Titok Sugiarto[5], Alex Lukmanto Suherman[6]

Research Center for Physics, National Research and Innovation Agency, Banten, Indonesia[1,3, 5, 6]
Department of Computer Engineering, School of Electrical Engineering, Telkom University, Bandung, Indonesia[1,2,4]

***Abstract*—In this paper a new approach of unique markers to detect and track moving objects with the encoding LED marker is presented. In addition, an LED spotlight system that can shoot light in the direction of the target is also proposed. The encoding process is done by making a unique blinking pattern on the LED, and thus the camera and a servomotor as an object recognition system would only recognize a unique marker given by an LED. In this work the camera with OpenCV could capture a unique marker in all variant of blinking patterns. A unique marker is important on object recognition system so the camera could identify the object marked by our unique marker and ignore all other markers that might captured by the camera. In addition, analysis of the PWM signal from an LED is carried out here to determine the characteristics of the LEDs in each color, determine the accuracy of the duty cycle, and the use of the bright-dim method on the LEDs. The results show that the highest accuracy is obtained when a 50% duty cycle is used with the on and off time are set to be 1 second for all LED colors. The benefit of the proposed system is confirmed by implementing an integrated control system as a unique marker. The effectiveness of the blinking LED against other laser interferences is also discussed.**

***Keywords*—Encoding LED; PWM signal; target markers; object recognition system**

I. Introduction

The optics and optoelectronics have made significant progress over the years, particularly towards more powerful and efficient laser devices [1]. Much of these advances are responsible for making them indispensable in modern warfare. There has been a large-scale growth of electro-optics devices and systems for various applications, such as target tracking and designating, range finding, and target acquisition [2][3]. Moreover, [4] and [5] suggested that the use of lasers for military purposes continues to grow every year. There are numerous benefits associated with electro-optics technology, for example, detecting markers by using a light-emitting diode (LED) [6][7].

LED has some benefits, such as low power consumption and more brightness compared to the others [8]. LED technology also presents an option for information display on a flat screen, offering a wide viewing angle and a bright and clear image suitable for outdoor applications. LEDs have become the necessary option in lighting environmental renovation. LEDs have many advantages, including high fidelity, high rendering, and supporting green technology [9][10]. In [11] it is reported that LED pulse response behavior, taking into account the junction capacitance and the spreading resistance distributed over the entire junction area, this research makes possible to be able to control LED in the future. LED has recently become a new light source in many areas due to its efficiency and durability. It has the merits of being environmentally friendly and low power consumption. Furthermore, LEDs are used in various sensors and other technological developments, such as LED markers [12][13].

Recently, LEDs cannot only emit continuously, but also dim and blinking. The author in [14] reported the dimming is accomplished by adjusting the average current in the LEDs through pulse width modulation of a switch in series with the load. Meanwhile, [15] suggested that a camera with image processing software is used to recognize the target. The color, size, and variant of blinking of the LED markers detected by the camera is the main limiting factor of this method.

In [16] it is reported that the object recognition system by using single camera. The quality of the camera greatly influences the object detection limits. Meanwhile, the increase of interest and need for the camera generates increasingly complex object tracking algorithms. The object tracking system is often used for security, surveillance, traffic monitoring, navigation, and human-computer interactions [17][18]. Moreover, LED marker is a marker of a particular light. This marker is usually used just before the bullet is fired.

LED needs the electric current to emit the light. The higher the current flowing in the LED, the brighter the light produced. However, it should be noted that the amount of current allowed is 10mA - 20mA and at a voltage of 1.6V-3.5V according to the character of the wavelength. If the current is higher than 20mA, the LED will burn. Changing the electric current on LED according to a certain pattern at a certain time can produce LED blinking [19][20]. In this work, the aim to develop an object recognition system and unique marker by using a variety of LED blinking. The LED spotlight and object recognition based on computer vision were studied, resulting in the fact that the correlation was small.

II. Method

A. Scheme of LED Blinking

The diode forward current can control the LED luminance. The most convenient method for LED dimming without altering the current is the pulse-width-modulation (PWM). PWM regulates the output voltage on the LED to get different average voltages. Meanwhile, PWM adjusts the output voltage

signal on the led to get different average voltages. Two factors limiting the shortest PWM pulse duration are driver response time and LED response time. Typically, driver response time is clearly stated by a manufacturer. LED manufacturers are not aware of the need for faster LED response times. Usually, the manufacturer considers it to be fast enough to satisfy dimming requirements. However, the LED manufacturers do not specify the response time. Therefore, the investigation was needed to evaluate the current state of the response time of visible light LEDs.

In this work, LED is controlled by a microcontroller, which can show the continuous or blinking light, as shown in Fig. 1. Here, the LED is controlled by an Arduino Uno and will generate a PWM. In addition, the Arduino Uno can adjust the blinking signal with a variety of frequencies and duty cycles. For the optimum condition, the maximum speed of the duty cycle on LED depends on the speed rate of spectrometer capture led light.

The spectrometer used in this work has a limitation of speed rate in 10 µs. Therefore, the speed of the duty cycle was set at over 10 µs. The spectrometer has captured the spectrum of LED and converted it to the PWM signal. The comparison between the PWM signal in the input and the output was analyzed to generate the precision and resolution numbers.

B. Object Recognition System

The detection of objects using markers is one of the easiest methods of detection. Markers are artificial objects that are designed to be easily recognized and identified. The marker usually is at a known position or size, serving as a reference point or a measure for other objects. In this work, a camera with image processing software called OpenCV is used to recognize the target. The motor servo is used to steer the laser beam. An external camera is used to recognize the target. This camera can capture pictures with a resolution of 1920 × 1080 pixels for 30 fps and a 76° angle of view. However, the camera also needs a particular program to recognize the object intended as the target. There are several methods of OpenCV that can be used for target recognition. Since the target is dynamic (moving objects), the shape and size of the target, which is recognized by the camera, might slightly change when it moves.

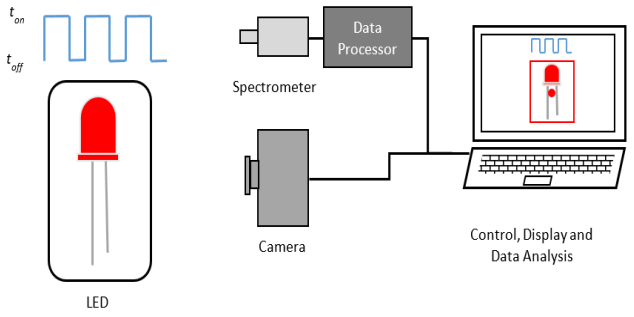


Fig. 1. Encoding LED Blinking Mechanism. A Camera Captured the Display of LED, and a Spectrometer Captured the PWM Signal.

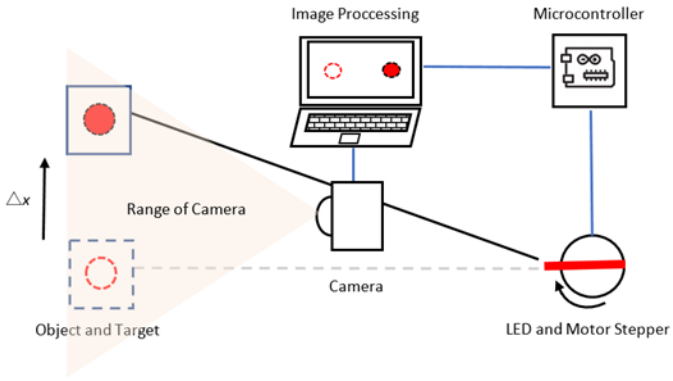


Fig. 2. Object Recognition System consists of the Camera, Image Processing, LED as a Marker Embedded to the Servo Motor and Microcontroller.

The object recognition system consists of multi-colored LED that can emit continuous light, external cameras for capturing an object, and a spectrometer for capturing a PWM signal. When the LED was on, the camera captured the visualization of the colored LED, and the spectrometer captured the PWM signal simultaneously as shown in Fig. 2. Visually, the light looks on and off with the particular period. The PWM signal goes up (t_{on}) and down (t_{off}) on a particular blinking type, and the duty cycle follows the rules as shown in (1). When the light is on, t_{on} is always straight without breaking up, and there is no signal at the t_{off}, vice versa. The blinking system is the combination of both t_{on} and t_{off} that can produce the PWM signal. Typically, the camera will capture the visual object when the light is on (t_{on}) and will be lost when the light is off (t_{off}).

III. Result and Discussion

A. Characteristics of LED Blinking

In this work, LEDs of various colors have been characterized and analyzed for their resolution and precision values. The colors used in this experiment are red, yellow, green, and blue. The duty cycle is the fraction of one period when a signal is active, and its calculation follows (1). Moreover, the PWM duty cycle's accuracy is related to the duty cycle's correction. The precision analysis of Duty Cycle (DC) follows (2).

$$DC = \frac{t_{on}}{t_{on}+t_{off}} x\ 100\% \quad (1)$$

$$Accuracy\ of\ PWM\ DC = \left[1 - \left(\frac{DC_{inp} - DC_{out}}{DC_{inp}}\right)\right] \times 100\% \quad (2)$$

Here, the duty cycle is the ratio of time a load, and the t_{on} and t_{off} are the conditions when the light is on and off. The duty cycle correction for the color LED can be calculated by comparing the input duty cycle and the output duty cycle. The four colors have different wavelengths, resulting in different accuracies. It can be seen on the t_{on} and t_{off}, where the time variations are carried out to see the maximum speed captured by the photo detector and spectrum processing software. The concept of t_{dim} was proposed to capture all the time on Computer Vision as shown in Fig. 3.

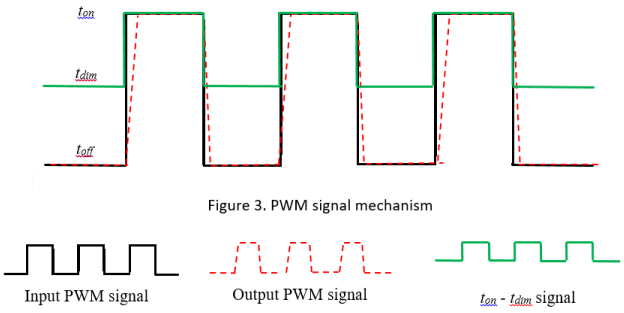


Fig. 3. The Mechanism of PWM Signal, separately by the Various Color.

Initially, the LED has a 50% duty cycle. It means that in the blinking system, the length of time when it turns on is the same as the length of time it turns off. Following this, the PWM signal was analyzed. In this work, the duty cycle is calculated by t_{on} = 1s and t_{off} = 1s. Furthermore, the variety of duty cycles and range of time were analyzed as well.

When the duty cycle is gradually decreased, the accuracy of the duty cycle decreases, as shown in Fig. 4a. The smaller the duty cycle, the t_{off} is higher than a $t_{on,}$ the duty cycle's accuracy drops significantly. Due to the photo detector's response, these phenomena are also affected by the speed ratio (t_{on} and t_{off}). Meanwhile, the duty cycle data after 50% show no significant accuracy shift, as shown in Fig. 4b. Therefore, 50% duty cycle is a simpler form of PWM signal and gives the higher accuracy of the duty cycle. The result indicates that when the LED switched from on to off, the spectrometer still receives the LED beam signal in a short time. The increasing value of the duty cycle correlates with the decrease's correction value. The result shows that at a longer t_{on}, the photo detector will capture the object better. The combination of t_{on} and t_{off}, correction values, and accuracy of duty cycle describes the encoding system for object recognition systems.

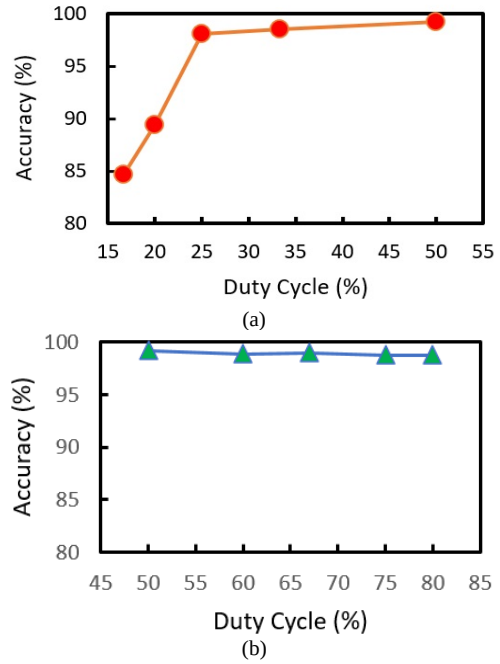


Fig. 4. Duty Cycle Dependency of Accuracy on the Red LED.

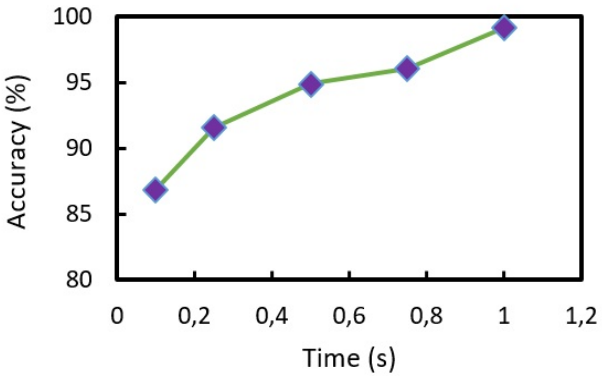


Fig. 5. The t_{on} and t_{off} Dependency of Accuracy on the Red LED under 1 s.

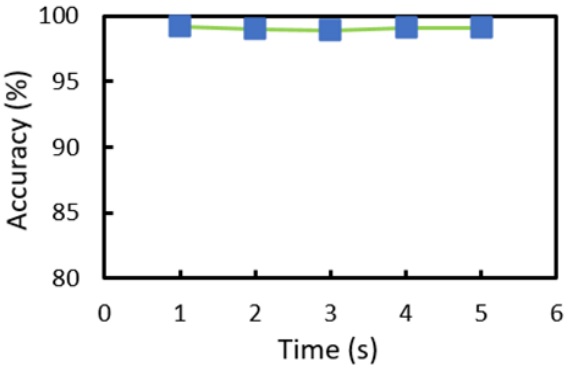


Fig. 6. The t_{on} and t_{off} Dependency of Accuracy on the Red LED after 1 s.

On the data of the 50% duty cycle, the time dependencies of accuracy were analyzed. When the t_{on} and t_{off} are getting smaller, the blinking speed increases, and the accuracy of the duty cycle drops significantly, as shown in Fig. 5. Meanwhile, the variant t_{on} and t_{off} data after one second with the periodic time indicates no significant accuracy shift, as shown in Fig. 6.

Therefore, one second of t_{on} and t_{off} is the cut-off point of time signal and gives the higher accuracy of the duty cycle. Here, all the measurement is on the operating range of the spectrometer. The transformation of the accuracy of the duty cycle due to the variant of t_{on} and t_{off} describes the blinking system. Due to the photo detector's response, these phenomena are also affected by the speed ratio (t_{on} and t_{off}).

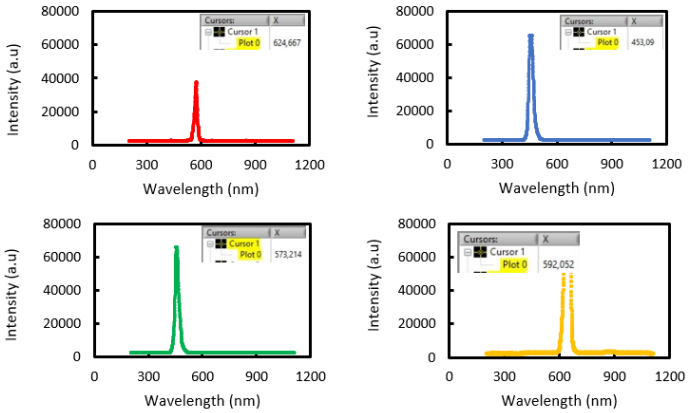


Fig. 7. The Various Wavelength on the Red, Blue, Green and Yellow LED.

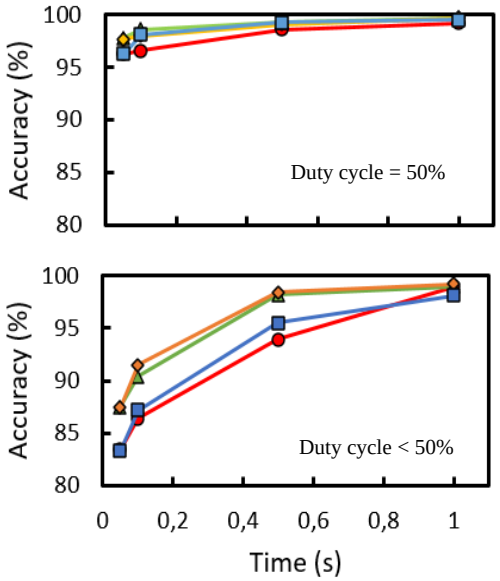


Fig. 8. Time Dependency of Precision Value on the Several Wavelengths LED.

The enhanced measurement was also analyzed with the same method based on the experimental result of duty cycle and variant of t_{on} and t_{off}. Fig. 7 shows that the red LED has 624.67 nm in wavelength. Meanwhile, blue, green, and yellow have 453.09 nm, 573.21, and 592.05, respectively.

On the 50% duty cycle data, the time dependencies of accuracy were analyzed on the various color of the LED, as shown in Fig. 8 (upper side). When the t_{on} and t_{off} are getting shorter, almost the color of LED has lower accuracy, especially at the $t_{on} = t_{off} = 0.05$ s and 0.1 s. When the variant of t_{on} and t_{off} is higher, all the colors of LED get closer to similar accuracy. This result shows LEDs with the various wavelengths has similar accuracy at the highest duty cycle. Moreover, the shortest time range produces the smallest accuracy when the duty cycle is lower than 50%, as shown in Fig. 8 (lower side). Meanwhile, the duty cycle after 50% shows the similar accuracy of duty cycle. The time dependencies accuracy increases significantly when the time range increases periodically. On the t_{on} and t_{off} equal 1 s, almost the variant of LED has a similar accuracy. Note that the red and blue LED has the lowest accuracy in almost all conditions.

B. Object Recognition of LED Blinking

LED with continuous light always keeps and locks on the computer vision as shown in Fig. 8. However, when the LED blinking is on, the LED is on t_{off}, and the camera cannot recognize the object as shown in Fig. 9. In this work, the report evidence that the combination of t_{on} and t_{off} and the value of the duty cycle can lead computer vision always to capture objects and lock them in the visual display.

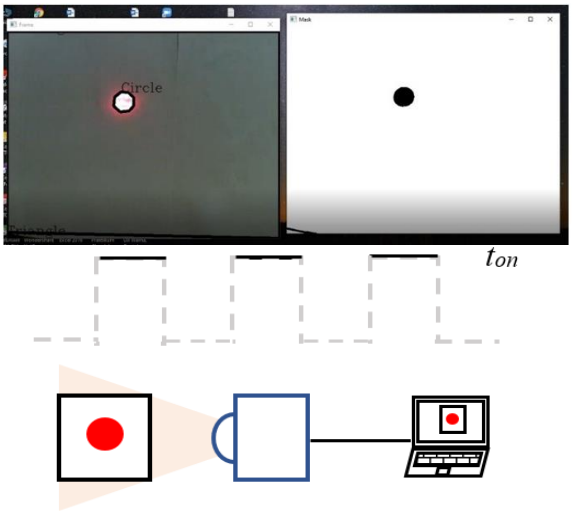


Fig. 9. The t_{on} Condition, the Marker Captured Well by the Camera.

The display in Fig. 9 consists of the original frame and masking the frame. The original frame shows the object and the bounding of object contour for the object recognition system. Then the masking frame display is the original frame with the reducer noise filter. The masking frame is presented in the black and white mode. The object is represented in the white color, and all the background will show in black color area. In this condition, the PWM signal is always in upper position (t_{on}). However, t_{on} condition of LED in range of camera was captured continuously as sown in Fig. 9.

Meanwhile, Fig. 10 shows the original and the masking frame cannot recognize the object contour for t_{off} condition. In this case, the PWM signal is always in under position (t_{on}). However, t_{off} condition of LED in range of camera was not captured continuously. Here, the dotted circle is the representative of the LED position.

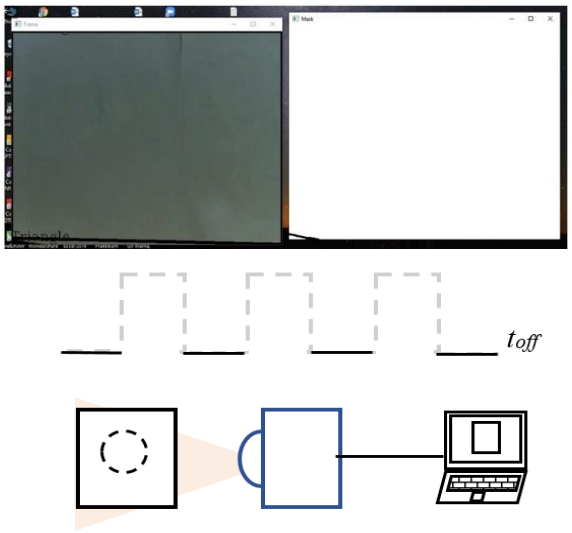


Fig. 10. In the t_{off} Condition, the Marker is not Captured All the Time by the Camera.

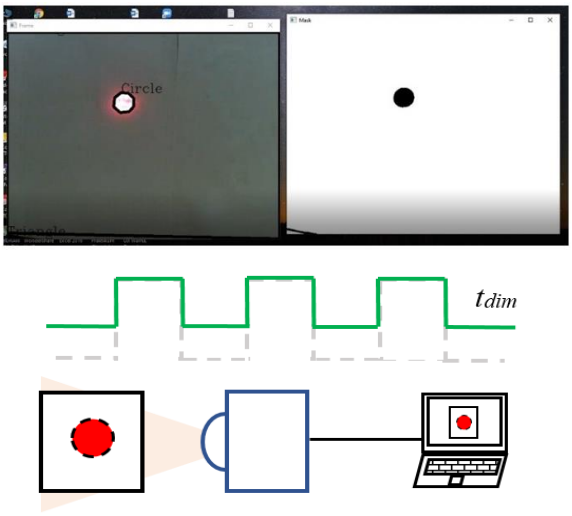


Fig. 11. The Combination between t_{on} and t_{dim}, the Marker Captured Well by the Camera.

The unique blinking of encoding LED as a marker is captured perfectly on the camera as well as shown in Fig. 11. As long as the unique blinking consists of the bright-dim method, a camera always locks the *x*-axis and *y*-axis. However, the bright-dim condition was proposed to solve the t_{off} condition as shown in Fig. 10. In bright-dim condition, the PWM signal is always in upper and middle position. However, t_{bright} - t_{dim} condition of LED in range of camera was captured continuously. LED with bright-dim condition always keeps and locks on the computer vision as shown in Fig. 11.

Fig. 12 shows the unique blinking concept and the evident of it sent in other files. This unique marker will be recognized, and the distractor markers will not be recognized by camera.

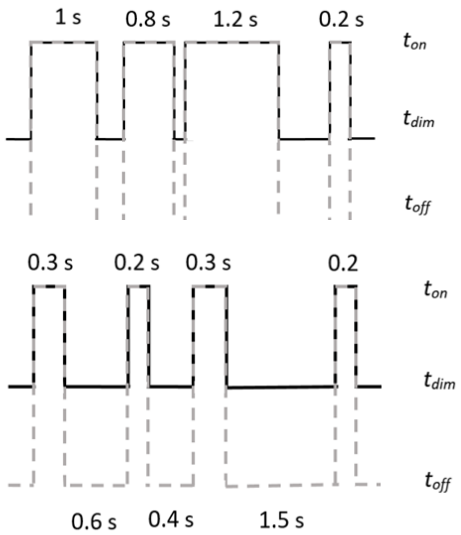


Fig. 12. The Variant of PWM Signals that can be produced to Unique Markers. A Combination of Bright-Dim Captures the Markers All the Time on a Camera.

IV. Conclusion

A new object recognition approach with the unique LED marker is studied rapidly to propose an encoding marker. Camera and image processing would only recognize the unique marker. Characterization LED blinking has been successfully analyzed by controlling the PWM signal with a variety of duty cycles. The accuracy of the duty cycle is one of the parameters to describe the characteristics of LED blinking. The experiment results show that the highest accuracy is obtained on a 50% duty cycle with the range of time 1 s on any LED color. The object recognition system captured all variant LED blinking with the bright-dim method.

This method potentially applied for the future work to the smart weapon in the military system, especially the movement of three-dimensional objects that can applied to air-ground smart missiles.

Acknowledgment

The Author's would like to thank all of member LASER group, Research Center for Physics, National Research and Innovation Agency for the entire instrument that used in this work.

References

[1] B. G. Disha and G. Indumathi, "Automated test system for laser designator and laser range finder," *IJERT*, vol. 4, no. 5, pp. 126-130, 2015.

[2] L. Goldberg, J. Nettleton, B. Schilling, W. Trussel, and A. Hays, "Compact laser sources for laser designation, ranging and active imaging," in *Proc. SPIE* 6552, Orlando, Florida, USA, 2007, pp. 65520G1–65520G8.

[3] H. Kaushal and G. Kaddoum, "Application of laser and tactical military operatins," *IEEE Access*, vol. 5 pp. 20736-20753, 2017.

[4] L. Lazov, E. Teirumnieks, and R. S. Ghalot, "Applications of laser technology in the army," *J. Def. Manag*, vol. 11 Iss. 4 No: 210. 2018.

[5] M. L. W. B. Anderberg, "Laser weapons: The dawn of a new military age," New York, USA: Springer, 2013.

[6] A. K. Maini, "Battlefield lasers and opto-electronics systems," *Def. Sci. Journal*, vol. 60, no. 2, pp. 189-196. 2010.

[7] P.H. Putman, "When old is new again," *Video Systems* pp. 37–43, 2002.

[8] E.F. Schubert, "Light-emitting diodes," Cambridge University Press, 2003, p. 313.

[9] G. Leschhorn and R. Young, "Handbook of led and ssl metrology, instrumentation system," GmbH, 2017.

[10] H. I Hsieh, H. W. "LED current balance using a variable voltage regulator with low dropout ds control," *Applied Sciences*, pp. 1-14. 2017.

[11] I. Hino and K. Iwamoto. "LED pulse response analysis considering the distributed CR constant in the peripheral junction," *IEEE Trans. on Elec. Dev*. Vol. 26, pp. 1238-1242. 1979.

[12] W. Kurdthongmee, and T. Lamsub. "An automatic system for non-uniform brightness compensation of LED arrays: image processing routines to locate LED centers," *Walailak J. Sci & Tech*. vol. **5,** no. 2, pp. 203-216. 2008.

[13] L. Svilainis. "LED PWM dimming linearity investigation. displays," vol. 29, pp. 243–249, 2008.

[14] P. Nara, and D. S. Zinger. "An effective LED dimming approach," *in IAS Annual Meeting - IEEE Industry Applications Society*, vol. 3, pp. 1671-1676, November 2004.

[15] K. Adi, A. P. Widodo, C. E. Widodo, A. Pamungkas, and A. B. Putranto, "Automatic vehicle counting using background subtraction method on gray scale images and morphology operation," *in OP Conf. Series: J. of Physics: Conf. Series 1025*, 012025, 2018.

[16] R. T. Yunardi, A. W. Mardhiyah, M. H. Yahya, and F. C. S. Arisgraha, "Desain dan implementasi visual object tracking menggunakan *Pan and Tilt* vision system," *ELKHA*, vol. 11, no. 2, pp. 85 – 92, 2019.

[17] B. Micusik, and T. Pajdla, "Simultaneous surveillance camera calibration and foot-head homology estimation from human detections," *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1562-1569, 2010.

[18] H. G. Schantz, "A real-time location system using near-field electromagnetic ranging," *2007 IEEE Antennas and Propagation Society International Symposium*, Honolulu, Hawaii, USA, pp. 3792-3795, 2007.

[19] J. Biswas, and M. Veloso, "Depth camera based indoor mobile robot localization and navigation. *In 2012 IEEE International Conference on Robotics and Automation*, pp. 1697-1702, 2012.

[20] H. Adinanta, H. Kato, A. W. S. Putra, and T. Maruyama, "Enhancement of beam tracking response using color filtering method for optical wireless power transmission," *AIP Conference Proceedings 2256*, vol. 1, no. 6, 2020.

Inherent Feature Extraction and Soft Margin Decision Boundary Optimization Technique for Hyperspectral Crop Classification

Mr. M.C.Girish Babu, Dr. Padma M.C
Department of Computer Science and Engineering
PES College of Engineering
Mandya, India

***Abstract*—Crop productivity and disaster management can be enhanced by employing hyperspectral images. Hyperspectral imaging is widely utilized in identifying and classifying objects on the ground surface for various agriculture application uses such as crop mapping, flood management, identifying crops damaged due to flood/drought, etc. Hyperspectral imaging-based crop classification is a very challenging task because of spectral dimensions and poor spatial feature representation. Designing efficient feature extraction and dimension reduction techniques can address high dimensionality problems. Nonetheless, achieving good classification accuracies with minimal computation overhead is a challenging task in Hyperspectral imaging-based crop classification. In meeting research challenges, this work presents Hyperspectral imaging-based crop classification using soft-margin decision boundary optimization (SMDBO) based Support Vector Machine (SVM) along with Image Fusion-Recursive Filter (IFRF) and Inherent Feature Extraction (IFE). In this work, IFRF is used for reducing spectral features with meaningful representation. Then, IFE is used for differentiating physical properties and shading elements of different objects spatially. Both spectral and spatial features extracted are trained using SMDBO-SVM for performing hyperspectral object classification. Using SMDBO-SVM for Hyperspectral object classification aid in addressing class imbalance issues; thus, the proposed IFE-SMDBO-SVM model achieves better accuracies and with minimal misclassification in comparison with state-of-art statistical and Deep Learning (DL) based Hyperspectral object classification model using standard dataset Indian Pines and Pavia University.**

Keywords*—Crop classification; decision boundary; deep learning; dimensionality; feature selection; hyperspectral image; support vector machines

I. Introduction

Agriculture plays a very important role in improving major developing country economy. High productivity of food yield will aid in meeting food security. However, with global warming, it is hard to achieve a very high yield. A significant amount of crops is lost worldwide due to natural disasters such as drought, cyclones, and floods; leading to loss of life of farmers/people. A farmer requires timely relief of funds for disaster management. Allocating the right kind of funds is challenging, as farmer grows multiple crops within the same region. Thus, efficient crop identification methodologies are needed. Hyperspectral Imaging (HSI) is an efficient method used for crop identification. Extensive work has been done in recent times for crop recognition in agriculture environment such as Locally Adaptive Dictionary (LAD) through Multiscale Joint Collaborative Feature (MJCF) [1], spatial-spectral feature extraction through end-to-end deep learning framework [2], neural network learning framework for extracting adaptive Spatial-Spectral Features [3], Improved CNN framework combining Markov random fields for extracting spatial-spectral feature [4], Conditional Random Field and Deep Metric Learning for HSI classification [5]. However, following challenges such as high dimension size, presence of noise, and high similarity among spectral features, shapes, textures of different crops must be addressed in building an effective hyperspectral imaging-based crop classification method. Hyperspectral imaging consists of hundreds of Narrow Bands that are continuous with high spectral correlation. Thus, results in Hughes phenomenon, space, and computation complexity as shown in following work such as band selection through End-to-End deep learning architecture [7], hierarchical spatial-spectral feature maps through CNN [8], spectral-spatial feature extraction using CNN and information measure [9], and Active learning-based CNN model [10], a hybrid model combining Inception and Deep Residual Network [11] for HSI crop classification.

The crop classification accuracies can be improved by the utilization of feature extraction and feature selection methods. Existing methodologies predominantly used Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for reducing the feature size of hyperspectral images [12]. The ICA-based hyperspectral crop classification methodologies assure the extracted feature is independent; nonetheless, ICA induces high computation overhead and doesn't guarantee to retain spatial information. On the other side, PCA-based hyperspectral crop classification methodologies realize good classification accuracy when compared with ICA- based methodologies. The PCA- based methodologies aid in assuring stabilizing features with a limited size of high meaningful representation. Nonetheless, PCA-based HSI crop classification methodologies are not efficient in retaining useful spectral features. Thus, for retaining spectral features more efficiently Image fusion (IF) methodologies are used in recent work. However, IF-based methodologies achieve poor classification performance; this is because they are affected due to the presence of noise and

mixed pixel due to different illumination and climatic conditions [13], [14].

Recently, Deep Learning (DL) methodologies [15], [16] have been adopted for HSI crop classification [17], [18] with good accuracies [19], [20] which is studied in literature survey section. However, these DL-based methodologies induce high computation overhead and require a higher number of training parameters [21]. Further, induces high misclassification when data is imbalanced. For overcoming research problems it is important to extract meaningful features both spectrally as well as spatially; further, it is important to eliminate shading features from crop inherent features to classification accuracies. Here we used image fusion and recursive filter (IFRF) [22], [23] for obtaining semantic features across different bands i.e., spectrally. The usage of IFRF aided in reducing feature size with meaningful representation. Then we present an inherent feature extraction (IFE) method for distinguishing physical properties and shading elements of different crops. Existing models are trained using a Support vector machine (SVM) [18] for performing crop classification; the classification accuracies using SVM are affected due to misclassification [19]; especially when data is imbalanced and two objects exhibit similar physical features [20]. Thus, to address data imbalance issues and reduce misclassification in this we introduced a soft-margin decision boundary optimization model for SVM. The SMDBO-SVM based crop classification model aided in achieving high classification with less misclassification in comparison with the deep learning-based classification model.

The significance of using IFE-SMDBO-SVM is described below:

- Presented effective spatial-spectral feature extraction mechanism namely IFE. The IFE model can extract semantic features even under different illumination and climatic conditions.
- Presented soft-margin decision boundary optimization model for performing classification when HSI data exhibit data imbalance and also under mixed pixel environment.
- SMDBO-SVM based HSI achieves high classification accuracies with less misclassification (i.e., Kappa coefficient) in comparison with recent deep-learning-based HSI classification models.
- The SMDBO-SVM based HSI classification model reduces computation overhead in comparison with deep-learning-based HSI classification models.

The rest of the paper is organized as follows. Section II discusses various existing hyperspectral crop classification models and establishes the benefits and limitations, and hypotheses of the proposed method. Section III presents Inherent feature extraction and soft margin decision boundary optimization for Hyperspectral image-based crop classification methodology. In section IV, the performance efficiency of IFE-SMDBO in comparison with the existing HSI classification methodology is discussed. In the last section, the benefit of IFE-SMDBO is discussed and the future direction of work is discussed.

II. Literature Survey

This section presents some of the recent methodologies presented for performing crop classification using the hyperspectral image. In [2] presented HS-CNN (Hybrid Spectral Convolutional Neural Network) based HSI classification model. They first employed 3D-CNN for extracting spectral-spatial information followed by 2D-CNN. The classification performance of the HSCNN model heavily relies on both spectral and spatial information of HSI. The HS-CNN model can joint retain spatial-spectral feature sets from different brands. The hybrid CNN model aids in learning more abstract level spatial features with minimal overhead in comparison with the 3D-CNN model. In [6], presented recurrent neural network (RNN)-based HSI classification. Here the spectral information is considered as a sequence; however, they showed standard RNN models are difficult to train and are not efficient as spatial features are not used. Thus, they presented Shorten Spatial-spectral RNN Parallel-GRU (St-SS-pGRU) by combining convolution layers to achieve better HSI classification performance. In [16] showed that the 2D CNN just focused on extracting spatial features; however, neglects to extract spectral features. Similar, to [2], [21] presented a 3D CNN model that jointly considers extracting both spatial and spectral features; however, with reduced computational overhead by distributing spatial and spectral features extraction across different layers. In [15] showed CNN is widely used for HSI classification; However, they significantly because of high misclassification at the pixel level. In particular at the edges of neighboring crops; this is because the impact of adjacent pixels crops is different from target pixels. To address this, here they presented a center attention network (CAN) for HSI classification. The CAN-HSI can extract spatial and spectral features of both target pixel and adjacent pixels together in a simultaneous manner for performing HSI classification. In CAN major importance is given to highly correlated features concerning target pixels; thus aiding HSI classification performance. Further, CAM reduces parameters through a weighted sum of spatial and spectral features to reduce computation overhead without compromising on HSI classification accuracies.

In [14] showed DL-based method generally use patch-wise learning architecture for HSI classification. In recent times fast patch-free global learning (FPGA) frameworks have been modeled for HSI classification considering global spatial contextual information. Nonetheless, when HSI data is imbalanced the FPGA-based HSI classification finds it difficult in extracting discriminative features. To address they presented "spectral-spatial dependent global learning (SSDGL) architecture employing global joint attention (GJA) technique and global convolution LSTM (GCLSTM). In SSDGL for addressing data imbalance issues employed hierarchical tradeoffs sampling solution and weighted softmax loss function are modeled. The GCLSTM model is used for extracting LSTM dependencies of spectral features and later these dependencies are used for distinguishing spectral features for crop types. The GJA model is used for extracting attention areas for identifying the most discriminating features. In [17],

for improving the robustness of standard machine learning models, recent work has emphasized integrating traditional ML models into DL methodologies. Here they studied modeling Deep SVM (DSVM) for HSI classification. The DSVM is modeled by implementing four kernel functions as polynomial, neural, Gaussian radial basis function, and exponential radial basis function. The standard SVM model is used for interconnecting weights of the entire network; the interconnecting weights act as a regularization parameter.

The research hypothesis the problems that existing SVM-based Hyperspectral object classification [18] are modeled using hard margin decision boundary [19], [20]; thus, high induce misclassification for smaller classes. Thus, are not efficient when data is imbalanced and induce high computational overheads. Further, the classification outcome is improved through better representation of spatial and spectral information; the proposed research work addresses the aforementioned problems in designing a better hyperspectral object classification model in the next section.

III. Inherent Feature Extraction and Soft Margin Decision Boundary Optimization for Hyperspectral Image-based Crop Classification

This work presents the inherent feature extraction (IFE) and Soft Margin Decision Boundary Optimization (SMDBO) Technique for Hyperspectral Crop Classification. Here first the working model of IFE-SMDBO based hyperspectral crop classification is presented. Second, present an inherent feature extraction model to reduce spectral features and exploit inherent features spatially to distinguish between actual crop and shadowing elements. Then, it discusses the standard SVM model used for HSI classification and highlights its limitation. Finally, present an improved decision boundary mechanism namely soft-margin decision boundary optimization (SMDBO) for addressing data imbalance and mixed pixel problems in HSI classification.

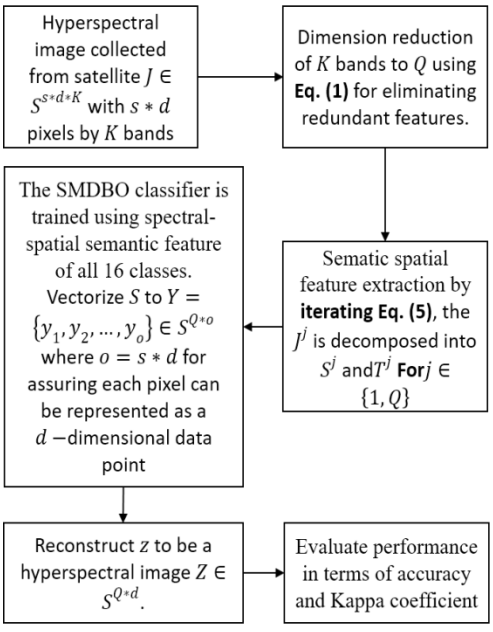


Fig. 1. Block Diagram of IFE-SMDBO based Hyperspectral Crop Classification Technique.

A. Working Model of IFE-SMDBO based Hyperspectral Crop Classification

The step involved in the proposed hyperspectral crop classification using the IFE-SMDBO model is shown in Fig. 1. The first step loads the HSI and reads the entire band information. Second, the HSI size is reduced spectrally using Eq. (1). Third, semantic features are extracted in an iterative manner using Eq. (5). Then, the sematic spatial-spectral feature bare trained using modified SMDBO and performs classification. Finally, the image is reconstructed for validating the accuracy of HSI classification models. The algorithm of the proposed IFE-SMDBO model is shown in Algorithm 1.

Algorithm 1. IFE-SMDBO based hyperspectral crop classification technique.

Input. Hyperspectral image collected from satellite $J \in S^{s*d*K}$ with $s*d$ pixels by K bands.

Output. A classified outcome (i.e., labeled Hyperspectral image Z).

Step 1. Start.

Step 2. Dimension reduction of K bands to Q using **Eq. (1).**

Step 3. For$j \in \{1, Q\}$ do

Step 4. By **iterating Eq. (5)**, the J^j is decomposed into S^j andT^j.

Step 5. End for.

Step 6. Vectorise S to $Y = \{y_1, y_2, \dots, y_o\} \in S^{Q*o}$ where $o = s*d$ for assuring each pixel can be represented as a $d-$dimensional data point.

Step 7. For obtaining labels $z \in S^o$**soft-margin decision boundary optimization-based support vector machine learning (SMDBO-SVM) algorithm** is used.

Step 8. Reconstruct z to be a hyperspectral image $Z \in S^{Q*d}$.

Step 9. Stop.

B. Band Selection and Feature Extraction

Effective selection of band plays a very important role in achieving high accuracies with minimal computation overhead for performing crop classifications. Existing HSI classification methodologies used PCA for reducing band size; however, PCA fails to provide a higher number of useful features. Let consider hyperspectral data with K bands which are reduced to Q bands. Here we employ IFRF [23] for reducing band size and assuring eliminating noisy and redundant pixels spectrally through the following equation.

$$J^l = \frac{\sum_{m=(l-1)n+1}^{ln} J^m}{n},\ l = \left\lfloor \frac{K}{Q} \right\rfloor, \quad (1)$$

where n defines the sub-group band size considered, l defines band indices of reduced spectral bands, m defines band indices of actual spectral bands, and $\lfloor \cdot \rfloor$ a value closer to $-\infty$.

Next, the inherent properties of different crops are extracted by eliminating shading elements in obtaining high-quality features spatially using the following equation.

$$J_q = S_q T_p, \tag{2}$$

where q defines pixel index, $T \in \mathcal{S}^{s*d}$ describes inherent features shading component, $S \in \mathcal{S}^{s*d}$ defining inherent feature component, and $J \in \mathcal{S}^{s*d}$ defines intensity feature. The variable S_q and T_p in the above equation are unknown; however, J_q is a known variable. Generally, the reflectance value changes rapidly in edges and remains constant otherwise; similarly, the pixel with the same value will have the same reflectance value. Keeping the aforementioned context in consideration the S_q is computed as follows

$$S_q = \sum_{r \in \mathcal{O}(q)} b_{qr} S_r, \tag{3}$$

where r defines pixel index and b_{qr} affinity matrix features for measuring similarities between J_q and J_r. the adjacent pixel obtained through Gaussian window as follows

$$GW = \exp\left(-\frac{\|q-r\|_2^2}{2\sigma^2}\right) \tag{4}$$

and σ defines the size considered. Further, defining affinity graph (AG) plays a very essential part in semantically extracting inherent characteristics. Using Eq. (2) and Eq. (4), the meaningful feature is extracted through linear properties as follows

$$\begin{cases} S_r = \sum_{r \in \mathcal{O}(q)} b_{qr} S_r, \\ \tilde{T}_q = \frac{1}{J_q} S_r, \end{cases} \tag{5}$$

where $\mathcal{O}(q)$ defines neighbor pixel q, $\tilde{T}_q = \frac{1}{J_q}$ after obtaining the estimated value of S_r and T_q. Thus each pixel physical properties of different crops are retained, where shading properties are not related to semantic feature sets properties and using inherent features the spatially useless feature can be eliminated.

C. SVM Classification

The feature space is represented as $Y \in S^e$, the index of different crops are represented as $Z = \{-1 + 1\}$, and respective crop distribution over $Y * Z$ is repressed as E. Let us consider that there are o feature points in respective hyperspectral data and n training features as described below

$$T = \{(y_1, z_1), (y_2, z_2), \dots, (y_n, z_n)\}, \tag{6}$$

Here training features selected are identical based on the distribution of E. For predicting the sample considered the following function is defined

$$f(y) = x^U \alpha(y), \tag{7}$$

where x represent the forecaster, $\alpha(y)$ represent corresponding feature mapping of y to kernel L as described below

$$L_{jk} = \alpha(y_j)^U \alpha(y_k). \tag{8}$$

In precise Y represent the matrix where its j^{th} column is $\alpha(y_k)$ which is defined as follows

$$Y = [\alpha(y_1), \alpha(y_2), \dots, \alpha(y_n)], \tag{9}$$

And z is its column vector which is defined as follows

$$y = (z_1, z_2, \dots, z_n)^U \tag{10}$$

The classification margin for describing a crop feature is computed using the following equation

$$\beta_j = z_j x^U \alpha(y_j), \;\; j = 1,2,3, \dots, n. \tag{11}$$

The state-of-art SVM based classification model generally considers that crop features are separable and the hyperplane has the capability in distinguishing the training crop features T with no errors; thus, the SVM classification margin is obtained using the following equation

$$\min_x \tfrac{1}{2}\|x\|^2 \;\; such\ that\ z_j x^U \alpha(y_j) \geq 1, \;\; j = 1,2,3, \dots, n, \tag{12}$$

That maximizes its minimum margin.

D. Soft Margin Decision Boundary Optimization Model

Using above Eq. (12) will lead to high misclassification when used for classifying crops under a mixed cropping environment, when crops exhibit similar features, and when data is imbalanced. Further, there exist scenarios where a very limited feature is available for some crops and the high number of features for other crops; leading to concept drift and data imbalance issues. Using a hard-margin-based SVM classification model defined in the above equation gives a very poor result. Thus, for addressing this paper introduce soft margin decision boundary optimization SVM (SMDBO-SVM) for classifying crops considering concept drift and data imbalance issues. The SMDBO-SVM model optimizes the margin/boundary by minimizing the margin difference and simultaneously maximizing the margin average. Using Eq. (11), the margin difference is computed as follows

$$\hat{\beta} = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \left[z_j x^U \alpha(y_j) - z_j x^U \alpha(y_j) \right]^2$$

$$= \tfrac{2}{n^2} (n x^U Y Y^U x - x^U Y x x^U Y^U x). \tag{13}$$

similarly, the margin means is obtained as follows

$$\bar{\beta} = \tfrac{1}{n} \textstyle\sum_{j=1}^{n} z_j x^U \alpha(y_j) = \tfrac{1}{n} (Yz)^U x, \tag{14}$$

The Eq. (2) decision boundary can be optimized using the following equation

$$\min_x \tfrac{1}{2}\|x\|^2 + \delta_1 \hat{\beta} - \delta_2 \bar{\beta} \text{such that } \alpha(y_j) \geq 1, \;\; j = 1,2,3, \dots, n \tag{15}$$

where δ_1 and δ_2 are parameters used for bringing good tradeoffs.

In non-distinguishable scenarios, the training crop features T can't be distinguished with zero error and ideal hyperplane can't be obtained by minimizing objective function (tradeoffs model of error minimization and margin maximization). For addressing in SMDBO-SVM the error minimization are an

additional parameter used for penalizing misclassified crop features, which is described below

$$\min_{x,\mu} \frac{1}{2}\|x\|^2\, \delta_1\hat{\beta} - \delta_2\bar{\beta} + D\sum_{j=1}^{n}\mu_j \text{ such that} z_j x^U\, \alpha(y_j) \geq 1-\mu_j,\ \mu_j \geq 0,\ \ j = 1,2,3,\dots,n, \quad (16)$$

where μ represent the slack parameter for quantitating feature loss and is computed defined as follows

$$\mu = [\mu_1, \mu_2, \dots, \mu_n]^U \quad (17)$$

D represent the regularization variable that is used for controlling the penalty given to misclassifications. The higher the error, the higher penalty is given to it.

Using Eq. (13) and Eq. (14) into Eq. (16) will result in quadratic programming problem as follows

$$\min_{x,\mu} \frac{1}{2}x^U x + \frac{2\delta_1}{n^2}(nx^U YY^U x - x^U Yzz^U Y^U x) - \delta_2\frac{1}{n}(Yz)^U x + D\sum_{j=1}^{n}\mu_j \text{ such that } z_j x^U \alpha(y_j) \geq 1-\mu_j,\ \mu_j \geq 0,\ \ j = 1,2,3,\dots,n. \quad (18)$$

The ideal forecasting/prediction model x^* for the optimization problem of Eq. (18) is defined using the following equation

$$W^* = \sum_{j=1}^{n}\varphi_j\alpha(y_j) = Y\varphi, \quad (19)$$

where φ described its coefficient which is described as follows

$$\varphi = [\varphi_1, \varphi_2, \varphi_3, \dots \varphi_n]. \quad (20)$$

Using Eq. (19) into Eq. (18), Eq. (18) is updated as follows

$$\min_{\varphi,\mu} \frac{1}{2}\varphi^U R\varphi + q^U + D\sum_{j=1}^{n}\mu_j \text{ such that } z_j\varphi^U L_j \geq 1-\mu_j,\ \mu_j \geq 0,\ \ j = 1,2,3,\dots,n, \quad (21)$$

where the parameter R is computed as follows

$$R = 4\delta_1\frac{[nL^U L - (Lz)(Lz)^U]}{n^2+L}, \quad (22)$$

Then, q is computed as follows

$$p = \frac{-\delta_1 Lz}{n}, \quad (23)$$

The kernel matrix L is computed as follows

$$L = Y^U Y, \quad (24)$$

and L_j represents the j^{th} column of kernel matrix L.

In general, the state-of-art SVM model are used for binary classification purpose; however, in this, we propose SMDBO-SVM as a multiclass classifier. Let consider $Z = \{1,2,3,\dots,m\}$ as a set of crop classes in hyperspectral data; then, $m(m-1/2)$ hyperplane is built all probable pairwise classifier using SMDBO-SVM. Here the SMDBO-SVM model first performs binary classification among two classes j and k through discriminant function $f_{jk}(y) \in \{-1,1\}$ where $j \neq k$ and belongs to Z. Further, it is important for computing weighted function $T_j(y_q)$ for respective individual class $j \in Z$, before making any decision of predicted value y_q. Thus, a weighted strategy is modeled for distinguishing different crops from one another is described as follows

$$T_j(y_q) = \sum_{\substack{k=1 \\ k\neq j}}^{m}\text{sign}\{g_{jk}(y_q)\}, \quad (25)$$

where sign(·) represents the sign function used for binary representation of value. The decision of classified crop y_q is done based on the highest weighted crops as described below

$$j^* = \arg\max_{j\in Z}\{T_j(y_q)\}. \quad (26)$$

The semantic feature extracted from HSI data is trained using the SMDBO-SVM model to aid in attaining better crop classification performance in comparison with the state-of-art crop classification model which is experimentally proven below.

IV. Simulation Analysis and Results

This section evaluated the effectiveness of IFE-SMDBO-SVM based hyperspectral crop classification over various recent state-of-art hyperspectral classification model [1], [4], [6], [10], [14], [15], and [17]. Total two publically available benchmarks HSI datasets such as Indian pines and Pavia University are used for analyzing HSI classification models. The performance of different classification models is measured using the most widely used metric in many existing HSI classification models such as average accuracy, overall accuracy, Kappa coefficient, and computation time. Attaining a higher accuracy value of accuracy and higher value of Kappa coefficient indicated good performance. Alongside, reducing time indicates the model is suitable for real-time deployment.

TABLE I. The Ground Truth Data of the Indian Pines Dataset with 16 Classes

Number	Classes	Total Samples
1	Alfalfa	46
2	Buildings Grass Trees Drives	386
3	Corn notill	1428
4	Corn mintill	830
5	Corn	237
6	Grass pasture	483
7	Grass trees	730
8	Grass pasture moved	28
9	Hay windrowed	478
10	Oats	20
11	Soybean notill	972
12	Soybean mintill	2455
13	Soybean clean	593
14	Stone Steel Towers	93
15	wheat	205
16	woods	1265

A. Dataset Description

The Indian Pines dataset is collected through an AVIRIS sensor deployed over the northern-west side of Indiana. The hyperspectral data is collected by setting a wavelength of $0{:}4-2{:}5\times10^{-6}$ meters with 224 bands and 145×145 pixels. The reason for using IP is because the majority of the area covered is the agriculture environment i.e., $2/3^{rd}$ and the remaining $1/3^{rd}$ measured areas are forest and other vegetation that is grown naturally. Further, IP data encompasses small roads, houses, low-lying buildings, and two-lane highways. Alongside, there are crops with early stages of growth which is less than 5% of overall data collected in IP. The ground truth data is composed of a total of 16 crops (i.e., labels) as shown in Table I. Similar, to [14]-[17], the water absorption bands are eliminated and spectral bands size are reduced to 200.

The Pavia University hyperspectral data is collected through the ROSIS sensor. The PU dataset is measured with a spatial resolution of $1{:}3$ meters, with total 103 spectral bands, and composed of 610×610 pixels. Before analysis some data are eliminated they don't provide any information similar to [14]-[17]. The ground truth data is composed of total 9 classes as shown in Table II.

TABLE II. THE GROUND TRUTH DATA OF PAVIA UNIVERSITY DATASET WITH 9 CLASSES

Number	Classes	Total Samples
1	Asphalt	6631
2	Bitumen	1330
3	Bare − S	5029
4	Gravel	2099
5	Meadows	18649
6	Painted − M − S	1345
7	Shadow	947
8	Sum	42776
9	Self − B − B	3682

B. Comparative Analysis for IndianPines Dataset

Here experiment is conducted using Indian Pines Dataset for validating the performance achieved using IFE-SMDBO-SVM and other state-of-art HSI crop classification methods such as CNN-AL-MRF, CAM, FPGA, SSDGL, and DSVM. The accuracies achieved for different classes of an object by the individual model are shown in Table III. From the experiment, it can be seen the proposed IFE-SMDBO-SVM achieves much better results than other HSI crop classification methods such as CNN-AL-MRF, CAM, FPGA, SSDGL, and DSVM in terms of accuracies and Kappa coefficient. Further, the IFE-SMDBO-SVM induces very little computation overhead in comparison with CNN-AL-MNF.

TABLE III. COMPARATIVE ANALYSIS OF IFE-SMDBO-SVM OVER RECENT HSI CROP CLASSIFICATION METHODOLOGY FOR INDIAN PINES DATASET

Class name	CNN-Al-MNF (2020) [10]	CAM (2021) [15]	FPGA (2020) [14]	SSDGL (2021) [14]	DSVM (2020) [17]	IFE-SMDBO-SVM
Alfalfa	92.71	87.8	97.22	100	**100**	**100**
Corn notill	92.98	98.05	93.07	99.63	**100**	99.98
Corn mintill	88.7	97.99	89.46	99.24	**100**	99.97
Corn	97.7	94.37	**100**	**100**	**100**	**100**
Grass pasture	92.9	98.39	95.63	99.56	99.43	99.56
Grass trees	98.89	99.7	97.56	**100**	98.89	99.88
Grass pasture moved	76.74	**100**	**100**	**100**	**100**	**100**
Hay windrowed	97.87	**100**	**100**	**100**	98.72	**100**
Oats	38.89	77.78	100	100	100	99.97
Soybean notill	92.27	98.17	96.64	99.68	95.75	99.41
Soybean mintill	95.07	98.33	96.74	99.36	100	99.46
Soybean clean	90.51	97.94	91.65	99.11	99.63	**100**
wheat	96.53	**100**	**100**	**100**	**100**	99.85
woods	99.28	98.77	99.91	**100**	**100**	**100**
Buildings Grass Trees Drives	88.4	92.51	99.72	100	95.45	**99.87**
Stone Steel Towers	97.12	98.81	**100**	**100**	**100**	**100**
OA (%)	98.79	98.1	96.18	99.63	98.86	**99.7**
AA (%)	94.28	96.16	97.33	99.79	99.24	**99.87**
Kappa (%)	-	97.84	95.64	99.58	-	**99.66**
Time (s)	8109.34	-	-	-	-	**12.5**

C. Comparative Analysis for Pavia University Dataset

Here experiment is conducted using Pavia University dataset for validating the performance achieved using IFE-SMDBO-SVM and other state-of-art HSI crop classification methods such as CNN-AL-MRF, CAM, FPGA, SSDGL, and DSVM. The accuracies achieved for different classes of an object by the individual model are shown in Table IV. From the experiment, it can be seen the proposed IFE-SMDBO-SVM achieves much better results than other HSI crop classification methods such as CNN-AL-MRF, CAM, FPGA, SSDGL, and DSVM in terms of accuracies and Kappa coefficient. Further, the IFE-SMDBO-SVM induces very little computation overhead in comparison with CNN-AL-MNF.

TABLE IV. COMPARATIVE ANALYSIS OF IFE-DMDBO-SVM OVER RECENT HSI CROP CLASSIFICATION METHODOLOGY FOR PAVIA UNIVERSITY DATASET

Class name	**CNN-Al-MNF (2020) [10]**	**CAM (2021) [15]**	**FPGA (2020) [14]**	**SSDGL (2021) [14]**	**DSVM (2020) [17]**	**IFE-SMDBO-SVM**
Asphalt	82.88	99.54	97.83	**100**	99.55	99.96
Meadows	**100**	99.78	99.95	**100**	99.36	**100**
Gravels	98.32	93.15	91.28	**100**	99.43	**100**
Trees	99.76	98.63	95.85	99.67	99.45	**100**
Painted metal sheets	99.85	**100**	**100**	**100**	95.64	99.97
Bare soil	**100**	99.78	99.76	**100**	**100**	**100**
Bitumen	98.83	98.08	99.73	**100**	97.66	**100**
Self-blocking bricks	**100**	96.06	98.05	99.92	98.92	**100**
Shadows	-	99.89	97.86	**100**	99.11	99.92
OA (%)	99.15	98.97	98.68	99.97	98.17	**99.98**
AA (%)	97.45	98.32	97.82	99.95	98.79	**99.98**
Kappa (%)	-	98.64	98.25	99.96	-	**99.98**
Time (s)	1378.57	-	-	-	-	**8.5**

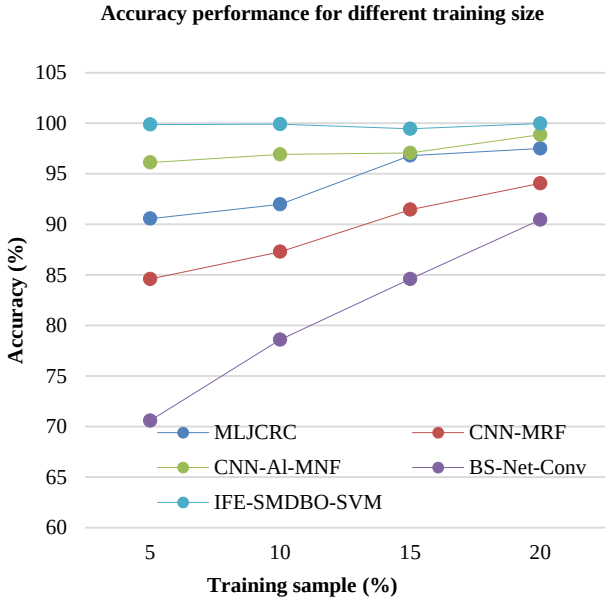


Fig. 2. The Classification Outcome was attained using IFE-SMDBO-SVM and Various Existing Classification Methods.

D. Effect of Varying Training Sample size

This section presents a comparative analysis of the proposed IFE-SMDBO-SVM classification over state-of-art HSI classification methodologies considering the effect training sample size. Here the training sample size is varied from 5 to 20% and the experiment is conducted as shown in Fig. 2. The existing HSI classification methods are MLJRC (Yang et al., 2018) [1], BS-Net-Conv (Cai et al., 2020) [7], and CNN-MBF (Cao et al., 2018) [4], and CNN-Al-MNF (Cao et al., 2020) [10]. From the result achieved it can be seen that among deep learning methodologies CNN-Al-MNF achieves very good performance with accuracies of 96.12% and 99.41% considering the training sample size of 5% and 20%, respectively. On the other, BS-Net-Conv achieves very poor performance with accuracies of 70.58% and 90.45% considering the training sample size of 5% and 20%, respectively. The IFE-SMDBO-SVM achieves very good performance accuracies with accuracies of 98.7% and 99.97% considering the training sample size of 5% and 20%, respectively when compared with other state-of-art HSI classification algorithms such as MLJRC, BS-Net-FC, BS-Net-Conv, and CNN-MBF. From the result, the SFR-HSI is very efficient when there is a very limited training sample available.

E. Effect of Inherent Feature Extraction Method

This section evaluates the effect of using IFE in a classification task. The classification accuracies obtained by IFE-HSI and other existing HSI crop classification methodologies are graphically shown in Fig. 3. The effect of using and not using IFE is shown in Fig. 4. From Figure, it can be seen how IFE aids the classification accuracies enhancement. Thus, it can be stated the IFE-SMDBO-SVM model can learn crop inherent features more efficiently by eliminating the shadow component.

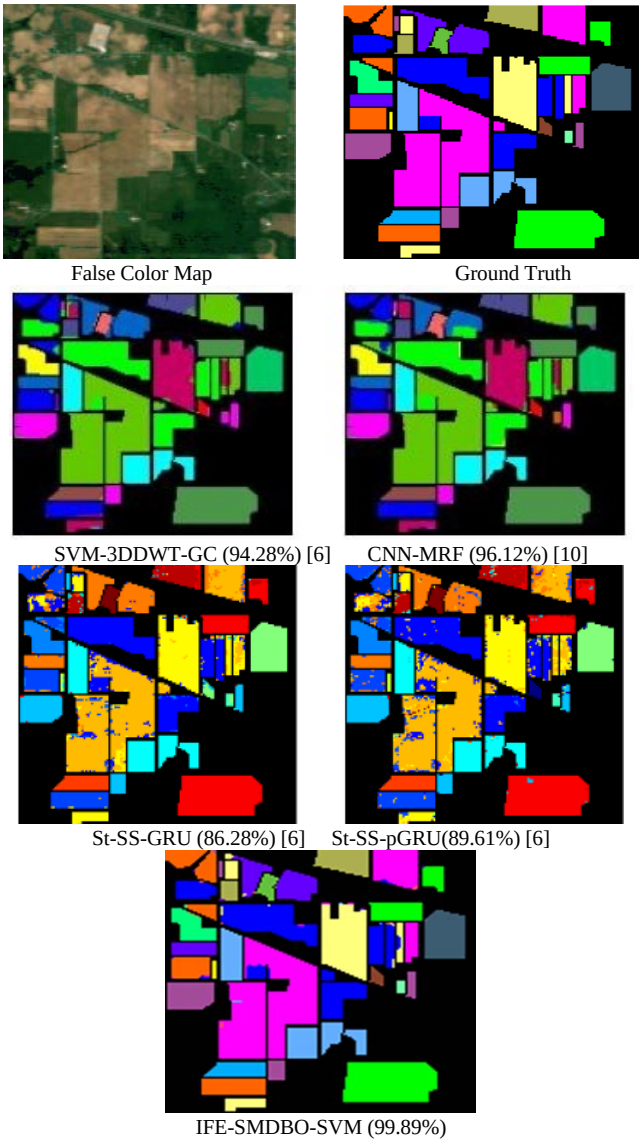


Fig. 3. Classification Maps were Obtained by All Methods on the Indian Pines Dataset (Overall Accuracies are reported in Parentheses).

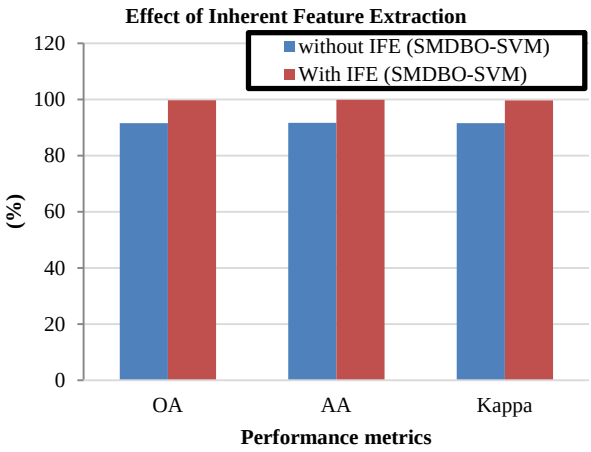


Fig. 4. Effect of IFE on Hyperspectral Object Classification Performance.

V. Conclusion

Designing hyperspectral crop classification with high accuracies with minimal computation time is challenging. In achieving high accuracies existing methodologies used machine and deep learning models; however, induces high training and computation overhead. In addressing computation overhead, dimension reduction technique has been emphasized; however, these model does not attain good accuracies due to poor spatial-spectral feature representation. This paper presented a hybrid design namely IFE-SMDBO-SVM using dimension reduction and machine learning model together for bringing tradeoffs among achieving higher accuracies with minimal time. The IFE-SMDBO-SVM works significantly well even with a fewer number of training samples; further, modeling of soft margin decision boundary aid in addressing feature imbalance issues during classification. The IFE-SMDBO-SVM much better result accuracies, Kappa coefficient, and computation time in comparison with recent HSI classification models such as SS-pGRU, CNN-Al-MNF, CAM, FPGA, SSDGL, and DSVM. The proposed model attains a much superior OA performance of 98.89% which is better than and slightly better than CNN-Al-MNF 2020. However, the accuracies of existing HSI methodologies are highly dependent on a higher number of samples and induce high computation overhead. However, IFE-SMDBO-SVM can work efficiently even with a small number of training samples with high speed. Thus, the proposed IFE-SMDBO-SVM model is much efficient when compared with existing hyperspectral image classification. In the future would consider introducing artificial noise into the hyperspectral image to meet the real-time requirement of the agriculture environment and see how the proposed HSI classification model can perform.

References

[1] J. Yang and J. Qian, "Hyperspectral Image Classification via Multiscale Joint Collaborative Representation With Locally Adaptive Dictionary," in IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 1, pp. 112-116, Jan. 2018, doi: 10.1109/LGRS.2017.2776113.

[2] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[3] A. Santara et al., "BASS Net: Band-Adaptive Spectral-Spatial Feature Learning Neural Network for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 9, pp. 5293-5301, Sept. 2017, doi: 10.1109/TGRS.2017.2705073.

[4] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu and J. Paisley, "Hyperspectral Image Classification With Markov Random Fields and a Convolutional Neural Network," in IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2354-2367, May 2018, doi: 10.1109/TIP.2018.2799324.

[5] Y. Liang, X. Zhao, A. J. X. Guo and F. Zhu, "Hyperspectral Image Classification With Deep Metric Learning and Conditional Random Field," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 6, pp. 1042-1046, June 2020, doi: 10.1109/LGRS.2019.2939356.

[6] Luo, Haowen. (2018). Shorten Spatial-spectral RNN with Parallel-GRU for Hyperspectral Image Classification.

[7] Y. Cai, X. Liu and Z. Cai, "BS-Nets: An End-to-End Framework for Band Selection of Hyperspectral Image," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 3, pp. 1969-1984, March 2020, doi: 10.1109/TGRS.2019.2951433.

[8] C. Yu et al., "Hyperspectral Image Classification Method Based on CNN Architecture Embedding With Hashing Semantic Feature," in IEEE Journal of Selected Topics in Applied Earth Observations and

Remote Sensing, vol. 12, no. 6, pp. 1866-1881, June 2019, doi: 10.1109/JSTARS.2019.2911987.

[9] Lin, L., Chen, C. & Xu, T. Spatial-spectral hyperspectral image classification based on information measurement and CNN. J Wireless Com Network 2020, 59 (2020). https://doi.org/10.1186/s13638-020-01666-9.

[10] X. Cao, J. Yao, Z. Xu and D. Meng, "Hyperspectral Image Classification With Convolutional Neural Network and Active Learning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 7, pp. 4604-4616, July 2020, doi: 10.1109/TGRS.2020.2964627.

[11] Alotaibi, B., Alotaibi, M. A Hybrid Deep ResNet and Inception Model for Hyperspectral Image Classification. PFG 88, 463–476 (2020). https://doi.org/10.1007/s41064-020-00124-x.

[12] Ye, M., Ji, C., Chen, H. et al. Residual deep PCA-based feature extraction for hyperspectral image classification. Neural Comput & Applic 32, 14287–14300 (2020). https://doi.org/10.1007/s00521-019-04503-3.

[13] Gerland, Xin Wang, Guoqiang Wang, "Hyperspectral Band Selection Based on Adaptive Neighborhood Grouping and Local Structure Correlation", Journal of Sensors, vol. 2021, Article ID 5530385, 21 pages, 2021. https://doi.org/10.1155/2021/5530385.

[14] Zhu, Qiqi & Deng, Weihuan & Zheng, Zhuo & Zhong, Yanfei & Guan, Qingfeng & Lin, Weihua & Zhang, Liangpei & Li, Deren. (2021). A Spectral-Spatial-Dependent Global Learning Framework for Insufficient and Imbalanced Hyperspectral Image Classification.

[15] Z. Zhao, D. Hu, H. Wang and X. Yu, "Center Attention Network for Hyperspectral Image Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3415-3425, 2021, doi: 10.1109/JSTARS.2021.3065706.

[16] Ahmad, Muhammad & Shabbir, Sidrah & Raza, Rana Aamir & Mazzara, Manuel & Distefano, Salvatore & Khan, Adil. (2021). Hyperspectral Image Classification: Artifacts of Dimension Reduction on Hybrid CNN.

[17] Okwuashi, Onuwa & Ndehedehe, Christopher. (2020). Deep support vector machine for hyperspectral image classification. Pattern Recognition. 103. 107298. 10.1016/j.patcog.2020.107298.

[18] Kalaiarasi, G., Maheswari, S. Deep proximal support vector machine classifiers for hyperspectral images classification. Neural Comput & Applic (2021). https://doi.org/10.1007/s00521-021-05965-0.

[19] Pathak, D.K. & Kalita, Sanjib & Bhattacharyya, Dhruba K. (2021). Hyperspectral Image Classification using Support Vector Machine: a Spectral Spatial Feature Based Approach. Evolutionary Intelligence. 10.1007/s12065-021-00591-0.

[20] Guo, Y., Yin, X., Zhao, X. et al. Hyperspectral image classification with SVM and guided filter. J Wireless Com Network 2019, 56 (2019). https://doi.org/10.1186/s13638-019-1346-z.

[21] Yang X, Zhang X, Ye Y, Lau RYK, Lu S, Li X, Huang X. Synergistic 2D/3D Convolutional Neural Network for Hyperspectral Image Classification. Remote Sensing. 2020; 12(12):2033. https://doi.org/10.3390/rs12122033.

[22] M.C. Girish Babu, M.C. Padma, “A Efficient Solution for Classification of Crops using Hyper Spectral Satellite Images,” International Journal of Innovative Technology and Exploring Engineering (IJITEE)’, ISSN: 2278–3075 (Online), Volume-9 Issue-2, December 2019, Page No. 5204-5211, 2019.

[23] M. C. Girish Baabu, Padma M. C. Semantic feature extraction method for hyperspectral crop classification. Indonesian Journal of Electrical Engineering and Computer Science, Vol. 23, No. 1, July 2021, pp. 387~395, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v23.i1.pp387-395.

Arabic Sentiment Analysis for Multi-dialect Text using Machine Learning Techniques

Aya H. Hussein[1], Ibrahim F. Moawad[2], Rasha M. Badry[3]
Department of Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum 63514[1, 3]
Department of Artificial Intelligence Engineering, Faculty of Computer Science and Engineering, Galala University, New Galala City, Suez 43511[2]

***Abstract*—Social media networks facilitated the availability and accessibility of a wide range of information and data. It allows the users to share and express their opinions. In addition, it presents the appraisals of the top news and the evaluation of movies, products, and services. This headway has been controlled by a well-known field called Sentiment Analysis (SA). Compared to the research studies conducted in English Sentiment Analysis (ESA), little effort is exerted in Arabic Sentiment Analysis (ASA). The Arabic language is a morphologically rich language that poses significant challenges to Natural Language Processing (NLP) systems. The purpose of the paper is to enrich the Arabic Sentiment Analysis via proposing a sentiment analysis model for analyzing an Arabic multi-dialect text using machine learning algorithms. The proposed model is applied to two datasets: ASTD Egyptian-Dialect tweets and RES Multi-Dialect restaurant reviews. Different evaluation measures were used to evaluate the proposed model to identify the best performing classifiers. The findings of this research revealed that the developed model outperformed the other two research works in terms of accuracy, precision, and recall. In addition, the Bernoulli Naive Bayes (B-NB) classifier achieved the best results with 82% for the ASTD Egyptian-Dialect tweets dataset, while the SVM classifier scored the best accuracy result for the RES Multi-Dialect reviews dataset with 87.7%.**

Keywords*—Arabic sentiment analysis (ASA); arabic tweets; sentiment analysis (SA); natural language processing (NLP); machine learning (ML)

I. Introduction

Most Internet users tend to shift from traditional communication tools (e.g., traditional blogs or mailing lists) to Micro-blogging services. It has a free format of messages and is easy to use. Micro-blogging today has become a top-rated communication tool between Internet users, reflecting users' opinions [1]. These opinions represent any kind of information (political, sport, technology, etc.) that comes from different sources. Sentiment analysis (SA) aims to extract or predict the polarity of users' opinions in a specific area which is a challenging task [2], [3]. SA is considered an important area in Natural Language Processing and Artificial Intelligence to identify emotions and trends about a specific topic. Two main approaches are adopted for SA: machine-learning and Lexicon-based approaches [3]–[5]. The machine learning approach uses a supervised learning approach where a classifier is trained on a human-annotated dataset [6], [7]. Many sentiment analysis researches have been done, especially in the English language. However, there are a huge number of Arabic users on social media posting and sharing their opinions in the Arabic language, expressing feelings and opinions, which can affect many businesses and domains.

Simultaneously, there are many Arabic dialectal variants such as classical Arabic, the language of the Quran, and modern standard Arabic (MSA). The standardized official language is written in the news and taught in schools. In addition, dialectal Arabic (DA) is used in daily life and communications. The Arabic dialects are divided into (1) Egyptian-Dialect Arabic for Egypt and Sudan (EA), (2) Levantine Arabic for Lebanon, Syria, Palestine, and Jordan (LA), (3) Gulf Arabic for Gulf area (GA), and (4) Maghrebi Arabic for Morocco, Algeria, Tunisia, Mauritania, and Libya (MA). Furthermore, Arabic used in social media is usually a mixture of MSA and one or more Arabic dialects [8]–[10].

Arabic sentiment analysis has challenging issues based on two main vectors: Arabic-specific and general linguistic problems. Arabic morphological complexity, limited resources cause the Arabic-specific, and dialects, while the general linguistic issues include polarity fuzziness and strength, implicit sentiment, sarcasm, spam, reviews quality, and domain dependence [9], [11], [12].

The importance of this research is that a sentiment analysis model for analyzing and extracting Arabic text multi-dialect opinions is proposed based on machine learning algorithms and gets high accuracy results. The proposed model experimented using two different datasets (Egyptian and Multi-Dialects datasets). First, the Arabic text is preprocessed to enhance the classifier's performance, such as de-nosing, removing stop words, and applying the lemmatization technique. Then, feature weight and feature selection methods are used. Finally, several machine learning classifiers are applied to extract the text polarity.

The rest of the paper is organized as follows: Section 2 reviews the previous studies and related work, while Section 3 introduces the proposed model. The experimental results are presented in Section 4.

II. Related Literature

Numerous investigations on sentiment analysis approaches have been conducted. The English language has the largest number of research works, while the research efforts exerted for the other languages, including Arabic, are more restricted. This section examines the research work conducted in the field of Arabic Sentiment Analysis (ASA).

Most of the research efforts on ASA studies focused on text processing in a public domain or in news articles, while few efforts were developed in specific domains such as [1], [2], [13]–[18].

Some of the research studies achieved low accuracy results with ML classifiers, as in [1]. On the other hand, some of the research studies used two balanced classes to avoid bias and to achieve better results, such as [15] and [19].

Nabil, Aly, and Atiya [13] used an automatic approach to construct their sentiment dataset in a public domain. They collected 84000 Arabic tweets, and then they determined the most active Egyptian twitters to get the list of the top 30 users. Finally, they filtered the top recent Hashtags to get a list of 2500, and they called it ASTD; it consists of 10,006 Arabic Hashtags classified into machine learning algorithms "SVM, LR, M-NB, B-NB, KNN, SGD, Passive Aggressive, and Linear perceptron" into Subjective "positive 793, negative 1684, neutral 832" and Objective 6691 which has no opinion. Moreover, the objective class doesn't have any effect on sentiment, and its size is too big compared to subjective, positive, and negative classes. The used TF-IDF and CBOW as Text feature and accuracy results showed the best value with B-NB classifier with accuracy 74, 9%.

Abdellaoui, and Zrigui [14] used an automatic approach to construct their sentiment dataset. They collected 5,615,943 Arabic tweets, and then they determined the top 20 most used emojis on Twitter. After that, a list of the ten most used Emojis on Twitter is selected. They dealt with four different dialects, "Egyptian, Levan, Maghrebi, and Gulf," they also used various lexicons to translate dialects to modern standard Arabic MSA. After filtering, they called it TEAD; classifying it with machine learning algorithms "SVM, LR, M-NB, B-NB, DT and RF" into three classes "positive 3,122,615, negative 2,115,325, neutral 378,003 by using TF-IDF and CBOW as text feature and accuracy results showed the best value with SVM classifier with accuracy 84,8%. In this study, they translated dialects to MSA before preprocessing to facilitate the classification process. Further, the number of neutral classes is too small compared to others.

ElSahar and El-Beltagy [20] used an automatic approach to the annotated dataset. They collected four domains as follows "Hotel Reviews, Restaurant Reviews, Movie Reviews, and Product Reviews (PROD)." The dataset was divided into "15K, 8.6K, 1.5K and 15K Arabic reviews for each domain". They dealt with different dialects, "Egyptian, Gulf, and MSA." After filtering, they called each one as (HTL, RES, MOV, and PROD); it classified into two classes "positive, negative," using different machine learning algorithms as "Linear SVM, B- NB, LREG, SGD and KNN" and SVM showed the best accuracy as 82.4%. In this study, they tested the model for each domain separately, so they achieved good results.

Al Mukhaiti, Siddiqui, and Shaalany [1] utilized a new dataset by gathering data from different resources, such as Twitter, Facebook, and Instagram. Thus, overall, 58% of the reviews collected are from YouTube, 37% from Facebook, and 5% from Instagram. They manually annotated the filtered data as negative and positive and segregated them. The best result was 77.7% for accuracy. The study was in the general domain; also, the accuracy results are low despite using two classes.

El-Masri, Altrabsheh, Mansour, and Ramsay [2] utilized a new tool that applies sentiment analysis to Arabic text tweets using a combination of parameters. They tested their work in 8000 tweets with lexicon and machine learning results, and accuracy showed 66.5% with dictionary-based and 34% for SVM.

Oussous, Benjelloun, Lahcen, and Belfkih [15] decided to extract 2000 Moroccan reviews: 1000 positive and 1000 negative, and manually annotated them. They tested their system with machine learning and deep learning techniques. The best experimental results showed 80% with SVM and 95.5% for CNN.

Refaee and Rieser [8] made an Arabic dataset for conclusion investigation, which contains 2000 tweets; categorized into positive, the main half, and negative, the subsequent half. Two techniques were applied to the dataset: corpus-based "Administered Learning" and dictionary-based "Unaided Learning." Four regulated AI calculations were used, i.e., SVM, NB, D-Tree, and K-Nearest Neighbor. The SVM and NB got better outcomes, around 80%. Then again, the vocabulary-based methodology demonstrates that with a huge dictionary, the exactness results were improving. There El-Beltagy, Kalamawy, and Soliman [16] also developed an Arabic sentiment analysis task. The authors were ranked first in the SemEval 2017 task for Arabic SA. They used a set of hand-engineered and lexicon-based features, the classifier of choice was a complement NB classifier, and the accuracy result showed 77%.

Gamal, Alfonse, El-Horbaty, and Salem [17] used a dataset that included more than 151,000 different opinions in variant Arabic dialects, which are labeled into two balanced classes, namely, positive and negative. Various machine learning algorithms are applied to this dataset, including the ridge regression, which gives the highest accuracy of 99.90% with ridge Regression (RR) classifier and 98.95% with SVM. The study showed good results as they used two balanced classes.".

III. Proposed Model

The proposed model aims to extract the people's opinions in Arabic text. The opinions can be classified into three classes: positive, negative, and neutral. The proposed model is based on machine learning algorithms, where six different machine learning algorithms are exploited: Naïve Bayes (NB), Support Vector Machines (SVMs), Decision Tree (DT), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Random Forest (RF) [21], [22].

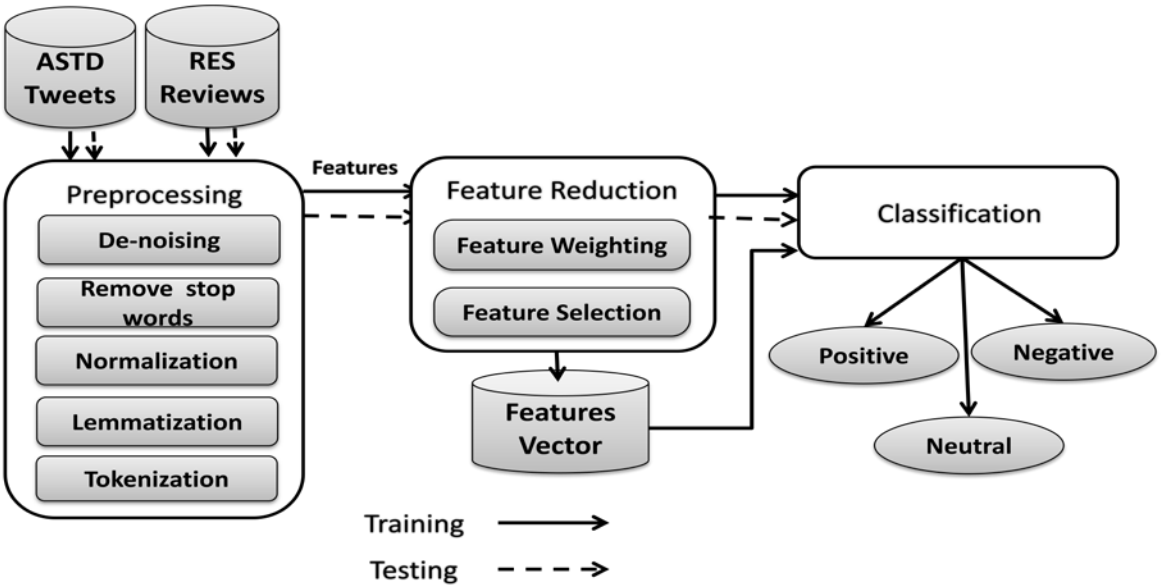


Fig. 1. The Proposed Model.

As shown in Fig. 1, the model consists of three modules, namely the preprocessing, feature reduction, and classification modules. In the preprocessing module, the data should be cleaned and transformed into a format that could be fit into the feature reduction phase. It consists of de-noising, stop-words removal, normalization, lemmatization, and tokenization steps. The feature reduction module is divided into feature weighting and feature selection, which are responsible for scoring each feature and for selecting the most effective features to build the features vector, respectively. Both Sections 3.1 and 3.2 describe the preprocessing and the feature reduction modules in detail, respectively.

Finally, the classification module is applied by using six different machine learning algorithms to classify tweets into three classes (positive, negative, and neutral), as will be explained in detail in Section 3.3. As shown in Fig. 1, the classification process is achieved in two phases: the training and the testing phases, which are represented as solid and dashed arrows, respectively.

A. The Preproccessing Module

The preprocessing process is typically conducted to convert the text into textual features that fit into the SA methods. The preprocessing was applied in a set of sequential steps on two different datasets: Egyptian-Dialect tweets and social media reviews datasets. These steps are tokenization, de-noising, normalization, stop-words removal, and lemmatization.

The preprocessing algorithm is illustrated in Algorithm (1), where the tokenization is the task of chopping a sequence of characters in a document up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation, space, and punctuation marks [19], [23]. Then, it applies the de-noising step that simply involves using a neural sequence transaction model to back translate the noised text to the original clean text [24]. The de-noising text includes removing URLs, which began by (http ://) until the following space, hash-labels subjects, mentions, punctuations, and special characters. In addition, the symbols for emotions are removed [25].

Algorithm 1 Pre-processing

```
//Tokenization
Input:          tweets_table
Output:         token_table
While (more tweets exit in tweet table) do
        { T=next tweet; P=1;
        While not end of T
                { I= next_space (T);    // determine the position of next
space in T
                  Token = tirm_substring (T,P,I,string);
                   Insert into token_table (tweet number,     token);
                  T=rest_string;}
//De-noising and normalization
Input:          tokens_table

Output:         Updated tokens_table
Arabic_punctuations = {'"', `, ÷, ×, <,>, _, (,), *, &, ^,], [, ,+ ,¦ ,~ ,{,} ,‘ ,“ ,/ ,-
,؛,؛,؛,‘,‘,-,–,“,…,”,!,|}
Emotions = {⌨, ↪, •,          ,.,.,.#...}
S=a set of Arabic vowels
While (more token exist in tokens_table) do
{T=next token;
F= the first letter of T;
While (F in Emotions or Arabic_punctuations or Emotions or S) do // remove
symbols from the beginnig of token
                                {trimleft (T.F.T):
                                 F= the first letter in T:}
                        E= the last letter of T;
While (E in symbols) do //remove symbols from the end of token
                                {trimright (T.F.T):
                                 E= the last letter of T;}
                               Concatenation (T, E, T);
 If (T starts with # or @ or the length of T<=2) then
                        Delete T from tokens_table}
//Stop word removing
While (more token exist in tokens_table) do
{T=next token;
If (T in stop_words_table then
 Delete the entry of T from token_table}
//Lematization
While (more token exist in tokens_table) do
{T=next token;
          If (T=3) then
                        Lemma=T
                        Root=T
                If (T>t) then
                        Remove suffixes from T
                        Lemma=T}

                }
```

While the normalization is responsible for transforming text into a single canonical form [19], for example:

- Removing Arabic short vowels (diacritics) [26] such as " ً، ٍ، َ".
- Removing the duplicate characters that doesn't influence the significance of the word such as "جمييييييييل" is converted to "جميل" "Beautiful" [26].
- Replacing the letter "ة" to the letter "ه ", replacing the letters "أ ، إ ، آ" to the letter "ا".
- Normalizing the words that can be written in different forms to one form. For example, the "emergency" word can be written in Arabic language as "الطوارى" or "الطوارئ". It should be mapped into "الطوارئ".
- Normalizing compound words by going along with them by the character "_ "; for example, "في عالم موازي", which will be standardized to "فى_عالم_موازى" ""in_aparallel_world".
- Furthermore, in removing stop words, there are some words which don't have any effect on sentiment analysis, such as "من ، في ، عن""from, in, about". These stop words are removed using a pre-defined list of stop words [15]. Here an example for removing stop words form a tweet " موقعه مناسب ممكن الوصول اليه بسهولة عن طريق التاكسي" "Convenient location, easily accessible by a taxi" comes to " موقعه مناسب ممكن الوصول بسهولة طريق التاكسي", which minimized data size.

The last step is lemmatization, which converts inflected words to their root form; It helps to minimize the dataset size, which improved the accuracy and performance. For example, "معاك لو تعيدها الف مره" "With you if you repeat it a thousand of times" begins to "مع لو أعاد ألف مر".

After the tweets/reviews data are preprocessed, a list of features is generated and can be represented as ([f1], [f2] … [fn]).

B. The Feature Reduction Module

The Feature reduction module aims to reduce the extracted features. It consists of two processes, which are feature weighting and feature selection. In feature weighting, the extracted features are weighted using the Term Frequency-Inverse Document Frequency (TF-IDF).

Finally, feature selection is applied to select the most candidate features. Then, the selected features will be used as an input for the classification modules.

1) Feature selection: Term frequency-inverse document frequency (TF–IDF) [27] was applied to find strongly affected words for sentiment by evaluating the weight of the feature, where each term 't' is weighted by a weight 'w' using TF-IDF [28]. The Term frequency computes the number of instances of "t" term that appears in the tweets/reviews document, as shown in Eq. (1).

$$Tf(t,d) = (fd(t))/(max[fd(t)]) \quad (1)$$

Where fd (t) is the frequency of the term "t" in tweet/review "d".

The Inverse Document Frequency (IDF) is used to estimate the importance of a given term. It measures how rare a given term in the whole tweet/ review document, using Eq (2):

$$IDF(t) = log(N/dft) \quad (2)$$

Where dft is the number of tweets /reviews with the term "t", and "N" is the total number of tweets / reviews per document.

Finally, the TF-IDF is calculated as the result of the multiplication of TF and IDF as shown in Eq(3):

$$TF\text{-}IDF = tf(t,d)\ IDF(t) \quad (3)$$

2) Feature weighting: After feature weighting, the feature selection process is applied to select the most important and relevant features. The used techniques are classified into two categories are supervised and unsupervised techniques. The supervised technique is used for labeled data to identify the relevant features for increasing the efficiency of supervised models like classification and regression. In unsupervised technique, it can be used for unlabeled data.

In addition, threshold-based approach is applied to identify the least number of features as shown in Table I.

TABLE I. TWEETS-BASED EXPERIMENT ACCURACY RESULTS USING DIFFERENT NUMBER OF FEATURES

NUMBER OF FEATURES		SGD	DT	M-NB	B-NB	SVM	LR	RF
2000	ACCURACY	68	50	60	68	45	62	62
5000		71	57	65	75	50	70	64
10000		78	74	78	82	78	78	77.8
15000		78	74	78	82	78	78	77.8

Cross-validation is used to tune the soft margin parameter C ($\lambda = 1/2C$). Higher values of C add a higher penalty for the misclassified points rather than maximizing the separation margin. Therefore, the optimization problem will lead to a larger number of selected features to reduce the misclassified errors then the threshold value is determined for each experiment to get the results for accuracy.

C. The Classification Module

The most important advantage of ML algorithms is that they deal with a complex problem and gets closer results than humans. In Arabic text processing, complex problems exist and need efficient solutions to solve these problems [6], [29], [30].

In the proposed model, six supervised machine-learning (ML) classifiers were exploited: NB, LR, SVM, DT, SGD, and RF. The used NB algorithms are Multinomial Naive Bayes (M-NB) and Bernoulli Naive Bayes (B-NB). These classifiers have been chosen since they are the most commonly used in the literature of SA [7], [31], [32]. These algorithms are applied to the training set to build the classification model.

SVM is one of the most robust prediction methods based on statistical learning frameworks [33]–[35]. While the NB assumption of attribute independence works well for text categorization at the word feature level [35], [36]. On the other hand, DT is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [37], [38].

Furthermore, the SGD is an iterative method for optimizing an objective function with suitable smoothness properties. RF is an ensemble learning method for classification, regression, andother tasks. It is operated by constructing many decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees [37]–[39].

IV. Experimental

This section presents the experiments conducted in this research and a discussion of the results. The Two experiments have been conducted using two different data sets: Arabic Sentiment Tweets Dataset (ASTD) [13] and Restaurant Reviews dataset (RES) [20].

A. Tweets-based Experiment

In this experiment, the ASTD [13] dataset is used, which contains 10,006 Egyptian-Dialect tweets that are divided as follows: 793 positive sentiment, 1684 negative sentiment, 832 neutral sentiments, and 6691 objective tweets. An example for a positive labeled tweet statement is "محبين البرنامج بيزيدوا", which is equivalent in English to "fans of El-Bernameg are increasing". Further, ASTD data were used in different ASA research because it is completely available. In the first experiment, as the objective class isnot effective in SA, only three classes were considered: positive, negative, and neutral with a total data size of 3,316 Egyptian-Dialect tweets.

Six different classifiers were applied in the tweets-based experiment: DT, SVM, RF, LR, M-NB, B-NB, and SGD. It has been applied to the ASTD Egyptian-Dialect tweets dataset with all the preprocessed steps described above, and hence a feature vector, which consists of 10000 top selected features has been created.

Table II illustrates the evaluation measures of the proposed model using the precision, recall, and accuracy measures. Since tweets in our model are divided into three classes, we have three precisions and recalls value for each class to be calculated [40], [41]by the following Eq. (4):

$$Precision_i Or\ Recall_i = \frac{Tweets\ correctly\ assigned\ to\ class_i}{Tweets\ attributed\ to\ class_i} \quad (4)$$

The B-NB scored the best accuracy with 82 %, followed by SVM, LR, M-NB, and SGD with 78% accuracy, as shown in Fig. 2.

Further, the tweets-based experiment results are compared with the related works that used the ASTD Egyptian-Dialect tweets dataset, as shown in Table III. The results show the different machine learning algorithms used by the proposed model and the related work with their evaluation measures.

TABLE II. Tweets-based Experiment Classifiers Results using ASTD Egyptian-Dialect Tweets Dataset

	SGD	DT	M-NB	B-NB	SVM	LR	RF
ACCURACY	78	74	78	82	78	78	77.8
PRECISION	77	72	80	82	78	77	70
RECALL	78	74	79	82	78	78	86

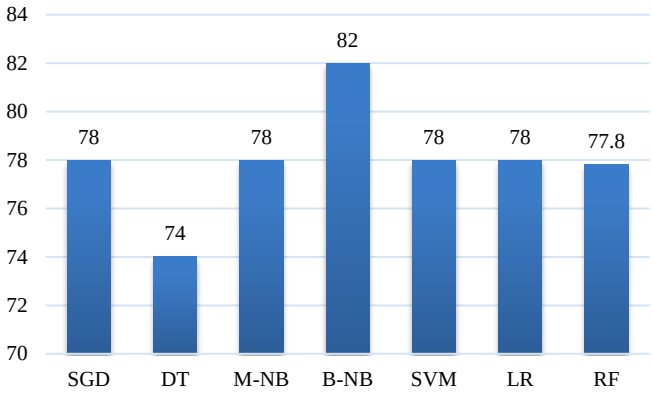


Fig. 2. Proposed Model Accuracy Results for Tweets-based Experiment.

TABLE III. Tweets-based Evaluation Results: the Proposed Model Versus the Related Work using ASTD Egyptian-Dialect Tweets

		SGD	DT	M-NB	B-NB	SVM	LR	RF
Abdellaoui&Zrigui[14]	accuracy	--	68.7	74.4	74.9	75.5	74.9	68.7
	precision	--	78	72	81	75	76	84
	Recall	--	73	72	74	76	74	73
Kaseb and Ahmed [42]	accuracy	--	--	--	--	64	--	--
	precision	--	--	--	--	58.3	--	--
	Recall	--	--	--	--	63.9	--	--
Nabil and Atiya[13]	accuracy	67.1	--	67	66.9	68.9	67.6	--
The proposed model	accuracy	78	74	78	82	78	78	77.8
	precision	77	72	80	82	78	77	70
	Recall	78	74	79	82	78	78	86

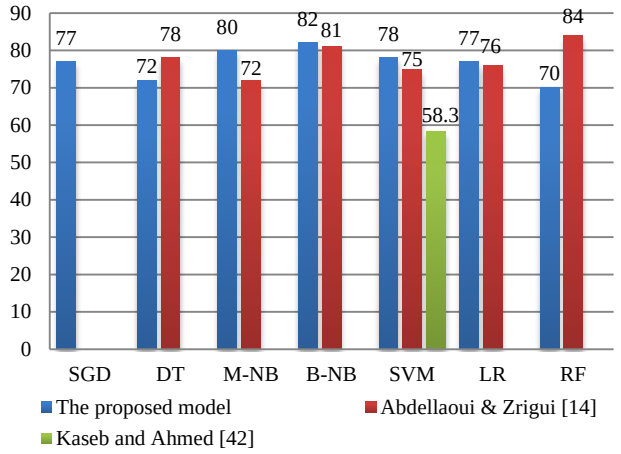


Fig. 3. Precision Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

It is noticed that the proposed model has better values with all classifiers compared with Nabil and Atiya [13] who classified the ASTD Egyptian-Dialect tweets dataset into four classes (positive, negative, neutral, and objective); while the proposed model classified the ASTD Egyptian-Dialect tweets dataset into three classes (positive, negative, and neutral), which are the most popular classes. Further, Nabil and Atiya didn't apply DT and RF classifiers; on the other hand, they are applied by the proposed model.

Both Fig. 3 and Fig. 4 show the precision and the recall for the proposed classifiers model versus two related works.

In Abdellaoui and Zrigui [14] research, SGD is not applied to the ASTD Egyptian-Dialect Tweets dataset, but the proposed model applies it. Moreover, the proposed model achieved better values with all classifiers compared with the above study.

On the other hand, Kaseb, and Ahmed [42] filtered and cleaned ASTD Egyptian-Dialect tweets to 1652 records. Unfortunately, they applied only one SVM classifier and achieved a lower accuracy of 64% compared with the proposed model with 78%.

As shown in Fig. 5, the proposed model with B-NB, SVM, LR, M-NB, and SGD achieved a better accuracy compared to the related work. B-NB achieved 82% versus 66.9% and 74.9% for the related works [13]&[14]. While SVM achieved 78 % versus 68.9%, 75.5%, and 64 % for the related works. Further, LR has higher accuracy with 78% compared to related works with 67.6% and 74.9%. M-NB achieved 78% with higher accuracy than the related works.

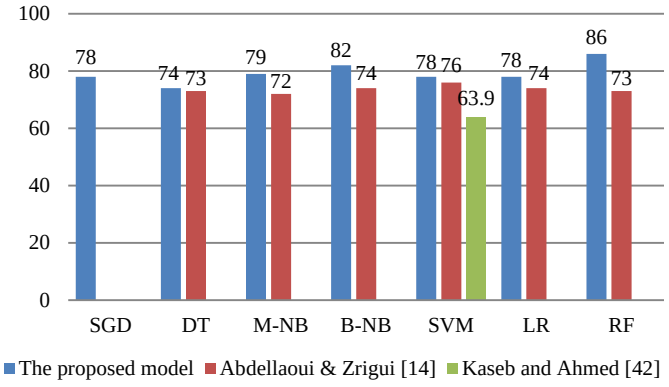


Fig. 4. Recall Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

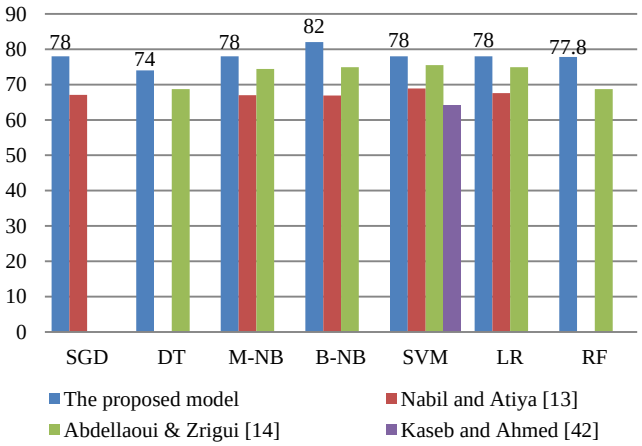


Fig. 5. Accuracy Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

In DT classifier, our module has higher accuracy rating than [14], but without knowing its recall, we cannot comfortably trust the results. Interestingly, recall and accuracy are often at odds with each other, as attempts to boost recall often negatively impact accuracy and vice versa.

B. Review-based Experiment

The Restaurant Review dataset RES [20] is used in the review-based experiment, which was collected from the trip advisor site with a total number of 10,871 reviews. The RES is divided as follows 8021 positive sentiments, 2625 negative sentiment, and 225 neutral reviews. An example of a positive sentiment tweet is "مطعم ممتاز و خدمة حلوى أوى و مكان متميز و معاملة راقية" which is equivalent in English to "Excellent restaurant, great service, excellent place and classy treatment" [20].

TABLE IV. COMPARING REVIEWS-BASED EVALUATION RESULTS AND THE RELATED WORK USING RES DATASET

		SGD	**DT**	**M-NB**	**B-NB**	**SVM**	**LR**	**RF**	**KNN**
ElSahar and El-Beltagy[20]	accuracy	78.4	--	--	82.1	81.4	70.4	--	49.5
The proposed model	accuracy	85.6	76.9	82	79.8	87.2	85.9	83.8	--
	precision	84	74	83	80	85	82	84	--
	Recall	86	77	82	80	87	84	86	--

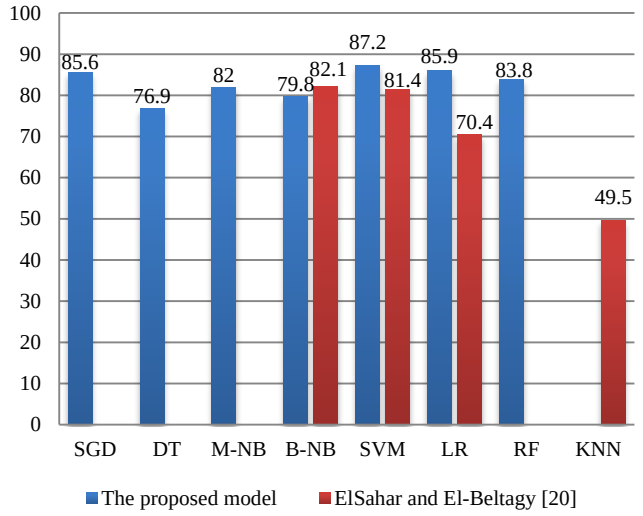


Fig. 6. Accuracy Value Results for Proposed Model Compared with Related Works for the Review-based Experiment.

The review-based experiment is applied using different classifiers DT, SVM, RF, LR, M-NB, B- NB, and SGD with a feature vector that consists of 15000 top selected features. The review-based experiment evaluation results are shown in Table III. It has been revealed that SVM scored the best accuracy with 87.2%, followed by LR with 85.9%, and SGD with 85.6% accuracy. On the other hand, the tweets-based experiment results are compared with the related works that use the RES dataset, as shown in Table IV. The table presents the different machine learning algorithms used by the proposed model and the related work with their evaluation measures. It is noticed that the proposed model has better values with most classifiers compared with ElSahar and El-Beltagy[20]. Also, ElSahar and El-Beltagy did not apply DT and M-NB classifiers, while the proposed model applies them.

Moreover, Fig. 6 Shows the accuracy of the proposed classifier model versus the previous works. The proposed model with SVM, LR, and SGD achieves better accuracy than the previous works. SVM achieved 87.2% versus 81.4% for the previous work while LR achieved 85.9 % versus 70.4% for the related work.

V. Conclusion

The paper has introduced a new model for Arabic sentiment analysis and the effect of different text preprocessing techniques on classification accuracy. The proposed model was evaluated using recall, precision, and accuracy measures. Two different types of Arabic datasets are used: (1) ASTD is an Egyptian-Dialect tweets, and (2) RES, which is Multi-Dialect reviews. Two main experiments have been conducted using machine learning algorithms (DT, SVM, RF, LR, M-NB, and B-NB). The first experiment was applied to the ASTD dataset with 3,316 Egyptian-Dialect tweets. It is noticed that B-NB scored the best accuracy with 82%, followed by SVM, LR, M-NB, and SGD with 78% accuracy. The second experiment was applied to the RES dataset with 10K Multi-Dialect Arabic reviews. In addition, SVM achieved accuracy with 87.2%, followed by LR with 85.9%, and SGD with 85.6%. These results revealed that the proposed model outperformed the related works in the two conducted experiments.

Further, the experiments showed that de-noising, stop words removal, lemmatization, and normalization slightly improved the classification's performance. The proposed model will use different techniques in future work, such as deep learning or lexicon-based approaches.

References

[1] "Dataset Built for Arabic Sentiment Analysis | SpringerLink." https://link.springer.com/chapter/10.1007/978-3-319-64861-3_38 (accessed Oct. 29, 2021).

[2] M. El-Masri, N. Altrabsheh, H. Mansour, and A. Ramsay, "A web-based tool for Arabic sentiment analysis," ProcediaComput. Sci., vol. 117, pp. 38–45, 2017.

[3] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data, Springer, 2012, pp. 415–463.

[4] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in 2016 3rd international conference on computing for sustainable global development (INDIACom), 2016, pp. 452–455.

[5] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowl.-Based Syst., vol. 89, pp. 14–46, 2015.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," ArXivPrepr. Cs0205070, 2002.

[7] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 1–5.

[8] E. Refaee and V. Rieser, "An arabic twitter corpus for subjectivity and sentiment analysis.," in LREC, 2014, pp. 2268–2273.

[9] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in 2017 1st international workshop on arabic script analysis and recognition (asar), 2017, pp. 114–118.

[10] E. Benmamoun, The feature structure of functional categories: A comparative study of Arabic dialects. Oxford University Press, 2000.

[11] I. Moawad, W. Alromima, and R. Elgohary, "Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval.," Arab. J. Sci. Eng. Springer Sci. Bus. Media BV, vol. 43, no. 12, 2018.

[12] A. Hamdi, K. Shaban, and A. Zainal, "A review on challenging issues in arabic sentiment analysis," 2016.

[13] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2515–2519.

[14] H. Abdellaoui and M. Zrigui, "Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis," Comput. Sist., vol. 22, no. 3, pp. 777–786, 2018.

[15] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," J. Inf. Sci., vol. 46, no. 4, pp. 544–559, 2020.

[16] [16] S. R. El-Beltagy, M. E. Kalamawy, and A. B. Soliman, "Niletmrg at semeval-2017 task 4: Arabic sentiment analysis," ArXivPrepr. ArXiv171008458, 2017.

[17] D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Twitter benchmark dataset for Arabic sentiment analysis," Int J Mod EducComputSci, vol. 11, no. 1, p. 33, 2019.

[18] L. Khreisat, "A machine learning approach for Arabic text classification using N-gram frequency statistics," J. Informetr., vol. 3, no. 1, pp. 72–77, 2009.

[19] S. Bedrick, R. Beckley, B. Roark, and R. Sproat, "Robust kaomoji detection in Twitter," in Proceedings of the Second Workshop on Language in Social Media, Montréal, Canada, Jun. 2012, pp. 56–64.

Accessed: Oct. 29, 2021. [Online]. Available: https://aclanthology.org/W12-2107.

[20] H. ElSahar and S. R. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis," in International Conference on Intelligent Text Processing and Computational Linguistics, 2015, pp. 23–34.

[21] G. Bonaccorso, Machine learning algorithms. Packt Publishing Ltd, 2017.

[22] T. O. Ayodele, "Types of machine learning algorithms," New Adv. Mach. Learn., vol. 3, pp. 19–48, 2010.

[23] R. M. Badry and I. F. Moawad, "A semantic text summarization model for Arabic topic-oriented," in International Conference on Advanced Machine Learning Technologies and Applications, 2019, pp. 518–528.

[24] Z. Xie, G. Genthial, S. Xie, A. Ng, and D. Jurafsky, "Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, Jun. 2018, pp. 619–628. doi: 10.18653/v1/N18-1057.

[25] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in LREc, 2010, vol. 10, no. 2010, pp. 1320–1326.

[26] "Key issues in conducting sentiment analysis on Arabic social media text | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/abstract/document/6544396/ (accessed Oct. 29, 2021).

[27] J. Ramos, "Using tf-idf to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, 2003, vol. 242, no. 1, pp. 29–48.

[28] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, "Efficient email classification approach based on semantic methods," Ain Shams Eng. J., vol. 9, no. 4, pp. 3259–3269, 2018.

[29] "A multi-objective feature selection method based on bacterial foraging optimization | SpringerLink." https://link.springer.com/article/10.1007/s11047-019-09754-6 (accessed Oct. 29, 2021).

[30] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," ACM Comput. Surv. CSUR, vol. 51, no. 6, pp. 1–36, 2019.

[31] M. Alruily, "Classification of Arabic Tweets: A Review," Electronics, vol. 10, no. 10, p. 1143, 2021.

[32] M.-Y. Cheng, D. Kusoemo, and R. A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning," Autom. Constr., vol. 118, p. 103265, 2020.

[33] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, 2013.

[34] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," Expert Syst. Appl., vol. 162, p. 113746, 2020.

[35] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," in 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016, pp. 106–111.

[36] M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm," Int. J. Adv. Comput. Res., vol. 3, no. 4, p. 139, 2013.

[37] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310–1315.

[38] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," Adv. Neural Inf. Process. Syst., vol. 25, 2012.

[39] L. Bottou and O. Bousquet, "The tradeoffs of large-scale learning: Optimization for Machine Learning, 351," 2011.

[40] K. A. Djaballah, K. Boukhalfa, and O. Boussaid, "Sentiment analysis of Twitter messages using Word2vec by weighted average," in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 223–228.

[41] K. Boukhalfa and O. Boussaid, "Sentiment analysis of twitter messages using Word2vec by weighted average".

[42] M. Farouk and G. Kaseb, "Extended-ATSD: Arabic Tweets Sentiment Dataset," vol. 14, pp. 4780–4785, May 2019.

A Systematic Review on e-Wastage Frameworks

Sultan Ahmad*[1], Sudan Jha[2], Abubaker E.M. Eljialy[3], Shakir Khan[4]
Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University Alkharj, 11942, Saudi Arabia[1]
School of Sciences, Christ (Deemed to be University), NCR, New Delhi, India[2]
Department of Information System, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University Alkharj, 11942, Saudi Arabia[3]
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia[4]

***Abstract*—The electronic devices that are targeted to the end users have become day to day essential parts. Traditional methodologies have changed drastically resulting in efficient mode of communication and fast information retrieval. As the demand and the production are exponentially growing, patterns of sales, storage and their destruction and then again, their collection have also been changed. This paper analyses many such behaviors of (electronic) waste management and recommends solutions like recycling management, different directives and policies required to be followed. Authors have emphasized on providing substantial information that can be useful to the regulating authorities responsible for waste management or the manufacturers of various electronic products and then the policy makers. With an extensive review of electronic wastages, authors have emphasized three variables (sales, stock and lifespan) for replacing/upgrading the older products with advanced versions. The root causes of electronic wastages are found in industrializing countries like India, China, Vietnam, Pakistan, the Philippines, Ghana and Nigeria whereas industrialized countries also play equally important role for its generation. This paper signifies the importance of e-waste management practice to reduce the emerging electronic waste hazards. Authors focus on today's demand of electronic devices, importance of e-waste management and management practices. The paper recommends key findings based on surveying data regarding the lack of regulation to manage the e-waste. The review concludes that the lack of regulation and improper awareness are the basic factors responsible for e-wastage and requires major focus to manage the e-waste.**

Keywords*—*e-Wastage; e-Wastage management; barriers; policy; findings; e-Wastage regulations; industrializing countries; industrialized countries

I. Introduction

E-wastage or E-waste is a phrase, widely used to cover items of all types of electrical and electronic equipment (EEE) and its parts that have been discarded by the owner as waste without the intention of reuse. In addition, used electronics which are meant for refurbishment, reuse, recycling through material recovery, or disposal are also considered e-waste. E-waste includes almost any household or business item containing circuitry or electrical components with either power or battery supply. To name a few, Microwaves, Fans, Refrigerators, washing machines, dryers, home entertainment devices, computers, laptops, tablets, mobile phones, Medical equipments like Dialysis Machines, Imaging Equipment, Autoclave, Defibrillator, Office and Information technology equipments like Copiers/Printers, IT Servers, Cords and Cables, WiFi dongles, switches, hubs, routers, Uninterrupted Power Supplies (UPS Systems) and other electronic utilities like Heating Pads, Remote Controls, Electrical Cords, Smart Lights, Treadmills, Smart Watches and many more. [www.ewaste1.com/what-is-e-waste/]

Few decades from now, the major proportion of e-waste was generated from the equipments that were discarded as they were no longer workable but it has been found in recent years that the current advancements in innovations in electronics and electrical technology has resulted into increasing demand for upgraded/faster/new electronic products for replacing/upgrading the older products with advanced versions. Although the exponential growth in electronics and electrical industry has benefited consumers to a great extent with massive technological advancements by making routine tasks easier and time-efficient but has also highly contributed in 'declaring' many workable electronic devices to be considered as obsolete. Thus, contributing to a great extent in generation of global electronic wastage. For instance, Video Cassette Recorder (VCR) players got replaced with Digital Versatile Disc (DVD) players and recently DVDs are being replaced by Blu-ray players. Although DVD players are working fine for normal videos but cannot showcase High-Definition (HD) videos with high quality like Blu-ray players. So, DVDs are now becoming a part of e-waste.

As there has been a sky-rocketing surge in the amount of e-waste due to dumping of electronic products by owners just because new models arrive every day especially for computers, laptops, smart phones and other electronic devices, the e-waste management has become a daunting task for both industrialized and industrializing countries. The Global E-waste Monitor 2020 report has presented statistics related to e-waste generation from 2014 onwards. In 2014, 44.4 MT of e-waste was reported to be generated globally which has risen to 53.6 MT in 2019. The report has predicted e-waste generation to gain manifold increase by the year 2030 and estimated a shocking figure of 74.7 MT.

A. Sources of e-Wastage

Generally, the sources of e-waste comprises of almost all types of electronics and electrical equipments. But due to advancements in technology, new electronic and electrical devices are being introduced that are added up accordingly to the list of sources. So, from time to time authors have given

*Corresponding Author

different categories of devices to be considered for as sources for e-waste.

B. e-Waste Management: A Global Challenge

E-waste management refers to the disposal of electrical and electronic devices in a secured and environment friendly way. Unlike, municipal waste, e-waste requires more sophisticated techniques to be followed for management and disposal due to crucial nature of components being used in electronic and electrical devices.

Most of the electrical and electronic devices especially IT hardware contain toxic, non-biodegradable and hazardous materials including mercury, lead, cadmium, beryllium, chromium, and chemical flame retardants, or or polychlorinated biphenyls (PCBs) which have the potential to leach into soil and water as well as could be hazardous if exposed to air.

C. Impact of E-Waste

The outrageous increase in the amount of e-waste being generated every year across the world has put forth many challenges for industrialized and industrializing countries and has created intense pressure for the execution of sustainable practices to redesign and recycle the products.

In developed countries, several conventions, directives and laws to regulate the e-waste disposal have been formulated for efficient e-waste management which includes Basel Convention, StEP initiative, 3Rs(Reduce, Reuse and Recycle) etc. Product manufacturing industries are taking back items which are collected by retailers and local governments for safe destruction, redesign, recycle and recovery of materials. Undoubtedly, industrialized countries have gone to great lengths to formulate efficient high-cost systems to handle e-waste which includes elaborated collection systems, deployment of clean recovery technologies, carefully engineered disassembly stations, plasma furnaces to prevent release of dioxins. But, despite following various conventions, directives to protect environment from the hazardous implications of e-waste, the majority of e-waste across Europe and North America still remains unrecycled (Barba-Gutierrez et al., 2008). Therefore, to ensure proper disposal of e-waste, developed countries are shipping e-waste to developing countries like China, India, Pakistan, and Nigeria due to availability of cheap labour for recycling.

II. Literature Survey and Background Study

In order to develop a conceptual framework, a systematic review of e-waste management research reported by various authors has been presented in this section. An extensive review is conducted comprising of research papers published between 2005 to 2021. A comparative analysis has been presented in the tabular form.

Both valuable metals and extraordinary materials, like Pb, Hg, As, Cd, Se, Cr, etc. are major factors in creating electronic wastages [1]. These non-ferrous and uncommon metals can be recuperated for additionally reuse and reuse, while the strengthened sap and epoxy sap (with generally bring down financial esteems) can be recouped for their warmth esteems through burning or reused as coatings, clearing and building materials subsequent to being powdered[2,3]. The yearly total entirety of overall EW age ranges from 20 to 50 million tons in created nations [4]. On the off chance that we allude to paper referenced at [5], the total residential electronic and electric waste is required to ascend to around 400 million units by 2015.

In 2010 Yu et.al [6] studied the existing framework for EW management in China counting regulatory policies and pilot projects. In 2012, Chibunna et. al [7], identified and talked about EW management challenges among establishments through a case study at UniversitiKebangsaan Malaysia (UKM).In 2012 Sthiannopkao and Wong [8] provided deeper knowledge about the different directives and policies needed to be followed while handling with EW. In 2013 Wang et. al [9], tried to address how to enhance EW estimates by giving techniques to increase information quality and proposed an advanced IOA method including each of the three variables (sales, stock and lifespan) and best available information focuses to prepare better datasets for modeling. In 2015 Reddy [10], concentrated upon informal EW recyclers who subsidize the environmental costs of Bangalore's IT blast. In 2016 Tansel [11], reviewed the challenges associated with increasing EWquantities. In 2016 Awasthi et. al. [12], explored the environmental contamination from EW recycling at numerous little formal and informal workshops in India. In 2017 Resmi and Fasila [13], proposed a novel calculation for establishing a standard methodology to manage and refurbish EW called EW Management and Refurbishment Prediction (EMARP), which can be adapted by refurbishing industries in order to improve their performance. Here Table I indicated the comparative analysis with advantage and disadvantages of different authors.

Both valuable metals and extraordinary materials, like Pb, Hg, As, Cd, Se, Cr etc. are major factors in creating electronic wastages [14]. These non-ferrous and uncommon metals can be recuperated for additionally reuse and reuse, while the strengthened sap and epoxy sap (with generally bring down financial esteems) can be recouped for their warmth esteems through burning or reused as coatings, clearing and building materials subsequent to being powdered. The yearly total entirety of overall EW age ranges from 20 to 50 million tons in created nations [15]. On the off chance that the total residential electronic and electric waste is required to ascend to around 400 million units by 2015.

TABLE I. COMPARATIVE ANALYSIS

SN	Year	Authors	Advantages	Limitations
1	2021	Tetiana Shevchenko et. al [16]	Proposed a smart reverse system for e-waste management based on intelligent information technology (IT) tools with an aim to minimize collection costs and ultimately, carbon dioxide emissions.	- The proposed smart e-waste reverse system has not been tested in real life scenario. - Lacks real life cost – benefit analysis.
2	2020	Yigit Kazancoglu et.al [17]	Proposed a collection and classification framework for efficient electronic wastage management by incorporating data-driven technologies that would certainly ensure social, environmental and economical sustainability in industrializing and less industrialized nations.	- The initial cost of investment is high. - Absence of adequate knowledge related to data-driven technologies for formulating sustainable solutions.
3	2020	Piotr Nowakowski and Teresa Pamuła [18]	Deep learning techniques have been used for categorizing and measuring the size of e-waste that would aid e-waste collection companies to prepare an effective collection plan.	- The research is confined to Large home appliances. - Small appliances and other categories of e-waste are not considered.
4	2020	Sudan Jha et. al [19]	Almost all aspects of smart cities have been implemented with IT enabled services.	- Despite inclusion of smart technologies in smart cities still they lack in efficient e-waste management.
5	2019	Charu Gangwara et.al [20]	The significant spread of cardiovascular morbidity, namely hypertension, among local residents and on site workers engaged in e-waste processing is explored. The significant correlations between the inhabitants' heavy metal concentration in the blood and corresponding metal concentrations in atmosphere are identified.	- No specific guidelines and safeguard plans have been suggested.
6	2018	Abhishek Kumar et.al [21]	Detailed analysis of e-waste management status, legislation formulated, and technology being used for e-waste recycling in India.	- No specific guidelines and safeguard plans have been suggested.
7	2018	Ashwani Kumar and Gaurav Dixit[22]	Identified '10 barriers' that restrict efficient e-waste management and used interpretive structural modeling (ISM) and Decision Making Trail and Evaluation Laboratory (DEMATEL) for understanding the hierarchal and contextual relationship among the barriers of e-waste management to aid in policy and decision making.	- The proposed model is highly susceptible to judgment and expertise of experts involved for framing the barriers. - Even more barriers could be explored and validated according using different statistical techniques.
8	2017	Stefan Salhofer [23]	Discussed different challenges being faced in e-waste management regarding collection and treatment of e-waste with an overview of technologies applied for the removal of hazardous materials for the recovery of valuable materials. The study focuses on three global areas, namely; Europe, China and Vietnam.	- Deals with hypothetical proposal which are yet to be implemented.
9	2016	Borthakur and Govind[24]	Studied on consumers' EW transfer behavior and awareness with respect to the case of India. Provides a detailed view on the current worldwide EW scenarios.	- Research is primary - Adequate detail.
10	2017	Rahman[25]	Investigated into the responsibilities of peoples played in the EW management in Bangladesh and also examined the existing policy gap and environmental management issues in terms of EW.	- Many issues have been highlighted which is not enough but need to provide solution for the issues.
11	2016	Debnath et. al. [26]	Tried to establish EW management as a parameter for green computing.	Various questions have been raised but did not provide any solution.
12	2013	Dwivedy and Mittal[27]	Attempted to understand the critical elements related with EW	Recycling program in the context of India
13	2013	Pariatamby and Victor[28]	Examined the strategically trends of EW management in Asia	Various questions have been raised but did not provide any solution.
14	2015	Jaiswal et. al. [29]	Individual awareness is addressedregarding EW handling.	Proposes a green framework for EW.

15	2015	Kumar and Rawat[30]	Analytical presentation for managing information for efficient management of EW.	It gives theoretical reasoning only. No realization on the theries proposed.
16	2015	Davis and Garb[31]	Endeavored to give broad field employments and optional written works to offer a logical order of administration positions towards casual EW hones.	The proposed method tackles a noteworthy however not overwhelming portion of the puzzle No other critical issues were addressed
17	2015	Dwivedy et. al. [32]	- Discusses on the "takeback" policies - The paper was based on reasonableness for the Indian conditions.	- Focused on copper links in the EW stream - Discusses upstream and downstream changes - limits and business components of a formalizing division
18	2016	Garlapati[33]	Presented an idea of worldwide EW details, health concerns of EW components alongside the waste management, recycling.	Various regulatory, regulatory bodies and their problems, and corresponding suggestions, recommendations related to EW are not addressed
19	2016	Singh et. al. [34]	Given a review of world's present CRTs squander situation, to be specific extent of the request and preparing, current exchange and reusing operations.	Various questions have been raised but did not provide any solution.
20	2017	Ikhlayel[35]	Attempted to evaluate the environmental effects and	Given a review of world's present CRTs squander situation, to be specific greatness of the request and preparing, current exchange and reusing operations.

Fig. 1. Comparative Analysis of Electronic Wastages w.r.t. the Electronic Product.

TABLE II. POLLUTANTS AND THEIR RESPECTIVE COMMON SOURCES

Pollutants	**Sources of pollutants**
As	Electric diodes, Light Emitting Diodes, Semiconductors, μ-waves, Solar panels
Cu	Printed circuit boards, conductors, etc.
Ba	Filters, insulators
Cd	Cathode Ray Tubes, Cell-batteries, soldering alloys
Pb	Almost all kind of transistors and batteries
Cr	Hubs, Routing devices, switches, etc.
As – Arsenic, Ba – Barium, Cd – Cadmium, Cr – Chromium, Co – Cobalt, Cu – Cuprum/Copper, Pb - Lead	

In an existing framework for EW management in China counting regulatory policies are discussed along with the pilot projects. It also identified EW management challenges in Malaysia (UKM) and provided deeper knowledge about the different directives and policies needed to be followed while handling with EW and addressed how to enhance EW estimates by giving techniques to increase information quality and proposed an advanced IOA method including each of the three variables (sales, stock and lifespan) and best available information focuses to prepare better datasets for modeling. The Global E-waste Monitor 2020 concentrated upon informal EW recyclers who subsidize the environmental costs of Bangalore's IT blast. The challenges associated with increasing EW quantities were of big concern that explored the environmental contamination from EW recycling at numerous little formal and informal workshops in India. It proposed a novel calculation for establishing a standard methodology to manage and refurbish EW called EW Management and Refurbishment Prediction (EMARP), which can be adapted by refurbishing industries in order to improve their performance.

Table I indicates the comparative analysis with advantage and disadvantages of different authors. Fig. 1 shows

Comparative analysis of electronic wastages w.r.t. the electronic products.

The toxic substances contained in Waste Electric and Electronic Equipment (WEEE) are mostly non-degradable plastics [36]. The sources are as shown in Table II.

Apart from these, casings, PCBs, cables etc. are inevitably the most common toxics among assembled electronic devices. Firing, igniting the electronic waste materials are most commonly used today. In developing nations, these practices are done around human inhabitants or residential areas of cities. This lead to the formation of unsafe gasses which seriously damage the respiratory systems of each and every surrounding living being.

In some cases, we observe dumping these materials and landfilling which is less harmful than the preceding one. However, these may sometime become poisonous due to cyanide, lead, arsenic and mercury composition that contaminates at the out scale. Table II features the significant contaminations that cause serious health dangers when recycled improperly [37]. Separated of the few mentioned here, numerous elements pose as health risks.

III. Challenges and Open Issues

In this section, different challenges and issues related to e-waste management have been highlighted to seek attention of stakeholders and general public. Different initiatives and agreements have been established between post industrialized, industrialized, less industrialized and industrializing countries to save our environment and human health from the dangerous effects of e-waste. Various eco-friendly tools and techniques are being formulated for efficient disposal of e-waste with minimal environmental consequences. Moreover, various governmental and non-governmental organizations around the globe have joined hands to spread awareness among communities about e-waste management and related consequences to make them understand its need and criticality.

Despite the active participation of Governmental organizations, non-governmental organizations, industries, other stakeholders, the e-waste management is an 'open challenge' for both developed and developing countries. There are various issues that pose hindrance in the execution of systematic e-waste management strategies. Furthermore, the alarming increase in the amount of e-waste being generated around the years which is expected to reach a limit of 55 million tons in the year 2030 is really 'jaw-dropping'.

It has been observed by author in that with advancement in technologies paved way for change in composition of electronic and electric devices. So, the traditional recycling and disposal e-waste techniques are no longer applicable for the extraction of constituent elements. In Fig. 2 below, change in composition of Personal computers (PCs) over the time has been depicted. In earlier PCs, mainly three constituent elements were present like glass, plastic and metals but in recent PCs, we could varied elements like Zinc, Lead, Iron, Aluminium, Copper, etc.

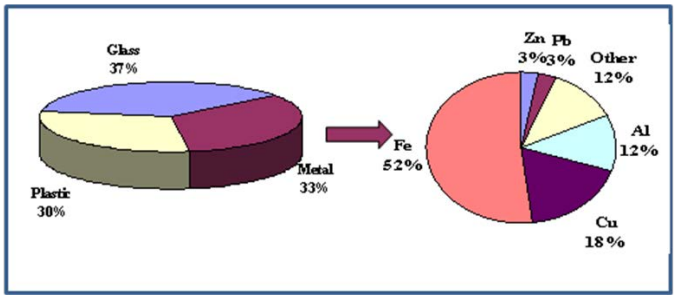


Fig. 2. Change in Composition of Personal Computers.

Therefore, it is evident that e-waste recycling and disposal methodologies need to revised and reformulated for the inclusion of changing composition of e-waste.

A. Switching to Transboundary Shipment

The management of e-waste in developing and transition countries is getting worse day by day. The major factors responsible for this are.

1) Illegal trafficking of waste materials
2) Unauthorized recycling of e-waste
3) Lack of skilled man power
4) No regulatory rules and regulations / protocols flourished for stakeholders and institutions that are involved in e-waste management.

It has become a normal trend that (a) the large amount of e-waste generated from developed countries are exported to developing countries or (b) those discarded by developing countries, are rapidly taken up by transition countries; and this has resulted into a very high impact on emerging economies of these countries.

B. International Treaties

One such example is Basel Convention in which it has been reported that.

- Only 20% of the global electronic wastage are recycled every year; i.e. 40 Mt electronic wastages are burned or destroyed or illegally traded. Further, 50% of these wastages consist of handheld devices and TVs along with household appliances.
- The rate of e-waste generated in emerging economies like India and China is increasing from 5% to 10% annually. These data do not include the imported electronic wastages whether they are legal or illegal.
- The first country in the world where e-waste management system was formally established is Switzerland where 11 kg/capita of e-waste are recycled. Their policy targets 4 kg/capita as regulated by the European Union (EU).

IV. Interoperability of Electronic Wastages

In this section, we characterize e-waste by investigating its criticality. The reuse of EW, state enactment, and the issue of the worldwide shipment of risky e-wastages are also adhered. Various hazardous development in the hardware business has prompted a quick raising issue of end-of-life (EOL) gadgets or

e-squander. Squandering of electronic is growing thus resulting in a high surge for the requirement of powerful hardware reusing programs.

It is assessed that 75% of electronic things are secured in view of defenselessness of how to supervise it. These electronic hurls out lie unattended in houses, work environments, dissemination focuses et cetera and regularly mixed with family wastes, which are at last orchestrated off at landfills. This requires implementable organization measures. In wanders organization of EW should begin at the explanation behind age. This should be possible by waste minimization methodologies and by supportable thing design. Waste minimization in endeavors incorporates getting a handle on:

1) Inventory organization.
2) Changing the production-process.
3) Volume reduction.
4) Recovery and reuse.

Challenges of EW Recycling: The toxic substances contained in Waste Electric and Electronic Equipment (WEEE) are mostly non-degradable plastics. The sources are as shown in Table III. Apart from these, casings, PCBs, cables etc. are inevitably the most common toxics among assembled electronic devices.

Firing, igniting the electronic waste materials are most commonly used today. In developing nations, these practices are done around human inhabitants or residential areas of cities. This lead to the formation of unsafe gasses which seriously damage the respiratory systems of each and every surrounding living being.

TABLE III. POLLUTANTS AND THEIR RESPECTIVE COMMON SOURCES

Pollutants	**Sources of pollutants**
As	Electric diodes, Light Emitting Diodes, Semiconductors, μ-waves, Solar panels
Cu	Printed circuit boards, conductors, etc.
Ba	Filters, insulators
Cd	Cathode Ray Tubes, Cell-batteries, soldering alloys
Pb	Almost all kind of transistors and batteries
Cr	Hubs, Routing devices, switches, etc.

As – Arsenic, Ba – Barium, Cd – Cadmium, Cr – Chromium, Co – Cobalt, Cu – Cuprum/Copper, Pb - Lead

TABLE IV. MAJOR POLLUTANTS AND THEIR EFFECTS OF HEALTH

Pollutants	**Major organic attack**
Lead	- Psychological degradation
Plastics	- Respiratory organs damage
Cadmium	Affecting - bones and joints - digestive systems
Acid Leachates	- Respiratory issues - Eye infection - Skin erosion

In some cases, we observe dumping these materials and landfilling which is less harmful than the preceding one. However, these may sometime become poisonous due to cyanide, lead, arsenic and mercury composition that contaminates at the out scale. Table IV features the significant contaminations that cause serious health dangers when recycled improperly. Separated of the few mentioned here, numerous elements pose as health risks.

Waste Electric and Electronic Equipment (WEEE) or EW joins a wide and growing collection of electronic devices going from tremendous family unit contraptions, for example, coolers, ventilation systems, PDAs, singular stereos, and buyer equipment to PCs disposed by the customers. These equipment needs to be well treated, mostly mercury, chromium, arsenic, lead, cadmium, and plastics responsible for various toxics.

V. SCHEMATIC FRAMEWORKS FOR E-WASTAGES PRACTICED TILL DATE

The electronic wastage has been practiced in a smaller note in most of the countries. Table III justifies this statement. The countries adhering to their policies also aren't able to implement the same as per the expectation. Fig. 3 is a schematic (conceptual) framework carried out based on the rigorous survey and relevant literature done by the authors. This clearly indicates that there are some of the unorganized sectors seem to be active in many regions. Due to the lack of proper policy, the authorities are not in a position to give the proper regulatory actions to be taken in care of these wastages. Even it is very disappointing to note that there is no reliable source available that can state about (a) the total number of organizations involved in such activities and (b) amount of e-waste that is recycled by the concerned dominants.

Descriptive analysis of our review work includes central tendency, mean and variation of variables of the study (See Fig. 1). Each of the variables are segregated into 24 aptitude questions.

A. Political or Regulating Factors (PRF)

The PRFs are analyzed based on four measuring questionnaires.

PRF1 – The degree of governments' initiation in managing EW

PRF2 – Properness of policy for EW Management.

PRF3 – Consciousness among of international community in EW management, and

PRF4 – Degree of following the international practices to handle the E-Waste

Fig. 4 depicts the bar diagrammatic view of PRF1, PRF2, PRF3 and PRF4. This clearly indicates the lack of significant policies for EW management. Proper policy helps to manage the e-waste in effective ways that is generating every year.

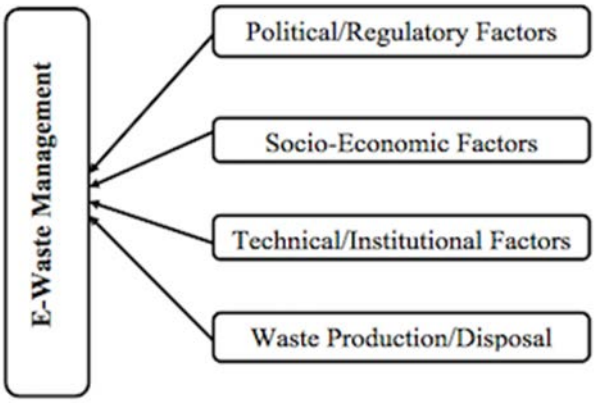


Fig. 3. Theoretical Framework of E-Waste.

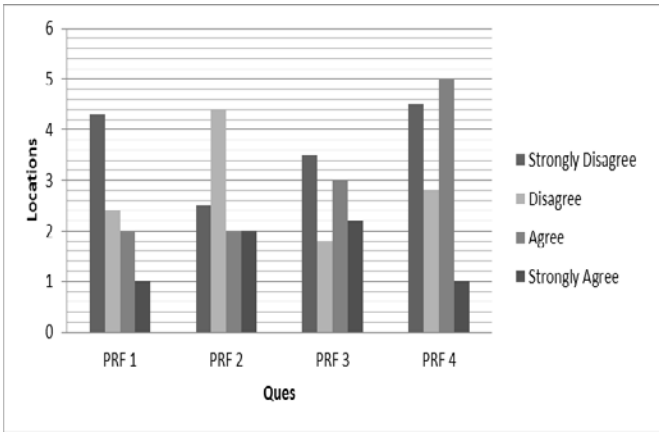


Fig. 4. Political/Regular Factor.

B. Waste Production/Disposal (WPD)

Four different questionnaires were been floored to analyze the effect of waste Production and Disposal.

WPD1 – referred to the scale at which the EW are generated on daily basis,

WPD2 – referred to the rate at which the electric or electronic equipment were discarded or donated per person,

WPD3 - referred to the status of discarded electric or electronic equipment – whether they are thrown away cumulating other waste, and

WPD4 - if those discarded or disposed electric and electronic equipment had any hazardous parts and that if any special method for safety disposal were applied.

Fig. 5 can be concluded with respondents' response that a special procedure is required utmost for disposal of electronic wastages. Due to the advanced technology, every year many electronics items are being produced and side by side every year electronics items are getting damaged and e-waste are produced. In developing countries, there is no advanced technology to manage this e-waste; i.e. damaged items are not disposed properly.

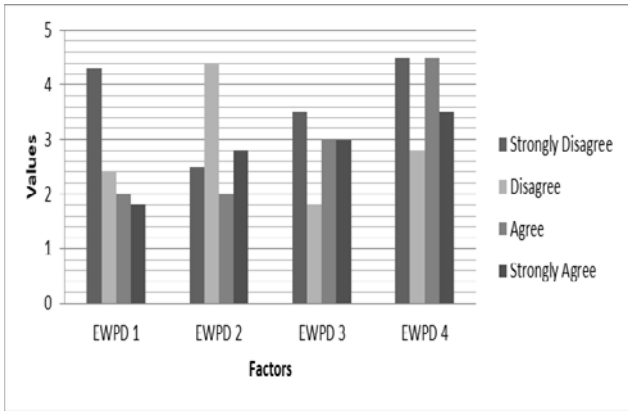


Fig. 5. Waste Production/Disposal.

C. E-Waste Management (EWM)

The four different questionnaires polled for EW management were as below:

EWM1 - awareness about EW management,

EWM2 – necessity of EW collector

EWM3 – convenience of the existing method for EW, and

EWM4 – accessibility of information regarding EW management.

Based on above queries, Fig. 5 is schematized. Fig. 5 depicts electronic wastage collector is the need of the hour which will ease the process of disposal. It will also reduce / resolve the confusion regarding purpose of e-waste.

VI. Findings and Recommendations

We have performed a rigorous review on electronic wastages. The reviews were based on five main factors. They are (a) Political / Regulating, (b) Social-Economic, (c) Technical / Institutional, (d) Waste Production / Disposal and (e) EW Management factors were examined. Based on analysis and various literature survey, the findings are listed as below:

1) Lack of regulatory body in tracing and handling EW
2) Lack of Solid governmental policies to tackle EW
3) Managing electronic wastes

a) from house hold, and
b) work-places

4) Reusability and finding out the possibility of reusable of these components or EW for future use without affecting any of the other factors

5) Regular updates regarding the current (and past) status of EW of various EEE

6) Finding out the correct sources that are responsible for EW

7) Assessing the barriers of EW management.

The survey was conducted among 151 individuals who were "understood as aware of EW" and data were collected in physical form. These individuals were categorized as individual working in technology related companies; electrical

or electronic (or both) shops; the individual working in handling the e-waste management. Out of 151 individuals, 61 females and 90 males comprising the percentage of 40.4% and 59.6%, respectively.

In addition to age, there are 37.7% of respondents who fall under the age group below 25 years, 56.3% respondents are aged 26-35 years, 2.6% respondents are of age group 36-45 years and 3.4% of respondents are of age group of 46 and 55 above year. 23.1% of respondent are student, 42% of respondents are Service holders, 11.9% of respondent are self-employed (Business) and, 22 % of respondent are engaged in other professions.

Based on the survey, data and findings indicated above, authors propose the recommendations as below:

1) Awareness / consciousness regarding managing EW is mandatory. There is a very less human resource working for EW, their efforts and activities need to be maximized one of which may be done through promotional activity.

2) It is well understood that the growth of generation of EW is exponential. Advanced technology as perturbed by UN (2014) should be availed specially to developing countries.

3) EW collector needs to avail to each sector for the E-Waste Management.

4) Technical manpower to be produced by the authorized body by providing related technical knowledge and thus generating skilled manpower in EW management.

5) The rapid advancement in ICT has result in improve capacity in computing devices but simultaneously decreases in the product lifetime as a result the generation of EEE components are growing rapidly. A concrete policy and budget in e-waste management should be allotted.

VII. Conclusion

Electric and electronic wastages can result into an environmental disaster. An analytical study was done regarding EW and it was found that despite of consciousness, and awareness about these kind of wastages, appropriate measures are still not initiated for proper monitoring, mitigating and managing the EW scientifically. Sources responsible for creating electric and electronic wastages were also discussed. No authorized sectors have been officially designated for the same in most of the developing countries which is the need of the hour. Therefore, this paper focuses on the importance of e-waste management and its management practice and identifies some key findings along with recommendations based on various socio-tech factors. Based on surveying data, the paper concludes that lack of regulation regarding e-waste is the major concern to manage the e-waste.

VIII. Future Work

To measure the reliability of this study, Cronbach's alpha will be used after the collection of data from wide sectors and a significant procedural conceptual model will be proposed.

Acknowledgment

We thank the Deanship of Scientific Research, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for help and support.

References

[1] Chancerel, P.; Meskers, C.E.M.; Hagelüken, C.; Rotter, V.S. Assessment of Precious Metal Flows During Preprocessing of Waste Electrical and Electronic Equipment. Journal of Industrial Ecology2009, 13, 791–810, doi:10.1111/j.1530-9290.2009.00171.x.

[2] Widmer, R.; H, O.-K. Global perspectives on e-waste. Environ Impact Assess Rev2005, 25.

[3] Zhou, L.; Xu, Z.M. Research progress of recycling technology for waste electrical and electronic equipment. Mater Rev2012, 26.

[4] Schwarzer, S.; AD, B. E-waste, the hidden side of IT equipment's manufacturing and use. UNEP Early Warning on Emerging Environmental Threats2005.

[5] Eugster, M.; Duan, H.B.; JH, L. Sustainable electronics and electrical equipment for China and the world. International Institute for Sustainable Development2008.

[6] Yu, J.; Williams, E.; Ju, M.; Shao, C. Managing e-waste in China: Policies, pilot projects and alternative approaches. Resources, Conservation and Recycling2010, 54, 991–999.

[7] Chibunna, J.B.; Siwar, C.; Begum, R.A.; Mohamed, A.F. The challenges of e-waste management among institutions: A case study of UKM. Procedia-Social and Behavioral Sciences2012, 59, 644–649.

[8] Sthiannopkao, S.; Wong, M.H. Handling e-waste in developed and developing countries: Initiatives, practices, and consequences. Science of the Total Environment2013, 463, 1147–1153.

[9] Wang, F.; Huisman, J.; Stevels, A.; Baldé, C.P. Enhancing e-waste estimates: Improving data quality by multivariate Input–Output Analysis. Waste Management2013, 33, 2397–2407, doi:10.1016/j.wasman.2013.07.005.

[10] Reddy, R.N. Producing abjection: E-waste improvement schemes and informal recyclers of Bangalore. Geoforum2015, 62, 166–174.

[11] Tansel, B. From electronic consumer products to e-wastes: Global outlook, waste quantities, recycling challenges. Environment international2017, 98, 35–45.

[12] Awasthi, A.K.; Zeng, X.; Li, J. Environmental pollution of electronic waste recycling in India: A critical review. Environmental pollution2016, 211, 259–270.

[13] Resmi, N.G.; Fasila, K.A. E-waste Management and Refurbishment Prediction (EMARP) Model for Refurbishment Industries. Journal of Environmental Management2017, 201, 303–308, doi:10.1016/j.jenvman.2017.06.065.

[14] Gangwar C, Choudhari R, Chauhan A, Kumar A, Singh A, Tripathi A. Assessment of air pollution caused by illegal e-waste burning to evaluate the human health risk. Environment international. 2019 Apr 1;125:191-9.

[15] Mohammed Yousuf Uddin, Sultan Ahmad and Mohammad Mazhar Afzal, "Disposable Virtual Machines and Challenges to Digital Forensics Investigation" International Journal of Advanced Computer Science and Applications(IJACSA), 12(2), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120299.

[16] Shevchenko, Tetiana, Michael Saidani, Yuriy Danko, Ievgeniia Golysheva, Jana Chovancová, and Roman Vavrek. 2021. "Towards a Smart E-Waste System Utilizing Supply Chain Participants and Interactive Online Maps" Recycling 6, no. 1: 8. https://doi.org/10.3390/recycling6010008.

[17] Kazancoglu Y, Ozbiltekin M, Ozen YD, Sagnak M. A proposed sustainable and digital collection and classification center model to manage e-waste in emerging economies. Journal of Enterprise Information Management. 2020 Jun 19.

[18] Nowakowski P, Pamuła T. Application of deep learning object classifier to improve e-waste collection planning. Waste Management. 2020 May 15;109:1-9.

[19] Jha S, Nkenyereye L, Joshi GP, Yang E. Mitigating and monitoring smart city using internet of things. Computers, Materials & Continua. 2020 Jan 1;65(2):1059-79.

[20] Gangwar C, Choudhari R, Chauhan A, Kumar A, Singh A, Tripathi A. Assessment of air pollution caused by illegal e-waste burning to evaluate the human health risk. Environment international. 2019 Apr 1;125:191-9.

[21] Kumar, A., Choudhary, V., Khanna, R. et al. Recycling polymeric waste from electronic and automotive sectors into value added products. Front. Environ. Sci. Eng. 11, 4 (2017). https://doi.org/10.1007/s11783-017-0991-x.

[22] Kumar A, Dixit G. An analysis of barriers affecting the implementation of e-waste management practices in India: A novel ISM-DEMATEL approach. Sustainable Production and Consumption. 2018 Apr 1;14:36-52.

[23] Salhofer S, Obersteiner G, Schneider F, Lebersorger S. Potentials for the prevention of municipal solid waste. Waste management. 2008 Jan 1;28(2):245-59.

[24] Borthakur, A.; Govind, M. Emerging trends in consumers' E-waste disposalbehaviour and awareness: A worldwide overview with special focus on India."Resources. Conservation and Recycling2017, 117, 102–113.

[25] Rahman, M.A. E waste management: A study on legal framework and institutional preparedness in Bangladesh. Thesis, North South University: Bangladesh, 2016.

[26] Debnath, B.; Roychoudhuri, R.; Ghosh, S.K. E-Waste Management–A Potential Route to Green Computing. Procedia Environmental Sciences2016, 35, 669–675.

[27] Dwivedy, M.; Mittal, R.K. Willingness of residents to participate in e-waste recycling in India. Environmental Development2013, 6, 48–68.

[28] Pariatamby, A.; Victor, D. Policy trends of e-waste management in Asia. Journal of Material Cycles and Waste Management2013, 15, 411–419.

[29] Jaiswal, A.; Samuel, C.; Patel, B.S.; Kumar, M. Go green with WEEE: Eco-friendly approach for handling e-waste. Procedia Computer Science2015, 46, 1317–1324.

[30] Kumar, S.; Rawat, S. Future e-Waste: Standardisation for reliable assessment. Government Information Quarterly2018, 35, S33–S42, doi:10.1016/j.giq.2015.11.006.

[31] Davis, J.M.; Garb, Y. A model for partnering with the informal e-waste industry: rationale, principles and a case study. Resources, Conservation and Recycling2015, 105, 73–83.

[32] Dwivedy, M.; Suchde, P.; Mittal, R.K. Modeling and assessment of e-waste take-back strategies in India. Resources, Conservation and Recycling2015, 96, 11–18.

[33] Garlapati, V.K. E-waste in India and developed countries: Management, recycling, business and biotechnological initiatives. Renewable and Sustainable Energy Reviews2016, 54, 874–881.

[34] Singh, N.; Li, J.H.; Zeng, X.L. Global responses for recycling waste CRTs in ewaste. Waste Management2016, 57, 187–197.

[35] Ikhlayel, M. Environmental impacts and benefits of state-of-the-art technologies for E-waste management. Waste Management2017, 68, 458–474.

[36] Sepúlveda, A. A review of the environmental fate and effects of hazardous substances released from electrical and electronicequipment during recycling: Examples from China and India. Environmental Impact Assessment Review2010, 30, 28–41.

[37] Janz, A.; Bilitewski, B. Hazardous substances in waste electrical and electronic equipment' in RakeshJohri, E-waste:Implications, Regulations and management in India and current global best practices; TERI, New, 2008.

SCADA and Distributed Control System of a Chemical Products Dispatch Process

Omar Chamorro-Atalaya[1], Dora Arce-Santillan[2]
Facultad de Ingeniería y Gestión
Universidad Nacional Tecnológica de Lima Sur (UNTELS)
Lima- Perú

Guillermo Morales-Romero[3], Nicéforo Trinidad-Loli[4], Adrián Quispe-Andía[5]
Universidad Nacional de Educación Enrique Guzmán y Valle
Lima-Perú

César León-Velarde[6]
Universidad Tecnológica del Perú (UTP)
Lima-Perú

***Abstract*—The objective of this article is to show the application of a supervision, control and acquisition system of an industrial network of chemical products, for which the design of the control logic and the architecture of the industrial network on Profibus-DP protocols is described and Ethernet, with a peripheral terminal station ET-200, through the Siemens CP433 programmable logic controller and level sensor sensors, coupled to radar-type transmitters with an accuracy of ± 0.5 mm. As findings of the implementation of the control system, it was possible to demonstrate the optimal regulation of the filling system of the 3-compartment trucks with a capacity of 300 Kilos each, generating the elimination of spills of the chemical product, as well as the reduction of polluting particles in the work environment. Finally, as a direct consequence, the productivity of the company was improved, which is a relevant aspect at the level of planning, management and direction.**

Keywords*—Distributed control; supervision; acquisition; chemical products; dispatch of chemical supplies

I. Introduction

Organizations dedicated to the distribution and dispatch of chemical products are susceptible to generating negative environmental impacts due to the lack of technological strategies that minimize risks of spillage [1], [2]. And it is necessary to be aware of and highlight the risks and consequences of not supervising or monitoring the correct process of transporting or transferring the chemical input for its storage and dispatch [3]. When these processes lack or are not linked to technological tools, but on the contrary respond to strictly manual processes, in which the personnel or operator is exposed to these chemical inputs, effects are totally adverse to health, since they are less likely intrinsic generates toxicity and pollutes the environment [4]. The maneuvers in the dispatch processes of chemical inputs when they are manual or executed by some personnel, increase the risk of spills, which, as already stated, affects the human being and the environment, but also affects production, due to actions corrective measures that must be generated at that moment [5]. It is clear that these organizations must get involved with technologies that reduce all types of spill risk, since the negative consequences are derived from there [6].

Industrial automation is currently used as a tool that contributes to improving various stages immersed in production processes, thus improving aspects related to precision in component dosing, accuracy in the sequencing of events, self-regulation and improvement of the location of the final product [7]. Automating a process guarantees the logical control relationship between sensors and actuators of a process, however through communication protocols it is necessary to seek to implement supervision, control and data acquisition systems [8]. These communications protocols guarantee the adequate interface in the sensors and actuators of a process, interrelating it through an industrial network, which allows the supervision and monitoring in real time of the variability of the indicators during the execution of the production process [9, 10].

Distributed control systems (DCS) allow an industrial plant to integrate its different production processes, linking the various programmable logic controllers (PLC), which are associated from a logical point of view to sensors and actuators, thus guaranteeing a fluid communication of data, which allows making decisions from the generation and processing of information [11, 12]. Referring to a distributed control system implies the generation of relevant data, which goes beyond operational aspects, becoming linked to very useful and important information systems at a strategic level in an organization [13]. With the purpose of supervising, monitoring and acquiring data for decision-making on aspects related to the operation or operation of an industrial process, SCADA systems (Supervisory Control and Data Acquisition) are used [14-17]. The application of SCADA systems is diverse, however among the processes that use it the most are electricity generation, natural gas, mining deposits, thermal systems, oil pipelines and chemical products dispatch [18-20]. One aspect to highlight in SCADA systems is that they allow in a friendly way to show the behavior of the critical variables in a process, becoming a strategic element at the industrial level, integrally linking geographically dispersed subsystems arranged in a plant [21-23].

It is then justifiable from the point of view of the monitoring, supervision and control capacity, the integration of DCS systems and SCADA systems [24-26]. As its application is diverse and by using different communication protocols as interface elements, its contribution is relevant, even more it is linked to the improvement of the operating and working conditions of the operators and their environment with the environment [27-29].

Under the above, the purpose of this article is to contribute through the description of the application of distributed control systems (DCS) under their integration to the SCADA system in a chemical products distribution company; seeking to reduce the remnants or spills of these inputs during their dispatch or filling in distribution trucks; guaranteeing the reduction of pollutants in the work environment and consequently eliminating any adverse effect on the personnel working in the plant.

II. Literature Review

In relation to relevant studies on this topic, we have in [30], which highlights the importance of the use of the SCADA system and its integration in the architecture of industrial processes, in which the performance and safety of processes related to oil industry or chemical inputs. In the same line of research in [31], a proposed solution is described for a process related to oil as a chemical input in which the SCADA system guarantees safety from the detection of abnormal events or events, thus avoiding any effect adverse on staff and work environment.

It should also be noted that this scientific article is supported by two studies already published in [32–34] which detail the composition of the industrial network that allows automating the chemical products dispatch process, as a solution to the reduction of pollutants due to spillage of the input due to a bad maneuver by the operators; These studies also highlight that as an effect of supervising, controlling and monitoring the dispatch system, from the perspective of improving the filling precision of the transport truck tanks, another positive effect, quite apart from the reduction of pollutants, is productivity improvement, a very relevant contribution to the organization, and which gives sustainability to a large-scale solution proposal such as the implementation of a distributed control system integrated into a SCADA system.

The implementation of intelligent systems through hardware and software, allow to guarantee the control of the operating parameters and variables of the industrial processes [35, 36].

III. Control Philosophy and Architecture of the Distributed Control Network

In order to understand the control logic that allows automating the chemical products dispatch process, in a simplified way the control philosophy represented in the flow diagram is shown in Fig. 1.

In the flow chart, two sensors are established that will determine the correct location of the containers for the filling process; if his position is not adequate, he will not proceed to execution; furthermore, only if the filling level of the container is adequate, a sensor will establish that the valves should be closed and proceed to fill the next container; It is necessary to indicate that the number of compartments to fill in a truck storing the chemical product is three.

In Fig. 2, the location and disposition of the three sources that contain the chemical product (Acetic acid) are schematized, and that from displacement pumps, valves controlled by electric actuator, overfill sensors, the same ones that were mentioned in the flow chart above; and making use of Siemens CP443 programmable logic controllers, I will proceed with the programming logic that establishes the link between the input and output elements of the process.

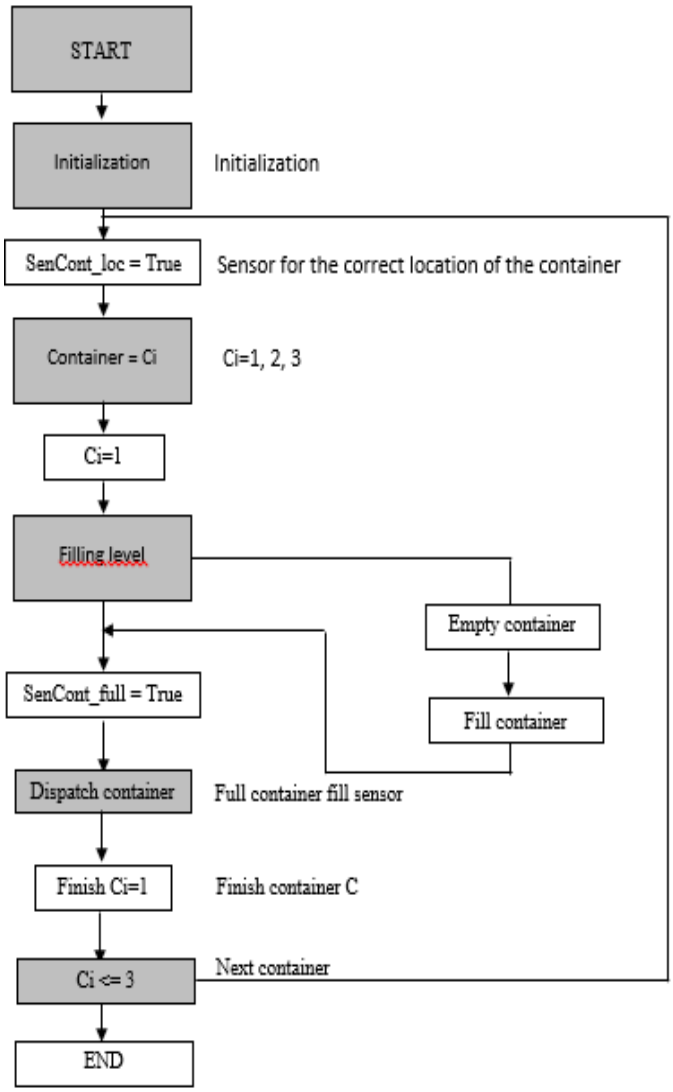


Fig. 1. Control Philosophy of the Chemical Product Dispatch Process.

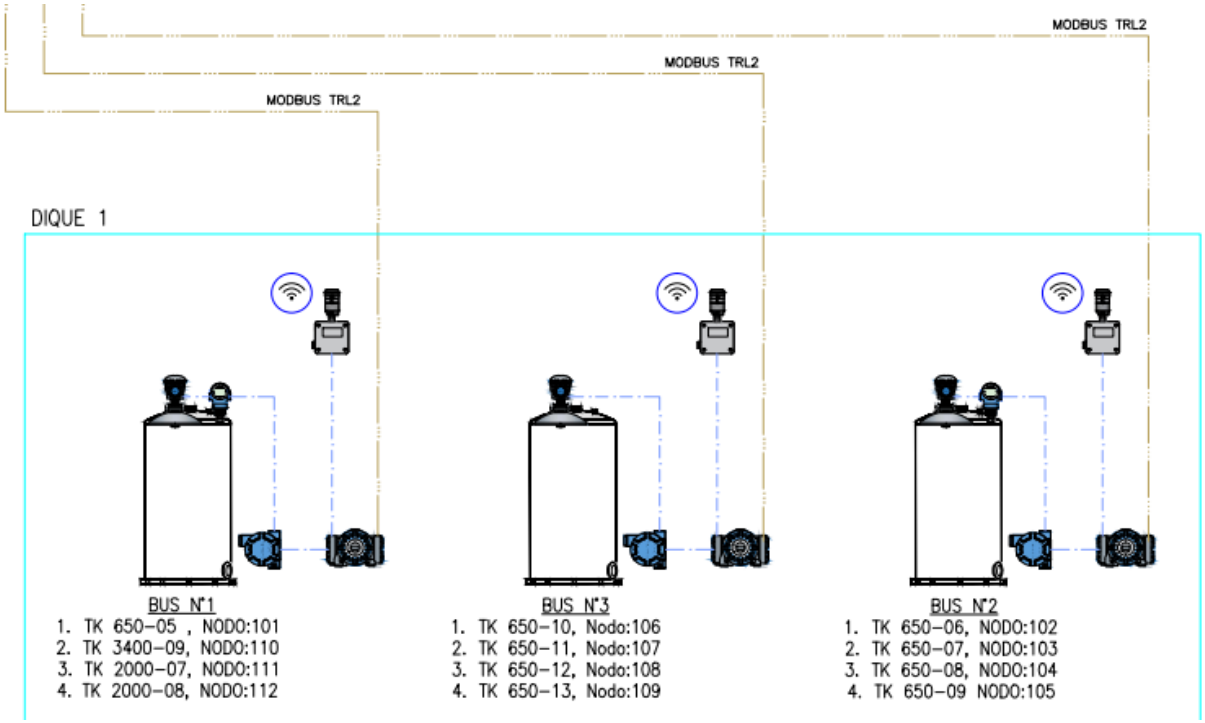


Fig. 2. Layout of Chemical Storage Tanks.

In Fig. 3, the connection between the sensors, actuators, scale zone and displacement pumps for filling with respect to the chemical product packaging trucks is shown; It should be noted that the pumps in operation operate in remote mode and in automatic sequence; in such a way that the valves also respond to a similar operating logic, that is, in remote mode and automatic sequence, which indicates that it is open; The displacement pump as well as the valves will always be monitored through the supervision and data acquisition system, the same as from the WinCC runtime Project system will allow the activation and shutdown of the mentioned elements and start the dispatch process. It should be noted that the valve is a 3" butterfly type that is located on the second level of the truck station between the 45 ° elbow and the start of the stainless steel hose. This system developed through the WinCC Runtime Project system will trigger an alarm 300 kilos before reaching the indicated weight for that compartment.

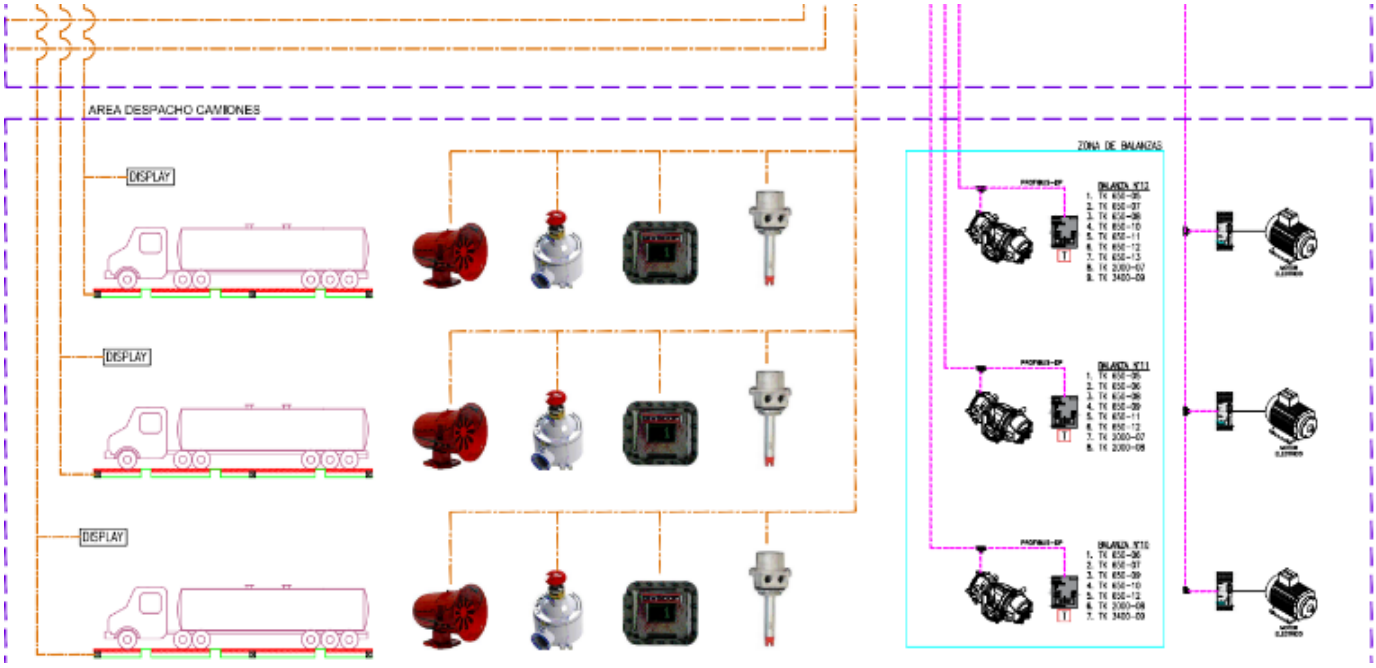


Fig. 3. Connection between Sensors, Actuators, Scale Zone and Displacement Pumps with respect to Packaging Trucks.

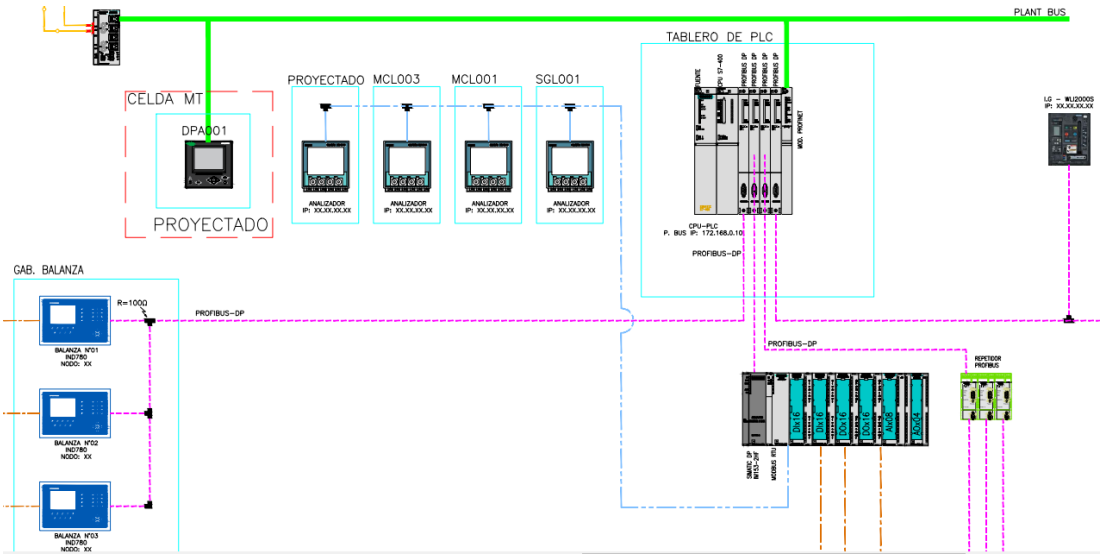


Fig. 4. Connection of Automated Industrial Network System Components.

The data server of the industrial network will be located in the server room and will be interconnected through a fiber optic network with the instrumentation equipment and located in the industrial instrumentation cabinets. The communication switch of the control system will allow the interconnection and exchange of information between the programmable logic controller and the data server, for this reason they must be installed in the same cabinet mentioned above, in accordance with the Control architecture, shown in the Fig. 4. For the monitoring of electrical variables, it is considered that all the data received from the multifunction meters will be sent to the data server for its proper use; in such a way that the displacement pumps must be composed of local control pushbuttons (located in the field, for starting and stopping the engine), selectors and status indicators. All remote start and stop signals coming from each programmable logic controller will be using the control logic and sent to the control center through Profibus DP communication.

IV. Control Board and Development of the SCADA System Configuration

A. Control Panel

In Fig. 5, the electrical control cubicle is shown in which the main switch that energizes the entire control system of the previously selected transfer pump is enabled.

Fig. 5. Control Cube where the Central Switch is Located.

In Fig. 6, it is shown that to confirm that there is no blockage enabled in the field, this due to the incorrect location of the chemical product packaging trucks, it is necessary that the state of the switches that do not have the insurance be verified in the field placed. When the safety lock is in stop mode, it will prevent the transfer pump from being controlled from the control room in automatic mode.

Fig. 6. Safety Switch Status.

Fig. 7. Automatic Remote Actuation Switch and Scale Platform.

In Fig. 7, the scale platform is shown, in which it will be verified that the selector is in the remote mode position, as shown in the figure. It should be noted that remote and automatic operation is achieved in the control mode shown.

B. SCADA System Configuration

A first relevant aspect for the configuration of the SCADA system is related to access security, the criteria of which has been considered at the time of making the configuration, for which the operation of the equipment has been restricted only to authorized users; in such a way as to avoid that personnel outside the operation of the plant can mainly start up equipment without prior permission; that is, only by entering the username and password can they be enabled. Fig. 8 shows the configuration window in which the enabling of the key is highlighted.

Once your user has been entered, the system will enable the access screens or menus, which are located in the upper part of the supervision system, it is available throughout the project. Through these menus we can enter any screen according to the user's requirement. These menus display a collection of screens classified by function; as evidence of the above, Fig. 9 is shown.

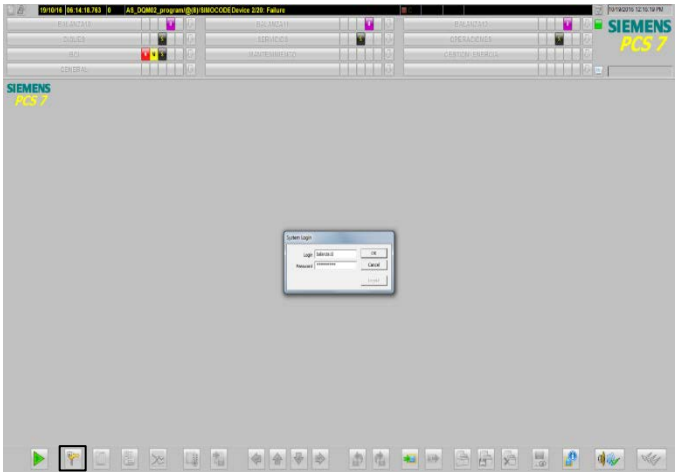


Fig. 8. Configuration of Restricted Access to the SCADA System.

Fig. 9. SCADA System Supervision Menus Configuration.

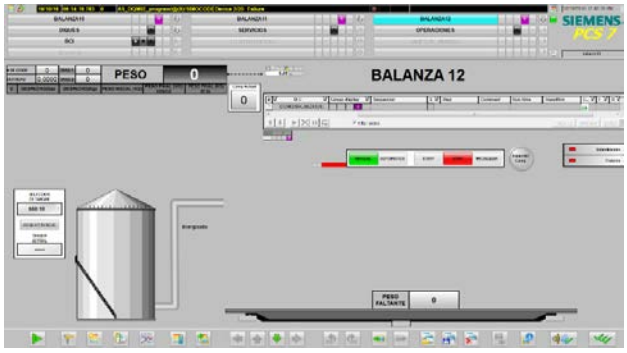


Fig. 10. Setting the Balance Selection in the SCADA System.

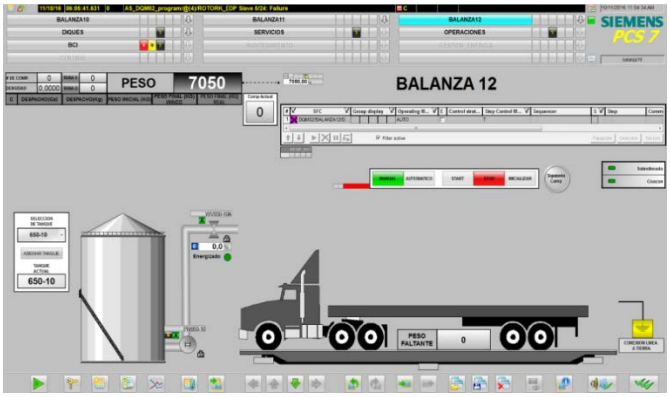


Fig. 11. Configuration of the Tank Selection and Start of the Dispatch Process, in the SCADA System.

Once the scale is entered from where the chemical product will be dispensed, the selected scale will be indicated by the background color of the selected tab. Once located on the balance, the following distribution will be available as shown in Fig. 10.

This is also configured to make the selection of the tank, which must be correctly located to start the dispatch process. Immediately, the actuator and the displacement pump will be enabled in automatic operating mode, with which the tanker truck or also called container truck will enter the scale platform recording the initial weight, so that next the sensor will be connected overfill and sensor ground, as shown in Fig. 11.

V. Result and Discussion

A. Results

As a result of the integration of the distributed control system and the SCADA system of the chemical product filling process in Fig. 12, the monitoring and supervision obtained in real time from the SCADA system is already shown, in which the valve is located at a 50% opening restricting the access of the chemical product, for when the missing weight is less than 300 kg. The displacement pump motor will keep its nominal speed, while in the field the beacon and siren will be active for 3 seconds, with the objective to notify the scale operator of the completion of the chemical product dispatch.

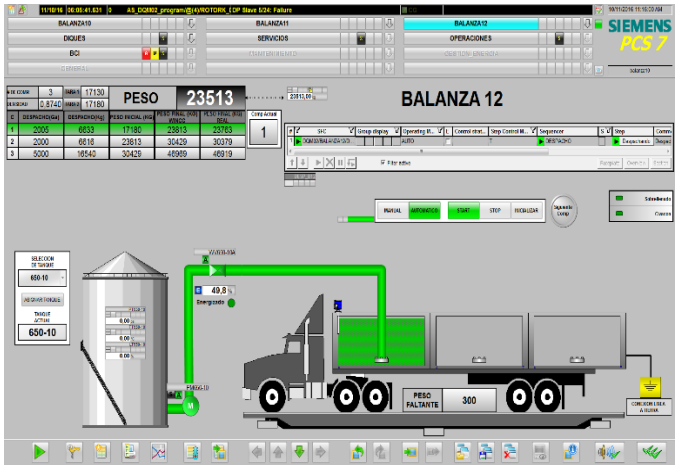


Fig. 12. SCADA System Commissioning: Filling System Control, Valve Restriction to 50%.

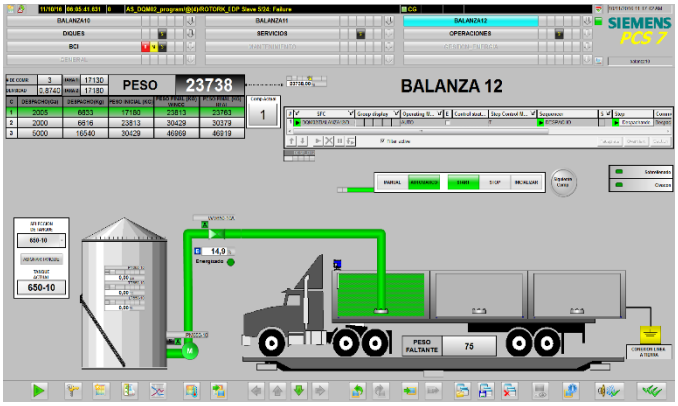


Fig. 13. SCADA System Commissioning: Filling System Control, Valve Restriction to 75%.

In Fig. 13, it is shown that when the missing weight is less than 150kg, the electric actuator will proceed to close to 75%.

In Fig. 14, it is shown that when the missing weight is less than 75kg, the electric actuator will proceed to close to 85%.

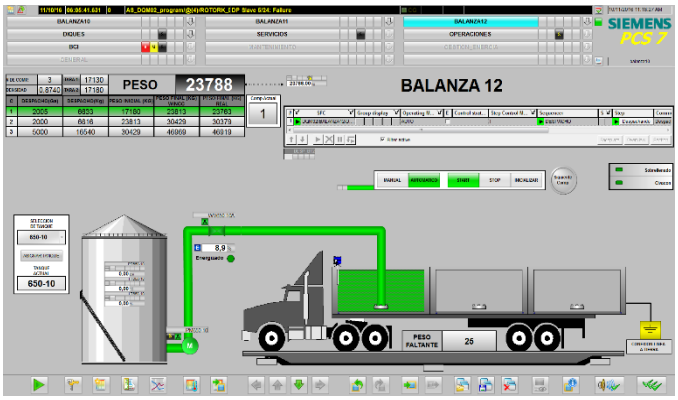


Fig. 14. SCADA System Start-Up: Filling System Control, Valve Restriction to 85%.

Upon reaching the missing weight, the electric actuator will fully close and the displacement pump will stop. The "SenCont_full" button will flash, showing that the dispatch in container 1 has ended, moving on to the next container. As evidence of the above, it is evidenced in Fig. 15.

Fig. 15. Commissioning of the SCADA System: Complete Filling of Container 1.

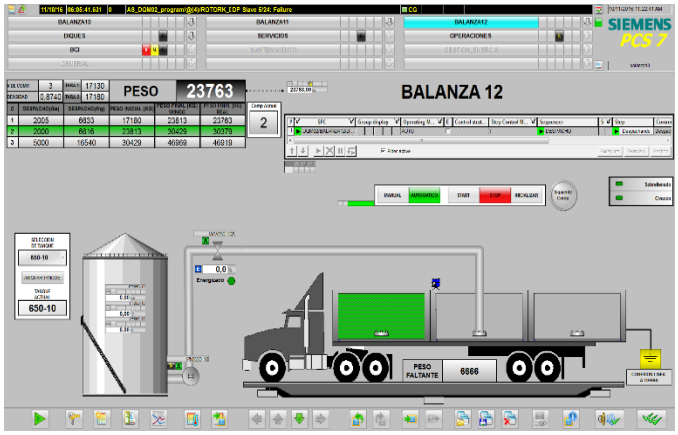


Fig. 16. Commissioning of the SCADA System: Filling Process of Container 2 of the Tank Truck.

In Fig. 16, it is shown that the tank truck tank 2 is automatically filled; for which precision and total elimination of spillage of the chemical product is guaranteed.

B. Discussion

According to the results evidenced in the screenshots of the SCADA system, it shows correct operability and optimal performance in the chemical product dispatch process, from the integration with the distributed control system developed under the Siemens CP443 controller and structured under an industrial network with Profibus DP protocol and remote station simenes ET 200, in this regard in [8] the author also obtains as a finding an optimal functioning in the integration of the distributed control system of a plant in which the flow of energy is monitored and regulated. raw material; highlighting that his proposed solution is developed using the Profibus DP and Ethernet protocol; however, he used as part of his solution a CP-343 controller also from Siemens and an ET200 peripheral junction terminal. The difference between the controllers used is due to the fact that in my solution proposal, I control a greater number of actuators, sensors and that I seek that the industrial rd has scaling capacity due to terminals projected to be implemented in the future.

Another aspect to highlight is the improvement in filling precision and eliminated any possibility of loss of the chemical product due to spillage, it has significant relevance at the management level because it generates an impact on productivity. In this regard, in [10] the author applies a SCADA system for dosing the entry of basic reagents to three tanks, for which he used, as in the proposed solution, a Siemens 400 controller, in which he evidenced in his research that the Automated system will contribute to the improvement of productivity in the organization. Also in [23] it is highlighted that from the implementation of a SCADA system to a process linked to the improvement of the dosage of components in the stage of crushing and pulverization of inputs, it contributed to the improvement of results in the productivity of the company.

A third point to highlight as part of the development of this article is the contribution to reducing environmental pollution in the work environment; The fact is that when the process of dispatch of inputs or chemical products is automated, there will

no longer be any spillage, so its impact is on the generation of polluting particles in the environment; In this regard, in [30] the author points out that in his study the distributed control system and the SCADA system showed an optimal functioning of the supervision and control process in liquid petroleum plants, guaranteeing the safety of the work environment and minimizing the conditions abnormal operation of the process.

VI. Conclusion

It is concluded that the distributed control system (DCS), implemented from a Siemens CP443 unit, the same that was structured under the philosophy of an industrial network with Profibus DP protocol, remote station Siemens ET-200, and integrated into a SCADA (developed through the WinCC Runtime Project system) in a chemical product dispatch plant (acetic acid), showed to work optimally and efficiently, guaranteeing a regulation of the container filling or dispatch system composed of 3 compartments of 300 Kilos, the same that regulates from the gradual opening of valves the filling level of acetic acid. The most significant contributions that are generated as an effect of this control are the elimination of the spillage of chemical product as well as the reduction of polluting particles in the work environment and finally improvement of productivity, a relevant aspect at the level of planning, management and direction of the company.

Acknowledgment

The recognition to J. Yataco-Yataco for his contribution in the information and contributions from his experience obtained in the company.

References

[1] J. Piñeros, “Chemical risk assessment applying the Colombian technical standard at the colsubsidio norte school facilities,” Thesis, Jose Francisco de Caldas District University, Colombia, 2020.

[2] A. Bakri, et al., “Addressing the Issues of Maintenance Management in SMEs: Towards Sustainable and Lean Maintenance Approach," *Emerging Science Journal*, Vol. 5, pp. 367-379, 2021. DOI: http://dx.doi.org/10.28991/ESJ-2021-01283.

[3] K. Sigueñas and M. Tipto, “Logistic improvement for the optimization of the discharge of hazardous materials toluene and xylene from IPE Tralsa,” Thesis, Maritime University of Peru, Peru, 2020.

[4] Z. Villavicencio, “Quality control applied to cleaning products manufactured by a chemical supply warehouse and recording the results in SAP-HANA,” Thesis, National University of San Agustín de Arequipa, Perú, 2018.

[5] R. Martinez, “Evaluation of the prevention of attention to spills in linear projects in the hydrocarbon sector: case of oil pipelines,” Thesis, University of America, Colombia, 2020.

[6] D. Gutiérrez, “Proposal for the improvement and update of safety and health in chemical inputs for A.G.P de Colombia S.A,” Thesis, University Francisco Jose de Calda, Colombia, 2016.

[7] C. Molina, W. Gonzáles and G. Cruz, “An approach to teaching from the CTS approach,” *University and Society Jurnal*, vol. 10, pp. 221-226, 2018.

[8] O. Chamorro-Atalaya, D. Arce-Santillan, T. Diaz-Leyva and M. Díaz-Choque, “Supervision and cotrol by SCADA of an automated fire system," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 21, pp. 92-100, 2021. DOI: http://dx.doi.org/10.11.591/ijeecs.v21.i1.pp92-100.

[9] L. Silva-Diaz, Y. Hernández-López and A. Vázquez-Peña, “Design of an automation system for the silage plant Hector Molina,” *Journal of Agricultural Technical Sciences*, Vol. 26, pp. 109-120, 2017.

[10] O. Chamorro-Atalaya, D. Goicochea-Vilela, D. Arce-Santillan, M. Díaz-Choque and T. Diaz-Leyva, “Automation of the burner of a pirotubular boiler to improve the efficiency in the generation of steam," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 21, pp. 101-109, 2021. DOI: http://dx.doi.org/10.11.591/ijeecs. v21.i1.pp101-109.

[11] J. Moscoso, “Integration of different distributed control systems (DCS) using standard software and OPC technology,” Thesis, National University of San Agustín de Arequipa, Perú, 2014.

[12] S. Marzal, “Conception and Integration of Architectures and communication protocols within Intelligent Microgrids supervision and Control systems,” Thesis, Polytechnic university of Valencia, España, 2019.

[13] J. Simó-Ten, J. Poza-Lujan, J. Posadas-Yague and F.Blanes, “Study and implementation of Middleware for distributed control applications,” XXXVIII Automatic Conference, Vol. 1, pp. 942-951, 2020. DOI: http://dx.doi.org/10.17979/spudc.9788497749.0942.

[14] D. Aguirre, “Development of a SCADA system for use in small and medium-sized companies,” Thesis, University of Piura, Perú, 2013.

[15] U. Sohail, M. Muneer and F. Khan, “An efficient aproach of load shifting by using SCADA”, *Advances in Science, Technology and Engineering Systems Journal (ASTES)*, Vol. 1, pp. 1-6, 2016. DOI: http://dx.doi.org/10.25046/aj010301.

[16] M. Syamsul, Susanto, D. Stiawan, M. Yazid and R. Budiarto, “The trends of supervisory control and data acquisition security challenges in heterogeneous networks,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 22, pp. 874-883, 2021. DOI: http://dx.doi.org/10.11591/ijeecs.v22.i2.pp874-883.

[17] M. Varghese, A. Manjunatha and T. Snehaprabha, “ method for improving ripple reduction during phase shedding in multiphase buck converters for SCADA systems,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 24, pp. 29-36, 2021. DOI: http://dx.doi.org/10.11591/ijeecs.v24.i1.pp29-36.

[18] E. Pérez-López, “SCADA Systems in the Industrial Automation,” *Technology in March Journal*, Vol. 28, pp. 3-14, 2015.

[19] A. Soetedjo, A. Lomi and Y. Nakhoda, “Smart grid Testbed using SCADA software and Xbee Wireless Communication,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 6, pp. 86-92, 2015. DOI: http://dx.doi.org/10.14569/IJACSA. 2015.060811.

[20] M. Lakhoua, B. Hamounda, R. Glaa and E. Amraoui, “Contributions to the Analysis and the Supervision of a Thermal Power Plant,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 7, pp. 93-101, 2016. DOI: http://dx.doi.org/10.14569/ IJACSA.2016.07 0213.

[21] N. Gaitan, “MCIP Client Application for SCADA in Liot Environment,” *International Journal of Advanced Computer Science and Applications (IJACSA)*,Vol. 6, pp. 158-163, 2015. DOI: http://dx.doi.org/10.14569/ IJACSA.2015.06 0921.

[22] T. Simona-Anda, “A solution for the Uniform Integration of field devices in an industrial Supervisory Control and Data Acquisition System,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 9, pp. 319-323, 2018. DOI: http://dx.doi.org/10.14569/IJACSA.2018.090344.

[23] M. Bernal and D. Jimenez, “Risk management proposal for SCADA in electrical systems,” *USBMed Journal*, Vol. 3, pp. 12-21, 2012.

[24] J. Ochoa-Hernandez, M. Barcelo-Valenzuela, F. Cirett-Galán and R. Luque-Morales, "A model to develop SCADA-type systems in productive environments,” *Computing and Systems Journal*, Vol. 22, pp. 1543-1558, 2018. DOI: http://dx.doi.org/10.13053/CyS-22-4-2823.

[25] I. Tawiah, U. Ashraf, Y. Song and A. Akhtar, “Marine Engine room alarm Monitoring System,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 9, pp. 319-323, 2018. DOI: http://dx.doi.org/10.14569/IJACSA.2018.090659.

[26] L. Rajesh and P. Satyanarayana, “Detecting Flooding attacks in communicacition protocol of Industrial Control Systems,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 11, pp. 396-401, 2020. DOI: http://dx.doi.org/10.14569/IJACSA. 2020.0110149.

[27] L. Barzaga, R. Mompie and B. Valdés, "SCADA systems for the automation of the CIGB production processes," RIELAC Journal, Vol. 37, pp. 20-37, 2016. http://scielo.sld.cu/scielo.php?script =sci_abstract& pid=S1815-59282016000100003.

[28] G. Tzokatziou, L. Maglaras, H. Janicke and Y. He, "Exploiting SCADA vulnerabilities usig a Human Interface Device," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 6, pp. 234-241, 2015. DOI: http://dx.doi.org/10.14569/IJACSA.2015.060731.

[29] Y. Chatei, E. Ar-Reyouchi, M. Hammouti and K. Ghoumid, "Dowlink and Uplink mesaage siza impact on round trip time metric in Multi-Hop Wireless mesh Networks," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 8, pp.223-229, 2017. DOI: http://dx.doi.org/10.14569/IJACSA.2017.080332.

[30] T. Simona-Anda, "Performances Analysis of a SCADA Architecture for Industrial Processes," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 8, pp. 456-460, 2017. DOI: http://dx.doi.org/10.14569/IJACSA.2017.081155.

[31] K. Chkara and H. Seghiouer, "Criteria to Implement a Supervision System in the Pretroleum Industry: A Case in a Terminal Storage Facility," *Advances in Science, Technology and Engineering Sytems Journal (ASTES)*, Vol. 5, pp. 29-38, 2020. DOI: http://dx.doi.org/ 10.25.046/aj050505.

[32] O. Chamorro-Atalaya, J. Yataco-Yataco and D. Arce-Santillan, "Industrial network for the control and supervision of the acetic acid dispatch process, and its influence on the reduction of chemical contaminants for operators," *Advances in Science, Technology and Engineering Systems (ASTES)*, vol. 5, pp. 13-20, 2020. DOI: http://dx.doi.org/10.25046/aj050103.

[33] O. Chamorro-Atalaya, A. Sanchez-Ayte, C. Dávila-Ignacio, O. Ortega-Galicio, N. Alvarado-Bravo and A. Torres-Quiroz, "Automatic control of motors through Simocode pro, and its effect on the performance of the process of filling and dispensing of chemical inputs," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS*), Vol. 23, pp. 179-187, 2021. DOI: http://dx.doi.org/10.11591/ijeecs.v23.i1. pp179-187.

[34] J. Yataco-Yataco, "Design of an industrial network for the control and supervision of the product dispatch process chemicals, to reduce the operator's exposure to harmful vapors in the company Chemical Deposits Mineros S.A.," Thesis, Alas Peruanas University, Perú, 2018.

[35] X. Qin and Y. Tang, "Design of Intelligent Drop Fuse," *Journal of Human, Earth, and Future*, Vol. 1, pp. 146-152, 2020. DOI: http://dx.doi.org/10. 28991/HEF-2020-01-03-04.

[36] I. H. Wayangkau, et al. "Utilization of IoT for Soil Moisture and Temperature Monitoring System for Onion Growth," *EmergingScience Journal*, Vol. 4, pp. 102-115, 2021. DOI: http://dx.doi.org/10.28991/esj-2021-SP1-07.

Supervised Learning through Classification Learner Techniques for the Predictive System of Personal and Social Attitudes of Engineering Students

Omar Chamorro-Atalaya[1], Soledad Olivares-Zegarra[2]
Facultad de Ingeniería y Gestión
Universidad Nacional Tecnológica de Lima Sur (UNTELS)
Lima- Perú

Alejandro Paredes-Soria[3], Oscar Samanamud-Loyola[4], Marco Anton-De los Santos[5], Juan Anton-De los Santos[6], Maritte Fierro-Bravo[7]
Facultad de Ciencias Económicas
Universidad Nacional Federico Villarreal
Lima-Perú

Victor Villanueva-Acosta[8]
Facultad de Ciencias Humanas
Universidad Autónoma del Perú
Lima-Perú

***Abstract*—In this competitive scenario of the educational system, higher education institutions use intelligent learning tools and techniques to predict the factors of student academic performance. Given this, the article aims to determine the supervised learning model for the predictive system of personal and social attitudes of university students of professional engineering careers. For this, the Machine Learning Classification Learner technique is used by means of the Matlab R2021a software. The results reflect a predictive system capable of classifying the four satisfaction classes (1: dissatisfied, 2: not very satisfied, 3: satisfied and 4: very satisfied) with an accuracy of 91.96%, a precision of 79.09%, a Sensitivity of 75.66% and a Specificity of 92.09%, regarding the students' perception of their personal and social attitudes. As a result, the higher institution will be able to take measures to monitor and correct the strengths and weaknesses of each variable related to satisfaction with the quality of the educational service.**

Keywords*—Supervised learning; classification learner; predictive system; personal and social attitudes; engineering students

I. Introduction

At present, the permanent search for educational quality is one of the main objectives of this sector [1], which is why strategies and methodologies designed to optimize student satisfaction factors [2], [3] have been implemented. Specifically, in the field of university higher education, the need to comply with quality standards in the educational service offered is evident [4].

Therefore, continuous self-evaluation of the dimensions related to student satisfaction is necessary, in order for this to improve institutional processes by identifying their strengths and weaknesses [5]-[8]. Taking into account the various influencing factors, it is complex to determine the strategic actions and decisions that correctly optimize these factors [9].

One of the factors that are related to student satisfaction is the self-perception of personal and social attitudes of university students. As indicated in [10] there is a need to evaluate the self-perception of personal and social attitudes in students, since this will generate the obligation to reflect on how the university community is contributing to social development. The importance of identifying the self-perception of the university's personal and social attitude lies in identifying what aspects need to be improved, in such a way that this leads to a significant contribution to higher education in the social context, allowing the student to meet their professional goals and personal [11], [12].

As indicated in [13], the higher educational level, has the duty to instill in university students the sense of social responsibility proper to the performance of professional activity. In this sense, the consequence of the acquisition of personal and social attitudes is a process that encompasses a significant portion in the lives of students in their passage through different organizations. Now if we develop this analysis in the context of educational virtualization, it leads us all to wonder to what extent this scenario has altered aspects related to personal and social attitudes of university students [14], [15] taking into account. As indicated in [16] that engineering students have perceived a greater change in the teaching process, due to the theoretical and practical nature of the subjects of the curriculum.

Given this, over time, information technology systems have been designed for different organizational sectors, from transactional to decision-making [17]. Within these information technology systems is the data mining tool. Data mining is conducive to the treatment of large amounts of information, and its purpose is to generate knowledge [18], [19]. Data mining uses databases, from which information is extracted in an automated way and through mathematical and statistical analysis deduces patterns and trends [20]. Data

mining makes data analysis easier, compared to traditional exploration, which, due to the large amount of data, makes this process much more complex [21].

Regarding the education sector, Educational Data Mining (EDM) is an emerging discipline that seeks to develop methods to explore data generated in the education sector, in order to achieve a better understanding of the characteristics of students and the way they learn [22]. Its development uses statistical techniques and artificial intelligence to detect patterns and anomalies in large amounts of data [23].

Within the field of artificial intelligence is Machine Learning, this learning is an automated process that extracts patterns from the data for the construction of models that allow prediction using supervised algorithms [24], [25]. One of the learning modalities that Machine Learning has is supervised learning, whose function is based on training the algorithm by granting it the questions, called characteristics, and the answers, called classes, in order that the algorithm combines them and can make predictions [26].

Within the two supervised learning techniques, is the classification, the classification algorithms look for patterns that will then allow them to classify the elements and determine which groups or classes they belong to. It should be mentioned that the values for these algorithms must be discrete values [27]. Among the classification algorithms is Kernel, which extends the regular logistic regression, used for binary classification, to deal with data that are not linearly separable [28].

Given what has been described, the need arises to design a predictive model of the personal and social attitudes of university students, which allow optimizing the services offered by the higher institution. Taking into account even more that the personal and social factors of the students are related to their academic performance [29], [30]. The research takes on even more relevance, due to the fact that, as indicated in [31], most predictive analysis research is related to primary and secondary education, so there is a small number of applications in higher education that serve the institutions as a base source for the improvement of educational quality.

In this sense, the main objective of this article is to determine the supervised learning model using the Classification Learner technique for the predictive system of personal and social attitudes of university students of professional engineering careers. To do this, it will analyze the performance metrics such as Accuracy (A), Precision (P), Sensitivity (S) and Specificity (R), to determine the most appropriate algorithm for the predictive model, also, the confusion matrix and the curve will be identified receiver operating characteristic (ROC) of the model.

The purpose of the research is to generate a significant contribution, for the taking of preventive and corrective actions that allow to comply with the quality standards of the educational service, whose improvement will be reflected in the satisfaction of the students and the academic performance.

The structure of the research development is divided into the methodology where the level of research is detailed, the participants, the data collection techniques, the validation of the collected data and the design of the supervised learning methodology through the technique classification. Next, the results and the discussion of them are presented, to finally make the conclusions of the investigation.

II. Research Methodology

A. Research Level

The research is descriptive in nature, since it focuses on determining the most optimal model of supervised learning using the Classification Learner technique for the predictive system of personal and social attitudes of university students of professional engineering careers, through the analysis of the performance metrics of the obtained algorithm (Accuracy (A), Precision (P), Sensitivity (S) and Specificity (R)).

The research starts from the identification of a problem, related to the improvement of the quality of the educational service, which is reflected in the satisfaction of the university students. Student satisfaction encompasses different dimensions, this study focuses on the self-perception of engineering students from a public university in Peru with personal and social attitudes. To do this, use of methods or tools already defined such as predictive systems through the modality of supervised learning.

This research also seeks to design a predictive multidimensional model that can be used to create and store new data for the higher institution. Based on this technological tool, it determines patterns and calculates association rules, providing support and reliability to the results obtained.

B. Participants

The participants in this research are made up of students from the sixth to the tenth cycle of professional engineering schools, with a total of 715 students. This selection criterion is part of a regulation established and approved by the public university of Peru. Due to the mandatory nature of the survey, it was possible to collect data from all participants.

TABLE I. Indicators that Measure the Self-Perception of Personal and Social Attitudes

Item	Indicators
1	Take on your studies with responsibility, seriousness and dedication
2	With the pride of belonging to the university
3	With the commitment of raising the name of the university
4	With the respect you show for the authorities, teachers and administrative staff
5	With the respect that you treat your colleagues
6	With the treatment you receive from your colleagues
7	With your interest to be better every day
8	With your commitment to the surrounding society

C. Data Collection Technique and Instrument

The data collection technique is the survey, and the instrument used to collect data regarding the self-perception of engineering students about personal and social attitudes is the questionnaire, which was carried out online, at through the university's virtual platform, due to the implementation of the online teaching-learning process in the context of the health emergency. In addition, access to the online questionnaire was given through the student code of each participant.

It is necessary to indicate that the questionnaire was made up of 8 indicators (characteristics), which are detailed in Table I. Likewise, the responses to the questionnaire for the present study were transformed into a 4-level Likert scale (1: dissatisfied, 2: not very satisfied, 3: satisfied and 4: very satisfied), these levels of satisfaction, in the present analysis represent the classes of the supervised learning predictive model.

D. Reliability of the Collected Data

The validation of the collected data is carried out by means of the Cronbach's alpha coefficient through the SPSS statistical software. Table II shows the reliability result, whose value is 0.962.

E. Data Processing Design

The methodology of the research process begins with the collection of data through the online questionnaire of the self-perception of engineering students about personal and social attitudes. The responses are stored in a database, the information of which is processed by the Open Data Base Connectivity (OBDC) driver and the Matlab R2021a software. In this way, the supervised learning process begins, through the Classification Learner technique, the best Machine Learning algorithm is identified, through the evaluation of performance metrics, after which the prediction is made, which will allow the analysis of the results. In Fig. 1, the scheme of the supervisory learning methodology used is shown.

TABLE II. CRONBACH'S ALPHA TEST

Reliability statistics	
Cronbach's alpha	***No. of elements***
0.962	8

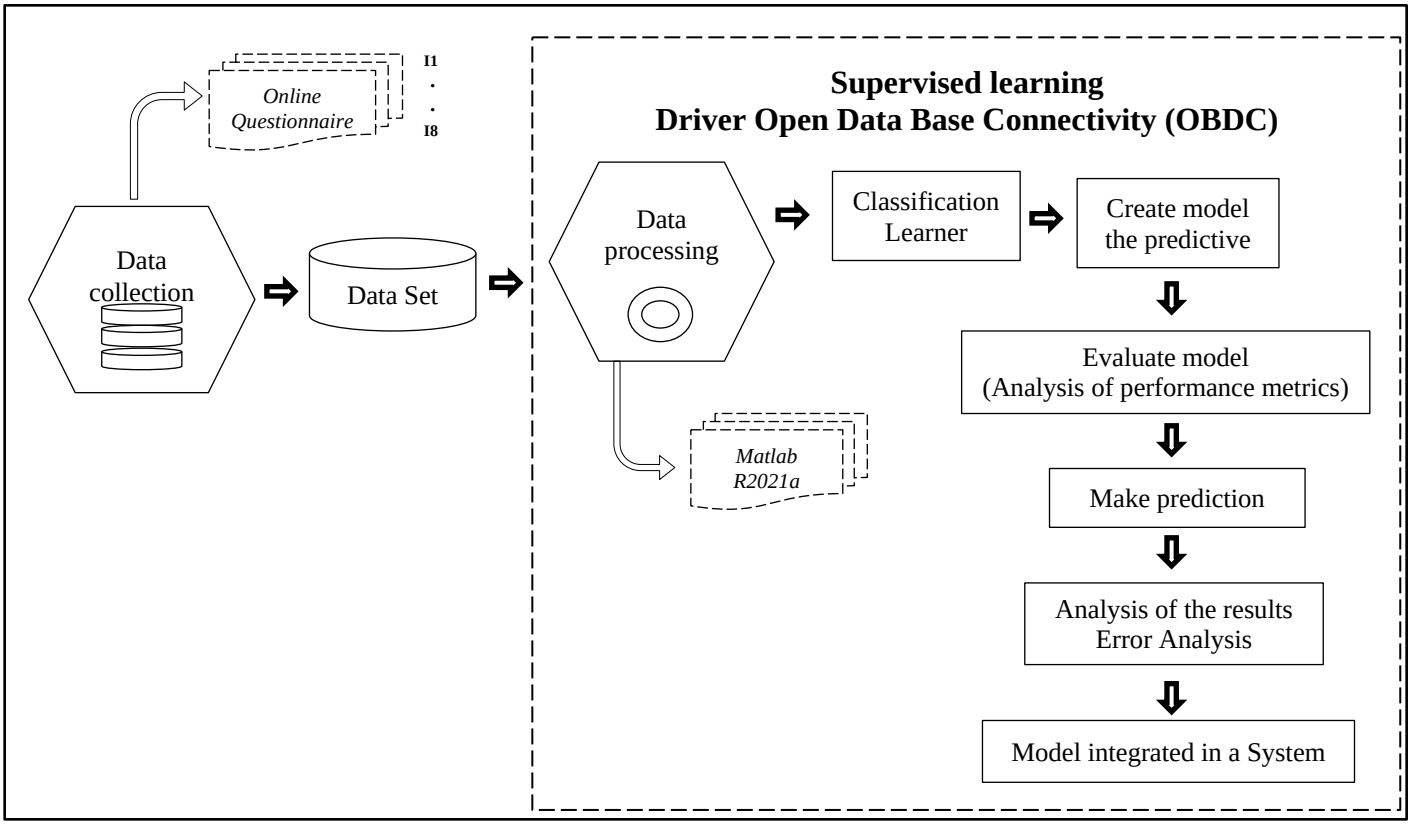


Fig. 1. Supervised Learning Methodology.

III. Result and Discussion

A. Determination of the Predictive Model

By means of the Matlab R2021a software, and through the Classification Learner technique of Machine Learning Toolbox 12.1, the best predictive model determined by the validation of the accuracy is identified, in statistical terms it is related to the bias of an estimate and is represented as the proportion of true results (true positives and true negatives) divided by the total number of cases examined (true positives, false positives, true negatives, false negatives).

In Fig. 2, the generated results are shown, which show that the Kernel algorithm: Logistic Regression Kernel, is the one that presents a better accuracy of 86.9% for the predictive system of the personal and social attitudes of engineering students This is followed by the algorithms Tree: Coarse Tree with an accuracy of 86.6%, SVM: Coarse Gaussian SVM with an accuracy of 86.4% and SVM: Linear SVM with an accuracy of 86.2%.

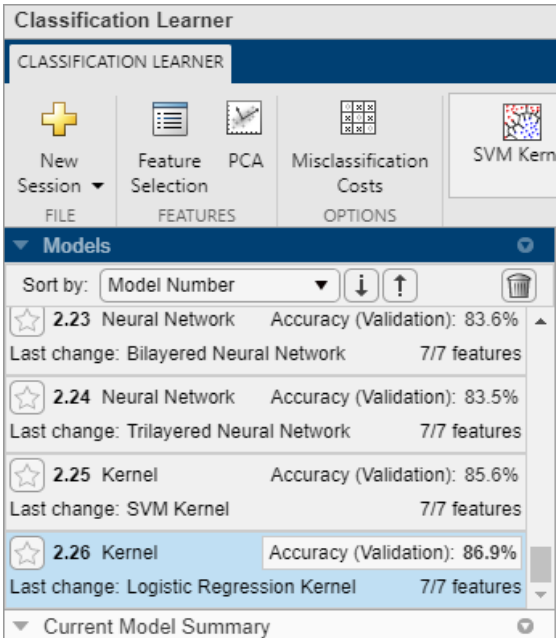


Fig. 2. Determination of the Classification Algorithm.

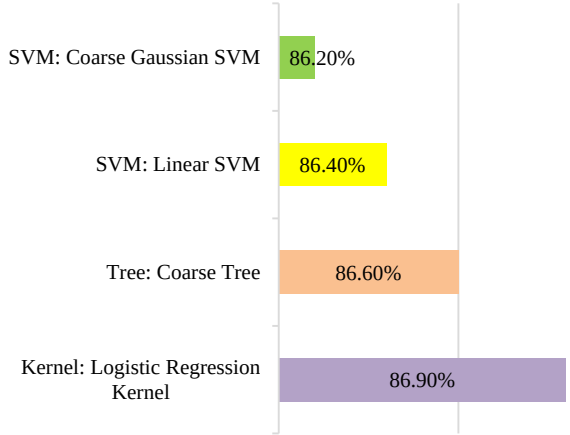


Fig. 3. Comparative Analysis of Accuracy Results.

Likewise, Fig. 3 shows the four algorithms with the highest accuracy, which identified the classification technique, where the Logistic Regression Kernel algorithm is observed as the best.

Being precision, the parameter that measures the percentage of cases that the model has hit, according to Fig. 3, it can be said that the predictive model through the Logistic Regression Kernel algorithm offers a total of 86.9% of the number of positive predictions which will be correct, that is, the value refers to how close the result of a measurement is to the true value.

B. Results of the Predictive Model Metrics

As part of the analysis of the performance metrics, we will visualize the confusion matrix of the Logistic Regression Kernel algorithm, the confusion matrix allows us to visualize the performance of a supervised learning algorithm and each column of the matrix will represent the number of predictions of each class (1: dissatisfied, 2: not very satisfied, 3: satisfied and 4: very satisfied), while each row represents the instances in the real class, in other words this analysis allows us to see what types of successes and errors our predictive system has .

In Fig. 4, the confusion matrix is shown, where the percentage of false negatives test (FNR) is displayed, also called the error rate, it is the probability that a true positive will miss it, the rate is also displayed of true positives (TPR), which measures the sensitivity metric, this metric comes to make the probability that a real positive result will be positive.

As can be seen in Fig. 4, of the 4 classes on which the predictive model acts, class 3 (satisfied) shows 89.9% sensitivity and 10.2% false negatives, this means that the predictive model has the ability to discriminate between a true positive (TP) from a false negative (FN), that is to say, there is an 89.9% capacity to be able to correctly detect satisfied students among dissatisfied students. On the other hand, class 2 (not very satisfied) shows a lower percentage of sensitivity equal to 64.4% with a percentage of false negatives of 35.6%, that is, it has the ability to detect only 35.6% of satisfied students among dissatisfied students.

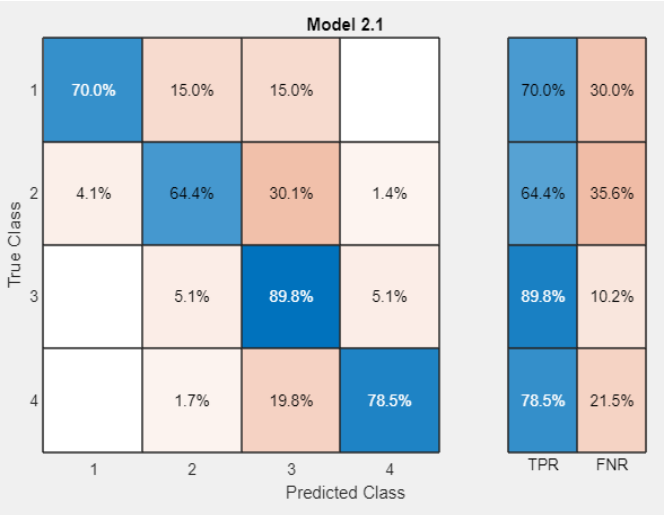


Fig. 4. Confusion Matrix based on TPR and FNR Rates.

In Fig. 5, the confusion matrix is shown, where the positive predictive value (PPV) and the false discovery rate (FDR) are displayed. It should be noted that the precision metric measures the quality of the machine learning model in classification tasks, it should be taken into account that the lower the dispersion value, the higher the precision of the model.

As can be seen in Fig. 5, the predictive model for class 3 (satisfied) shows the highest precision value, in this case it is 87.3% and a FDR percentage of 12.7%, that is, only 87.3% of students will be really satisfied with perceived personal and social attitudes, while 12.7% of examples will be wrong in the prediction.

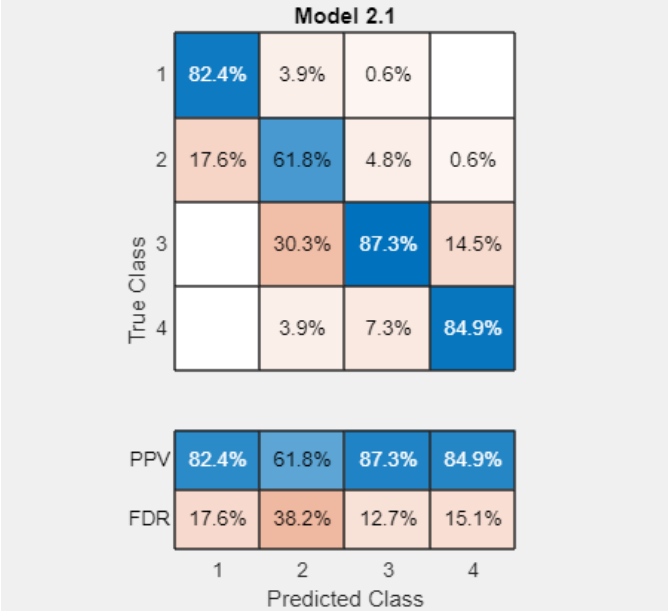


Fig. 5. Confusion Matrix based on PPV and FDR Rates.

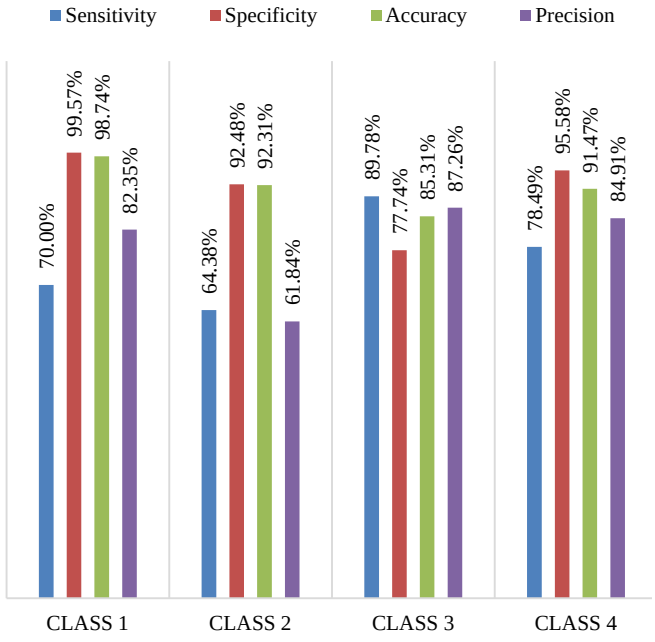


Fig. 6. Metrics of the Four Classes of the Predictive System.

In Fig. 6, the metrics of the predictive model of the four classes (1: dissatisfied, 2: not very satisfied, 3: satisfied and 4: very satisfied) are shown through the Logistic Regression Kernel algorithm, with this it can be said, in general, the Precision is 79.09%, the Sensitivity is 75.66%, the Specificity is 92.09% and the Accuracy is 91.96%.

Next, the analysis of the Receiver operating characteristic (ROC) curves of the Logistic Regression Kernel algorithm will be carried out, which constitutes a statistical method to determine the accuracy of the model, this test is carried out for three purposes, to determine the cut-off point of a continuous scale in the that the highest sensitivity and specificity is reached, evaluate the ability to differentiate satisfied and dissatisfied students, and compare the discriminative ability of two or more diagnoses that express their results as continuous scales. It is necessary to specify the more the value of the area on the curve (AUC), approaches 1, and the model will have a better performance and greater precision.

In that sense, in Fig. 7, the ROC graph for class 1 (dissatisfied) is shown.

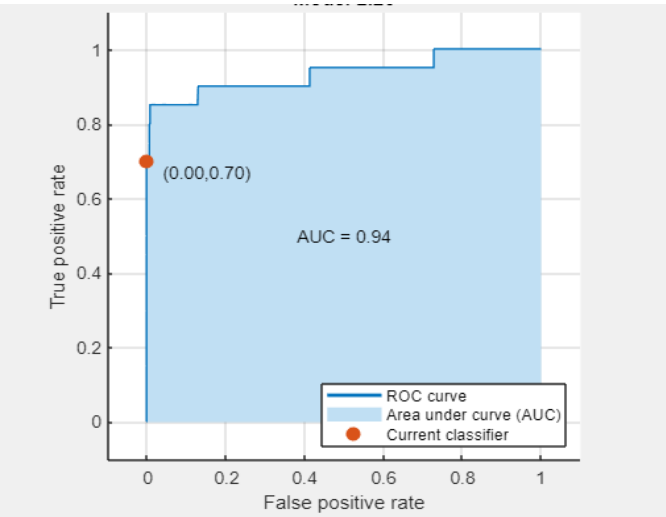


Fig. 7. ROC Charts for Class 1.

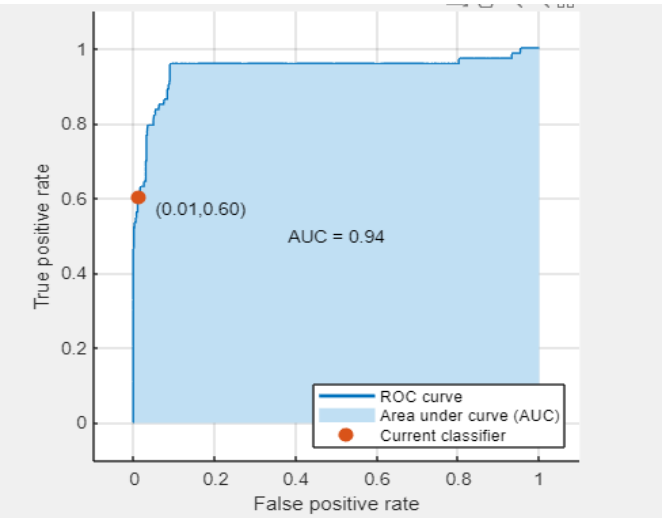


Fig. 8. ROC Charts for Class 2.

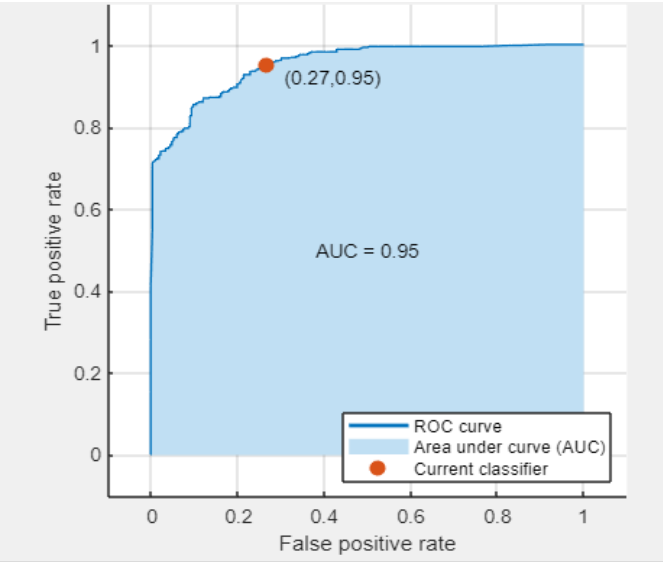


Fig. 9. ROC Charts for Class 3.

In Fig. 7, it is evident that for class 1, there is a sensitivity of 94%. In addition, the discrimination threshold is 0.70 for the true positive rate and 0.00 for the false positive rate, showing an optimal area value on the curve (AUC) of 0.93.

In Fig. 8, the ROC chart for class 2 (not very satisfied) is shown.

In Fig. 7, it is evident that for class 1, there is a sensitivity of 94%. In addition, the discrimination threshold is 0.70 for the true positive rate and 0.00 for the false positive rate, showing an optimal area value on the curve (AUC) of 0.93.

In Fig. 8, the ROC chart for class 2 (not very satisfied) is shown.

In Fig. 9, it is evident that for class 3, there is a sensitivity of 95%. In addition, the discrimination threshold is 0.95 for the true positive rate and 0.27 for the false positive rate, showing an optimal area value on the curve (AUC) of 0.95.

In that sense, in Fig. 10, the ROC graph for class 4 (very satisfied) is shown. For class 4, a sensitivity of 97% is evidenced. In addition, the discrimination threshold is 0.78 for the true positive rate and 0.03 for the false positive rate, showing an optimal area value on the curve (AUC) of 0.97.

Finally, it is shown in Fig. 11, the graph of parallel coordinates, used to plot multivariate data, this graph shows the relationship between the indicators of self-perception of engineering students on personal and social attitudes among the four classes of the predictive model (1: dissatisfied, 2: not very satisfied, 3: satisfied and 4: very satisfied).

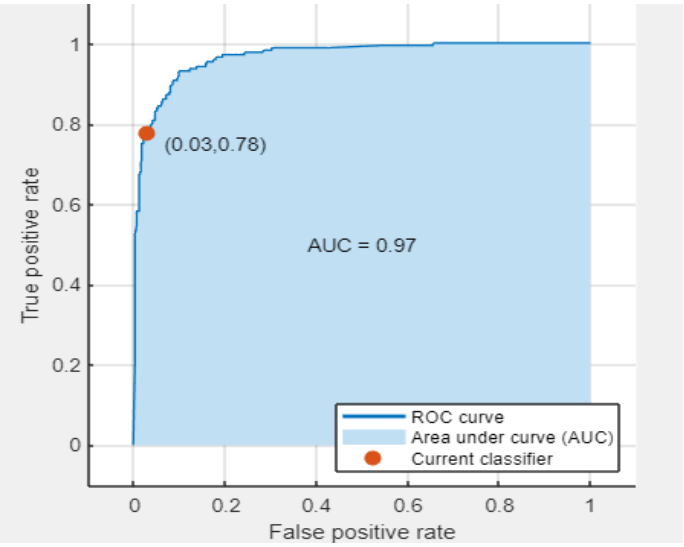


Fig. 10. ROC Charts for Class 4.

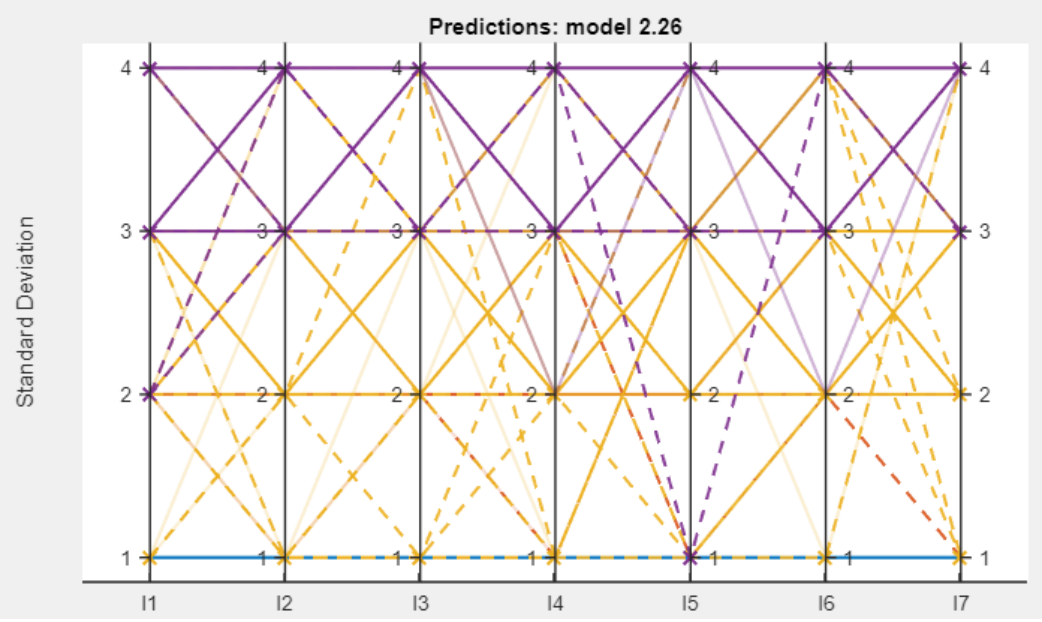


Fig. 11. Parallel Coordinates Graph.

It should be noted that the data are linked to 8 descriptors related to the levels of satisfaction of the self-perception of engineering students about personal and social attitudes, during the virtual teaching-learning process, said data are ordinal qualitative types, thus defining the four classes, likewise, as part of the determination of the predictive model using the Matlab R2021a software, the data collected was used based on the first 7 indicators, which are called predictors, and the indicator of the variable under analysis (which represents the satisfaction of the self-perception of engineering students on personal and social attitudes), is defined by the final indicator (I8), because it is the one that generally encompasses the variable under analysis, for this reason in Fig. 11 only displayed from indicator I1 to indicator I7.

C. Discussion

The results of the prediction model metrics through the Logistic Regression Kernel algorithm, in relation to satisfaction with personal and social attitudes, generally present, in its 4 classes, a precision of 79.09%, a Sensitivity of 75.66%, a Specificity of 92.09% and an Accuracy of 91.96%.

Also the ROC chart for class 4 (very satisfied) shows a sensitivity of 97%, a discrimination threshold is 0.78 for the rate of true positives and 0.03 for the rate of false positives, showing an optimal area value on the curve (AUC) of 0.97.

Regarding the optimal value of the performance metrics, it is sustained that they are optimal values, based on the research of [18] where it is stated that a model is made to predict the academic performance of incoming students through machine learning, the Results showed that it is feasible to predict performance, since the model has 69% accuracy. Similarly, in [26] it is indicated that the results show a precision of 82%, therefore, it can be pointed out that the predictive model will have optimal performance when implemented.

Regarding the ROC obtained of 97% in the present investigation, in [20], the author points out that his predictive model was good since its general precision was 75.42% and an area under the ROC curve of 0.805. This is based on the established theory that the closer this value is to 1, the better the model will perform and the more accurate it is. In the same way in [23], the author points out that an AUC of 60% of 91% or 99%, represents a better performance of the classifier algorithm, these results being favorable for the investigation.

In [29] it is pointed out that in relation to the accuracy of the model for the academic performance class, the value obtained was 62.45% lower than the value obtained in the research carried out by [24] who, in their prediction model, obtained the value 73%, very close to the result obtained by [25] which was 69%.

Comparing the results with the research carried out by [31] where a lower accuracy value of 80% was obtained in the predictive system for the training data and 76% for the validity data, this study presents higher performance results. In turn, in the investigation of [27], 82.87% accuracy was obtained using the decision tree algorithm, representing a lower value than the result of our investigation.

In the study carried out by [19] a data from 914 students was used to predict their final classification (passed or failed), with a predictive model of supervised learning and the Classification Learner technique and thus obtain a model to predict the results of the students, showing that with the network algorithm (Naive de Bayes) which showed optimal precision with 71.0%, compared to four analyzed techniques (neural networks, logistic regression, decision tree and Bayesian network) with a higher percentage for the class passed and minor for failed.

IV. Conclusion

With the results obtained, it is evident that it is possible to apply the supervised learning methodology uses Classification Learner techniques for the predictive system of the personal and social attitudes of the students, through the graph of parallel coordinates the association and / or relationship of the indicators of the variable under analysis with the four classes of the predictive model through the Logistic Regression Kernel algorithm. The results of the prediction model metrics in relation to satisfaction with personal and social attitudes are concluded, it generally presents an optimal performance of its validation metrics in its 4 classes, with a precision of 79.09%, a Sensitivity of 75.66%, a Specificity of 92.09% and an Accuracy of 91.96%.

It is recommended to extend the line of research to others with other indicators related to student satisfaction with the educational service, because not only does it allow the higher institution to have a database or reference through a rapid and reliable classification technique for take preventive and corrective actions, to improve the quality of education, but also the predictive system influences the reduction of dropout rates and improve the academic performance of students.

Acknowledgment

Thanks to the researchers who have contributed their knowledge in the development of this paper.

References

[1] F. E. Ceballos, J. E. Rojas, L. G. Cuba, L. P. Medina and A. R. Velazco, “Analysis of the quality of services in university centers”, *University, Science and Technology*, vol. 25, no. 108, pp. 23-29, 2021. DOI: https://doi.org/10.47460/uct.v25i108.427.

[2] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” *Telematics and Informatics*, vol. 37, pp. 13-49, 2019. DOI: https://doi.org/10.1016/j.tele.2019.01.007.

[3] D. Vlachopoulos, “Quality Teaching in Online Higher Education: The Perspectives of 250 Online Tutors on Technology and Pedagogy,” *International Journal of Emerging Technologies in Learning (IJET)*, vol. 16, no. 6, pp. 40-56, 2021. DOI: https://doi.org/10.3991/ijet.v16i06.20173.

[4] B. Bahati, U. Fors, P. Hansen, J. Nouri and E. Mukama, “Measuring Learner Satisfaction with Formative e-Assessment Strategies,” *International Journal of Emerging Technologies in Learning (IJET)*, vol. 14, no. 7, pp. 246-247, 2019. DOI: https://doi.org/10.3991/ijet.v14i07.9120.

[5] B. Semprún, and K. Ferrer, “Student satisfaction in a Biochemistry course: an evaluation after applying neurodidactic strategies”, *San Gregorio Journal*, vol. 1, no. 38, pp. 1-14, 2020. DOI: https://dx.doi.org/rsan.v1i38.1241.

[6] I. Amoako and K. Asamoah-Gyimah, “Indicators of students' satisfaction of quality education services in some selected universities in

Ghana," *South African Journal of Higher Education*, vol. 35, no. 4, pp. 61-72, 2020. DOI: 10.20853/34-5-4252.

[7] T. Chen, L. Peng, X. Yin, J. Rong, J. Yang, and G. Cong, "Analysis of user satisfaction with online education platforms in China during the COVID-19 pandemic," *Healthc.*, vol. 8, no. 3, p. 200, 2020. DOI: 10.3390/healthcare8030200.

[8] E. Alvarado-Lagunas, J. Luyando-Cuevas and E. Piccaso-Palencia, "Perception of students towards the quality of private universities in Monterrey," *Ibero-American Journal of Higher Education*, vol. 6, no. 17, pp. 58-76, 2015. DOI: 10.1016/j.rides.2015.10.003.

[9] K. Mukhtar, K. Javed, M. Arooj, and A Sethi, "Advantages, limitations and recommendations for online learning during COVID-19 pandemic era", Pak. J. Med. Sci., vol. 36, no. COVID19-S4, pp. S27-S31, 2020, doi: 10.12669/pjms.36.COVID19-S4.2785.

[10] W. Niebles-Núñez and M. Cabarcas-Velásquez, "Social Responsibility: Training Element for University Students," *Latin American Journal of Educational Studies*, vol. 14, no. 1, pp. 257-268, 2018. DOI: https://doi. org/10.17151/rlee.2018.14.1.6.

[11] A. Bernasconi and E. Rodriguez-Ponce, "Exploratory Analysis of Perceptions on Leadership Styles, Academic Climate and the Quality of Undergraduate Education," *Form. Univ.*, vol. 11, no. 3, pp. 29-40, 2018, DOI: http://dx.doi.org/10.4067/S0718-50062018000300029.

[12] P. Ramkissoon, L. J. Belle and T. Bhurosy, "Perceptions and experiences of students on the use of interactive online learning technologies in Mauritius," *International Journal of Evaluation and Research in Education*, vol. 9, no. 4, pp. 833–839, 2020. DOI: http://doi.org/10.11591/ijere.v9i4.20692.

[13] A. Ali and F. Mohammed, "Measuring quality of E-Learning and Desaire2Learn in the College of Science and Humanities at Alghat, Majma-ah University," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 1, pp. 61-72, 2020. DOI: 10.14569/IJACSA.2018.090170.

[14] H. Hafiza and R. Ibrahim, "Distance Education during COVID-19 Pandemic: The Perceptions and Preference of University Students in Malaysia Towards Online Learning," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 4, pp. 118-126, 2021. DOI: 10.14569/IJACSA.2021.0120416.

[15] C. Coman, L.G. Țîru, L. Meseșan-Schmitz, C. Stanciu, and M. C. Bularca, "Online teaching and learning in higher education during the coro-navirus pandemic: Students' perspective," *Sustainability*, vol. 12, no. 24, pp. 1-22, 2020. DOI: 10.3390/su122410367.

[16] F. Phang, et al., "Integrating Drone Technology in Service Learning for Engineering Students," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 16, no. 15, pp. 78-96, 2021. DOI: https://doi.org/10.3991/ijet.v16i15.23673.

[17] A. Ramos, M. Aldude, J. Estrada, V. Señas and L. Andrade-Arenas, "Analysis of the use of technological tools in university higher education using the soft systems methodology," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 7, pp. 412-420, 2020, DOI: 10.14569/IJACSA.2020.0110754.

[18] H. Sadiq, A. Dahan, Neama, F. M. Ba-Alwi and N. Ribata, "Educational data mining and analysis of students' academic performance using weka," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447-459, 2018. DOI: https://doi.org/10.11591/ijeecs.v9.i2.pp447-459.

[19] C. H. Menacho, "Prediction of academic performance applying data mining techniques," *Learning and Individual Scientific Annals*, vol. 78, no. 1, pp. 26-33, 2017. DOI: http://dx.doi.org/10.21704/ac.v78i1.811.

[20] R. S. Baker and K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-16. DOI: https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.12 40314.

[21] E. B. Costa, B. Fonseca, M. A. Santana, F. F. De Araújo and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247–256, 2017. DOI: https://doi.org/10.1016/J.CHB.2017.01.047.

[22] B. Bakhshinategh, O. R. Zaiane, S. ElAtia and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Education and Information Technologies*, vol. 23, no. 1, pp. 537-553. DOI: https://doi.org/10.1007/s10639-017- 9616-z.

[23] D. Moonsamy, N. Naicker, T. T. Adeliyi and R. E. Ogunsakin, "A Meta-analysis of Educational Data Mining for Predicting Students Performance in Programming," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 2, pp. 97-104, 2021. DOI: 10.14569/IJACSA.2021.0120213.

[24] O. D. Castrillón, W. Sarache, William and S. Ruiz-Herrera, "Prediction of academic performance using artificial intelligence techniques," *University education*, vol. 13, no. 1, pp. 93-102, 2020. DOI: https://doi.org/10.4067/S0718-50062020000100093.

[25] D. Buenaño-Fernández, D. Gil and S. Luján-Mora, "Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study," *Sustainability*, vol. 11, no. 10, p. 2833. DOI: https://doi.org/10.3390/su11102833.

[26] R. Katarya, J. Gaba, A. Garg, Aryan and V. Verma, "A review on machine learning based student's academic performance prediction systems," *2021International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 254-259, Coimbatore, India: IEEE, 2021. DOI: https://doi.org/10.1109/ICAIS50930.2021.9395767.

[27] P. Sokkhey and T. Okazaki, "Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 8, pp. 492-502, 2020. DOI: 10.14569/IJACSA.2020.0110862.

[28] F. Makombe and M. Lall, "A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 9, pp. 415-419, 2020. DOI: 10.14569/IJACSA.2020.0110949.

[29] B. Díaz-Landa, R. Meleán-Romero and W. Marín-Rodriguez, "Academic performance of students in Higher Education: predictions of influencing factors from trees decision," *Redalyc*, vol. 23, no. 3, pp. 616-635, 2021. DOI: https://doi.org/10.36390/telos233.08.

[30] M. Bravo, S. Salvo and C. Muñoz, "Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests," *Studies in Educational Evaluation*, vol. 44, pp. 50-59, 2015. DOI: https://doi.org/10.1016/j.stueduc.2015.01.002.

[31] M. Imran, S. Latif, D. Mehmood and M. S. Shah, "Student Academic Performance Prediction using Supervised Learning Techniques," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 14, no. 14, pp. 92–104, 2019. DOI: https://doi.org/https://doi.org/10.3991/ijet.v14i14.10310.

Workflow Scheduling and Offloading for Service-based Applications in Hybrid Fog-Cloud Computing

Saleh M. Altowaijri
Department of Information Systems, Faculty of Computing and Information Technology
Northern Border University, Rafha 91911, Kingdom of Saudi Arabia

***Abstract*—Fog and edge computing has emerged as an important paradigm to address many challenges related to time-sensitive and real-time applications, high network loads, user privacy, security, and others. While these developments offer huge potential, many efforts are needed to study and design applications and systems for these emerging computing paradigms. This paper provides a detailed study of workflow scheduling and offloading of service-based applications. We develop different models of cloud, fog and edge systems and study the scheduling of workflows (such as scientific and machine learning workflows) using a range of system sizes and application intensities. Firstly, we develop several Markov models of cloud, fog, and edge systems and compute the steady-state probabilities for system utilization and stability. Secondly, using steady-state probabilities, we define a range of system metrics to study the performance of workflow scheduling and offloading including, network load, response delay, energy consumption, and energy costs. An extensive investigation of application intensities and cloud, fog, and edge system sizes reveals that significant benefits can be accrued from the use of fog and edge computing in terms of low network loads, response times, energy consumption and costs.**

Keywords*—Workflow scheduling; workflow offloading; cloud computing; fog computing; edge computing; scientific workflows

I. Introduction

In recent years, new computing paradigms, named fog computing [1]–[5] and edge computing [6]–[11] have emerged as an extension of cloud architecture to the edge of the network to support the computational demands of real-time, latency-sensitive, and location-aware service-based applications (SBA) of largely distributed Internet-of-Things (IoT) devices/sensors. Fog and edge computing are considered among the most important archetypes in the current world. While some communities differ in the precise definitions of and differences between fog and edge computing or nodes, we prefer the following definitions. Edge computing refers to the computing at the edge of the networks, near the device and IoT layers. Fog computing refers to the computing at the intermediate layers between cloud data centers and IoT devices (many works have considered such definitions, see [3], for example). Edge and Fog layers have been proposed to bridge the gap between the cloud and IoT devices by enabling data management, computing, networking, storage, and application services at the intermediate layers and edge of the network while offering the possibility to interact with the cloud. Many applications have been proposed to benefit from fog and edge computing such as smart districts [9], SMS (Short Message Service) spam detection [12], networked healthcare [3], smart societies [2], QoS management in networks [8], Smart airport [9], Distributed Artificial Intelligence (AI) as-a-service (DAIaaS) [9], and many others applications [13]–[18]. However, the development and management of fog-based and edge-based systems for SBA face many challenges that need to be tackled. These include investigating and designing applications and systems for these emerging computing paradigms. One of the core challenging issues is workflow scheduling and offloading in such a dynamic, geo-distributed, heterogeneous environment where the set of computing nodes contains edge nodes, fog nodes, and cloud datacenters such as discussed in many works in the literature [19]–[23].

Further research is needed, for instance, for investigating dynamic scheduling of multiple workflows executions, i.e., invocations of multiple sets of linked elementary IT-enabled services, in hybrid edge-fog-cloud computing environments while ensuring the individual Quality-of-Service (QoS) requirements of all the workflows and their services and reducing services latency, energy consumption, and costs.

This paper provides a detailed study of workflow scheduling and offloading of service-based applications. We abstracted high-level challenges and requirements of cloud, fog and edge systems and developed different models of cloud, fog and edge systems, and study the scheduling of workflows (such as scientific and machine learning workflows) using a range of system sizes and applications intensities. Firstly, we develop several Markov models of cloud, fog, and edge systems and compute the steady-state probabilities for system utilization and stability. Secondly, using steady-state probabilities, we define a range of system metrics to study the performance of workflow scheduling and offloading including, network load, response delay, energy consumption, and energy costs. An extensive investigation of application intensities and cloud, fog, and edge system sizes reveals that significant benefits can be accrued from the use of fog and edge computing in terms of low network loads, response times, energy consumption, and costs.

The proposed workflow scheduling and offloading models can be utilized in practice to study a range of applications and derive several benefits. Firstly, different well-known standardized workflow can be plugged in our proposed workflow scheduling models to study their various performance behaviors including network load, average response delay, energy consumption, and energy cost, and this can be done for a range of cloud-only, cloud-fog, and cloud-

fog-edge systems. Some examples of standardized workflows include Montage workflow, SIPHIT workflow, epigenomics workflow, LIGO workflow, Cyber-Shake workflow, and more; see [23], for explanations and use cases of these workflows, and [19]–[22] for additional examples for practical utilization of our work. We elaborate this further in the methodology, results, and the discussion sections.

The rest of the paper is organized as follows. Section II reviews related work. Section III details the methodology and design. Section IV provides an analysis of the results. Section V discusses the practical utilization of the proposed models and concludes the paper. Section VI provides future research directions.

II. Literature Review

The focus of this paper is on combining the use of edge, fog and cloud computing for executing service-based applications. Fog computing has been attracting a lot of attentions in the last few years from researchers all over the world to bring out its potential. It has been seen as a complement of cloud computing to allow satisfying the increasingly sophisticated applications demanded by users that combine the use of time-sensitive services and intensive-processing services, such as big data analysis which can be performed only in the cloud. The coupling of fog and cloud computing requires providing scheduling and offloading mechanisms that allows to manage the execution of multiple workflows of interconnected services in fog and cloud resources.

The problem of scheduling and offloading in fog computing environments has been an active research topic for the last few years. Many researchers have been extensively focusing on providing solutions to this problem [24]–[32]. However, the research on fog computing, in general, and on workflow/task scheduling and offloading, in particular, is still in its early stages and the problem is not completely solved and there is still a lot of challenges that need to be addressed [33].

In [34], Zeng et al. tackled the problem of minimizing the maximum task completion time in Fog computing supported software-defined embedded system (FC-SDES) by jointly considering task scheduling and task image placement. The authors considered a scenario where tasks (requests) can be processed either on the client node or a fog (edge) node and task images can be saved on storage servers. Based on that scenario, they formulated their optimization problem as a mixed-integer non-linear programming (MINLP) problem. Then, in order to tackle its computational complexity, the authors proposed a heuristic algorithm for task completion time minimization based on the concept of “partition and join”. The main consideration in the proposed algorithm is that, by balancing the load between client nodes and fog nodes, the overall computation and transmission latency of all requests, therefore, their completion time, can be significantly minimized.

Similarly, Chen et al. [35], formulated the task offloading problem in ultra-dense network as a mixed-integer non-linear programming problem. Their aim, in this work, was to minimize the delay while saving the battery lifetime of user’s device. To do so, the optimization problem has been divided into two sub-problems, i.e., task placement and resource allocation sub-problems. Based on the solution of the two sub-problems, the authors proposed an offloading scheme which considers the battery lifetime of user’s device while reducing the task duration.

In [36], Huang et al. focused on providing a solution to the problem of computational offloading for multimedia workflows in mobile cloud computing. They proposed an energy-efficient offloading method using Differential Evolution (DE) algorithm to optimize the energy consumption of the mobile devices with time constraints.

Targeting the problem of task scheduling in smart factory, Wan et al. [37], introduced a method for energy-aware load balancing and scheduling (ELBS) based on Fog computing. They first formulated a load balancing optimization function by taking into account the energy consumption of the equipment in the smart factory. Then, they introduced a multi-agent system for achieve the dynamic scheduling of equipment workload with the task scheduling mechanism.

Considering a scenario where both edge/fog and cloud computing are used to serve mobile users, Zhao et al. [31], proposed to maximize the probability of tasks satisfying the delay requirement by jointly scheduling them either to the edge/fog network or to the cloud and allocating computational resources in the edge/fog network. So, they proposed to offload tasks with stringent delay bounds to resources in the edge level while the ones with loose delay bounds to resources in the cloud level. The proposed solution has been introduced to allow users with different delay requirement to be simultaneously served.

In [38], the authors presented a ranking-based method for task scheduling in fog-cloud computing networks. The aim of the proposal is to schedule user’s requests based on their different preferences and fog nodes’ constraints. To do so, the authors proposed to use linguistic quantifiers and fuzzy quantified propositions to rank fog nodes from the most to the least satisfactory one based on their requirements, then, the one that satisfies more user task preferences will be selected as a destination fog node.

Another work for task scheduling in hybrid fog-cloud computing has been proposed in [39]. In their work, Aburukba et al. modeled the problem of scheduling IoT service requests as an optimization problem using integer linear programming to minimize latency. They proposed a heuristic optimization approach in order to find feasible solutions with a good quality in a reasonable computational time. The genetic algorithm (GA) has been chosen and customized to schedule the IoT service requests to achieve the objective of minimizing the overall latency.

Even though there is many research works for scheduling and offloading in fog computing and hybrid fog-cloud computing, most of them fail to meet the scalability and mobility of nodes criterion. Also, they consider the scheduling of a single task which is not applicable for workflows composed of a set of linked tasks. The dependency between

tasks in workflows adds more challenge to the scheduling and offloading problem. More importantly, many efforts are needed to develop high-level understanding of cloud-fog-edge systems.

III. Methodology and Design

A. Workflow Scheduling Cloud-Fog-Edge Model

Consider a workflow scheduling system that is programmed to manage its capacity periodically -- such as monthly, weekly, daily, or hourly -- by scrutinizing the quantitative variations in the workload. A possible method that can be used for planning is building Markov models and solving these models for their steady-state probabilities.

Let us represent the demand of a workflow scheduling system in terms of the aggregate computational nodes by λ_A, where the subscript 'A' represents the 'aggregate' demand per hour. The demand λ represents the inter-arrival times that are exponentially distributed. The aggregate demand includes the demand for cloud, fog, and edge servers that are represented by λ_C, λ_F, and λ_E, respectively. We can write the aggregate demand in a mathematical form as in the following equation.

$$\lambda_A = \lambda_C + \lambda_F + \lambda_E. \quad (1)$$

Now consider that the aggregate hourly capacity of the collective cloud, fog, and edge system is μ_A, where, as for the arrival rate λ, the subscript 'A' represents the 'aggregate' hourly capacity in terms of the number of nodes. These nodes can be the physical nodes in the system or virtual machines. Similar to Equation (1), the following equation gives the breakdown of the aggregate capacity, which is the sum of the cloud, fog, and edge capacities, respectively.

$$\mu_A = \mu_C + \mu_F + \mu_E. \quad (2)$$

The total number of physical or virtual nodes in the system are represented by N_A, which is the sum of the total number of cloud, fog, and edge nodes, represented by N_C, N_F, and N_E, as formulated in Equation (3) below. Note that the capacities defined in Equation (2) are the hourly service capacities of the system while Equation (3) defines the number of cloud, fog, and edge nodes in the system.

$$N_A = N_C + N_F + N_E. \quad (3)$$

Note that "capacity" in this paper implies to be the server capacity in terms of the physical nodes or virtual machines the cloud, fog, and edge are able to provide in terms of their hourly rates. Note also that the three arrival rates or demands and capacities in the two equations given above can assume any reasonable values and their quantities do not affect our model. Indeed, not only that the model is independent of the values λ and μ can assume, the number of sub-arrival rates and capacities can also be extended to any number of clouds, fogs and edges. That is, using the model described above, we can model any number of clouds, fogs and edges by embedding in Equations (1) and (2) their individual arrival rates and server node capacities.

Fig. 1 depicts the CTMC (Continuous Time Markov Chain) transition diagram of our proposed cloud-fog-edge workflow scheduling model. There are three parts of the transition diagram, one part each for cloud, fog, and edge. The symbols used in the figure have already been defined in the earlier paragraphs and equations. The cloud layer model is depicted in the top row, as is evident by the use of "C" subscript in all the variables and parameters (arrival and departure rates, and node capacities), followed by the fog layer (use of the subscript "F") and edge layer (use of the subscript "E") in the second and third rows, respectively.

Inside the cloud layer, we have the system moving from the zero or idle state with no task in the system to be executed to one task, two tasks, until "c" tasks, where "c" could be any state between zero and N_C. Once the maximum number of tasks allocated to the cloud(s) have reached, the fog can start receiving tasks, moving from state one, to two, to "f" where "f" could be any state between one and N_F. Once the maximum number of tasks allocated to the fog(s) have reached, the edge can start receiving tasks, moving from state one, to two, to "e" where "e" could be any state between one and N_E. Note that any reasonable values can be assigned to the quantities in Equations (1), (2), and (3), with the exception that arrival rate cannot exceed the capacity otherwise the system will not be stable. Note also that the system can be equally modelled as first receiving the tasks in the edge layer followed by the fog and cloud. Similarly, we can conveniently place another set of parameters -- let us call them n_C, n_f, and n_e, to replace N_C, N_F, N_E, respectively – such that they can be any number between zero and N_x ($x \in \{C, F, E\}$. This would enable the model to allocate any maximum number of nodes in cloud, fog, and edge layers up to the maximum capacities of the three layers. That is, we can model such that any or all of the three layers do not have to work to their full capacities to avoid instability and provide higher reliability. Finally, note that the arrival and departure rates are dependent on the specific state the system is in but in the figure (Fig. 1) these are shown the same for simplicity (e.g., λ_C and μ_C).

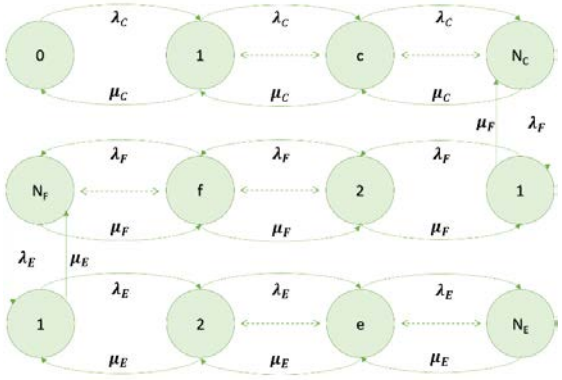


Fig. 1. The Cloud-Fog-Edge Workflow Scheduling Model – CTMC Transition Diagram.

$$\begin{pmatrix} -\lambda_C & \lambda_C & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu & -(\mu_C+\lambda_C) & \lambda_C & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 2\mu_C & -(2\mu_C+\lambda_C) & \lambda_C & 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 3\mu_C & -(3\mu_C+\lambda_C) & \lambda_C & 0 & \cdots & \cdots \\ \vdots & \vdots & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & N_C\mu_C & -(N_C\mu_C+\lambda_F) & \lambda_F & \cdots & \cdots \\ \vdots & \vdots & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & N_F\mu_F & -(N_F\mu_F+\lambda_E) & \lambda_E & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & N_E\mu_E & -(N_E\mu_E) \end{pmatrix}$$

Fig. 2. The Cloud-Fog-Edge Workflow Scheduling Model: CTMC Generator Matrix Q.

Fig. 2 depicts the CTMC generator matrix of our proposed cloud-fog-edge workflow scheduling model. The model itself is depicted in Fig. 1 and has been explained. This generator matrix, represented by A, can be used to calculate the steady-state probabilities vector, x, using the following equation.

$$Ax = 0. \tag{4}$$

The vector x is a probability vector and therefore the sum of the vector is normalised to the value of 1. The details about the numerical solution of Markov chains can be found in the nominal book [40], or our earlier works such as [41], and other works such as [42].

We define a range of cloud, fog and edge system sizes in terms of the computational nodes and task arrival and departure rates and solve them for the steady-state probabilities. We use these steady-state probabilities to compute the system utilization and stability as formulated in Equations (5) and (6). The system utilisation (U) is defined as the maximum relative state that has the highest probability for the system to be in, in terms of the number of computational nodes.

$$U = \frac{i}{N} \mid x[i] = \max(x). \tag{5}$$

The relative state in the equation above is computed by dividing the state number that has the highest probability in the vector (indicating that the system will be in this state with the highest probability) by the total capacity of the system (the hourly capacity in terms of the number of nodes); higher this number, higher will be the utilisation of the system in terms of the number of busy computational nodes. A higher utilisation is desired to make the best use of the available computational resources. However, a higher utilisation could reduce the stability of the system.

The system stability defines inverse of the maximum relative state that has the highest probability and can be computed by dividing 1 by the system utilization.

$$S = \frac{1}{U} \tag{6}$$

The average network workload in bytes per second is represented by NWL and its computation is formulated by the following equation.

$$\text{NWL} = \frac{\psi * \sum_{i=1}^{\kappa} (\eta_i * t_i)}{\kappa}, \qquad \eta_i \, \epsilon \, \eta, t_i \epsilon \, \{\kappa\}. \tag{7}$$

In the equation above, η_i is the network load of the task number i, given in bytes, t_i is the time the task number i takes to be transferred over the network, and η is the set of all tasks in a given state. There are κ tasks in a given state, and $\{\kappa\}$ is the set of all κ time durations required to complete all tasks. In our experiments, we have used a fixed size of 5MB for the network load for all tasks. The times for each task are also fixed according to the latency of different networks (cloud-fog latency is 100ms, fog-edge latency is 2ms, and edge latency is 0.1ms). Not that this is not a limitation of our approach; a distribution of task sizes and task network transfer time can easily be used in these equations. Finally, ψ is a factor that depends on the state the system is in and this is computed using the following equation.

$$\psi = 1 + \omega * U^{\sigma}, \quad \forall \psi, \psi \leq 1 + \omega. \tag{8}$$

The equation stats that ψ can be computed using system utilization U, ω, and σ, but its value cannot exceed $1 + \omega$, means that the value of U^{σ} cannot exceed 1. The parameter ω is a regulation weight given to the network workload calculation that regulates the factor ψ. We set it to 1.0. A lower value for this parameter will have a lower effect on the network load computation and vice versa. This can be set by the user based on their knowledge of the system or focus of the study. The parameter σ is set to 5 in our calculations. It is a dampening factor over utilization so that the effect of utilization is balanced over the various operational states of the workflow scheduling system. A higher value of the dampening fact σ will create higher dampening implying that the values of ψ will increase slowly for the earlier states towards the higher-numbered states (see Fig. 1 and Fig. 2).

The average response delay (RD) of the system can be computed by the following equation. The parameter ψ is the same factor that depends on the state the system is in and this was computed using Equation (8). The variable d_i is the delay of job "i" (the time it takes to complete the job) and there are τ jobs in the system, each with its own delay.

$$\text{RD} = \frac{\psi * \sum_{i=1}^{\tau} d_i}{\tau}, \qquad d_i \epsilon \, \{\tau\}. \tag{9}$$

The network energy consumption per hour (NE) is calculated by the following equation. The network energy consumption depends on the network load, NWL, and the estimated energy ζ. Several studies have reported the network energy consumption. We use the values reported in [9], [43], which is 0.54 kWh/GB. NWL has already been computed earlier. Since it was computed in MBps, we have added in the equation a denominator of 1000 to convert the network load into GBps. The value is multiplied by Ƭ which is the time factor, in this case it is 3600 (the number of seconds in an hour).

$$\text{NE} = \frac{Ƭ * \zeta * NWL}{1000}. \tag{10}$$

Finally, we compute the network energy cost (EC) per hour as given in the following equation. The cost depends on the network energy consumption, NE, calculated earlier, and the unit price of energy (γ), which we have taken from [44] as GBP 0.174 per kilowatt-hour (kWh).

$$\text{EC} = \gamma * NE. \tag{11}$$

IV. Results and Analysis

A. Workflow Scheduling System (Cloud)

We use the CTMC model described in Fig. 1 and Fig. 2 and model a cloud-only workflow scheduling system. We set the number of cloud nodes to 16,000, and the number of fog and edge nodes to zero each. This gives according to Equation (3), $\boldsymbol{N_A} = 16{,}000 + 0 + 0 = 16{,}000$. We study different CTMC system with varying aggregate arrival rates, λ_A, beginning from one task per hour ($\lambda_A = 1$) to 500 tasks per hour ($\lambda_A = 500$), up to 16,000 tasks per hour ($\lambda_A = 16{,}000$). The service rate or departure rate or hourly capacity of this system is kept at constant, which is 16,000 tasks per hour

($\mu_A = 16{,}000$). These settings result in 33 different CTMC models that we solve using iterative methods and compute the steady-state probabilities for each model. These are plotted in Fig. 3. The figures shows that the x-axis that plots the number of states varies between 0 and 16,000, and the y-axis provides probability values ranging from 1E-19 to 1.0 on a logarithmic base 10 scale. The system is idle in the zeroth state. Initially, with lower arrival rates, the probabilities of the lower-numbered states are higher compared to the higher-numbered states, and in these cases the maximum probabilities fall to a certain low, near-zero values (manifested in the vertically dropping lines before the state number 16,000). As we move towards higher arrival rates, the probabilities for the lower-numbered states start decreasing and the probabilities for the higher-numbered states start increasing. This trend continues until the probabilities rise towards the high-numbered states and do not fall vertically even after reaching the states nearer the state number 16,000. This shows that these cloud scheduling systems will be operating with high computational node utilization but with low stability and the risk for the system to drop the tasks off the system or its waiting queues.

Fig. 4 plots the utilization and stability of the 33 cloud workflow systems that we have described in the previous paragraph. Note that the utilization (y-axis, blue line) rises with the increasing arrival rates (x-axis), while the stability of the system (orange line) decreases with the increasing arrival rates. This is an expected behaviour from such systems.

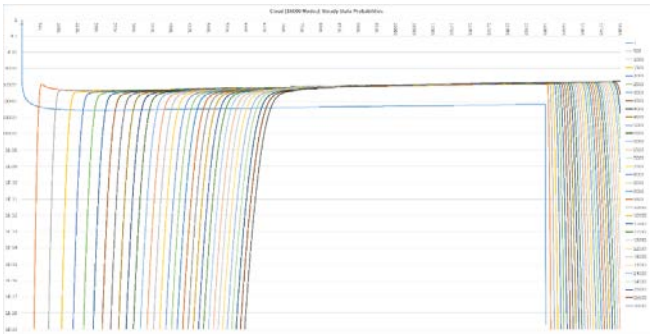


Fig. 3. Steady State Probabilities: Cloud Only System (16000 Nodes).

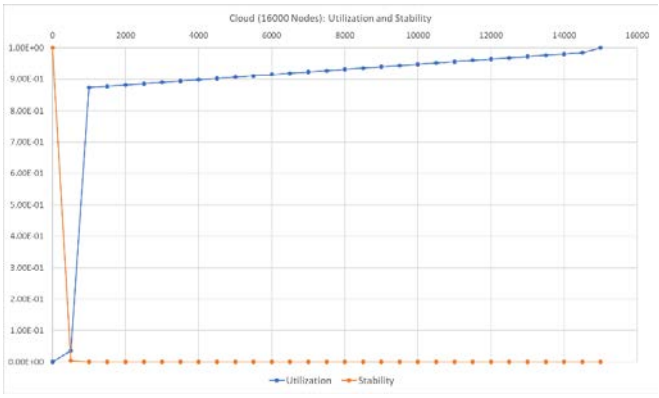


Fig. 4. Cloud (16000 Nodes): Utilization and Stability.

B. Workflow Scheduling System (Cloud-Fog-Edge)

We now model cloud-fog-edge systems where the workflows are scheduled to all three layers. Fig. 5 plots the steady-state probabilities of 23 workflow scheduling systems with different configurations. Some execute on cloud only (the top nine systems represented by Cloud-1, Cloud-1000, … Cloud-8000). While others run on fog (Fog-333 to Fog-333) and edge (Edge-166 to Edg-3166) layers. The numbers alongside Cloud-, Fog-, and Edge- represent the arrival rates for those nodes types. The strange numbering is used to avoid lines coming on top of each other and causing difficulty in reading and differentiating the plots. Note that since the fog and edge layers add to the capacity of the cloud, the probabilities for the higher-numbered states near the state number 16,000 are zero, indicating that those systems will not be unstable.

C. Network Load

The network workload (NWL) computations were explained earlier in Section III along with its Equation (7). In this section, we will study the network workload related performance of various cloud, cloud-fog, and cloud-fog-edge systems.

Fig. 6 depicts the network workload in GBps for a cloud-only workflow scheduling system containing 16,000 nodes. There are a total of 31 different systems that have been modelled and their network load has been computed. These 31 systems relate to different workloads on the systems in terms of the tasks being received by the system, beginning from 1 task, to 500 tasks, up to 15000 tasks per hour. The capacity in all of these 31 systems has been kept constant. The minimum network load is for the system with one task; it is actually 0.51 GBps but is rounded off to 1GBps in the figure. Note that the network load consistently rises to reach 15,315 GBps for the busiest workflow scheduling system. Note that the increase in the network load is due to the equal load of each job as have been explained in Section III. However, this increase can be varied by using a distribution of network loads related to different tasks, and these network loads and tasks can even be varied based on different system states. Moreover, note that the increase in the network load is not linear. This is due to the factor ψ, which is dependent on the steady-state probabilities and system utilisation.

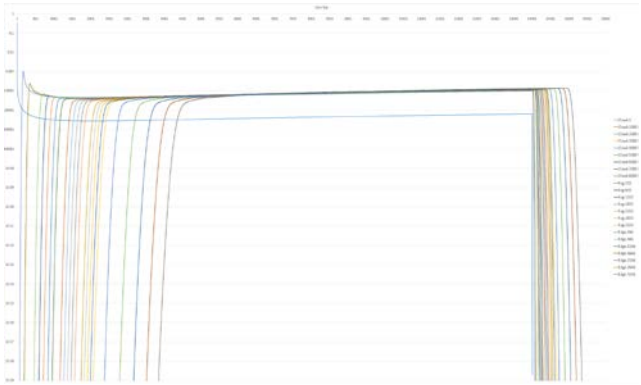


Fig. 5. Steady-State Probabilities: Cloud-Fog-Edge System (8000-4000-4000 Nodes).

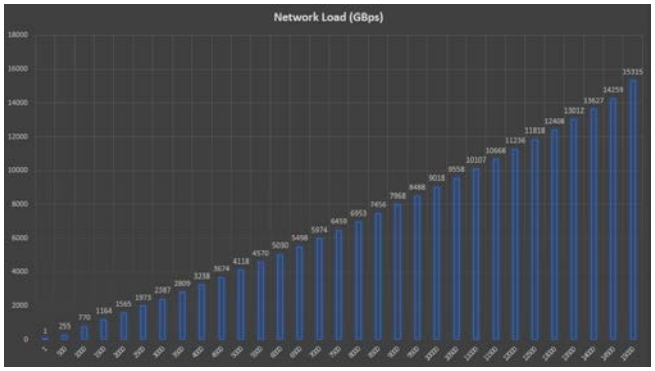


Fig. 6. Network Load in GBps (Cloud with 15,000 Nodes).

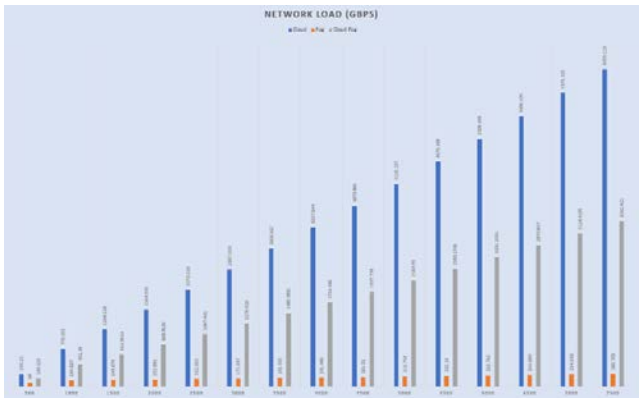


Fig. 7. Network Load in GBps (Cloud-Fog System with 7,500 Nodes Each).

Fig. 7 shows the network workload in GBps for a cloud-fog workflow scheduling system containing 7,500 nodes each in the cloud and fog. As mentioned earlier, an equal number of nodes are used for simplicity of explanation and it does not pose any limitations on the design of the systems. There are a total of 15 different systems for cloud and fog each that have been modelled and their network loads have been computed. These 15 x 2 systems relate to different workloads on the systems in terms of the tasks being received by the system, beginning from 500 tasks, up to 7,500 tasks. The capacity in all of these 30 systems has been kept constant. The cloud network load is depicted using blue bars and the fog network load is depicted using orange bars. The third set of bars in the grey colour represents the aggregate network load of the cloud-fog systems. The minimum network load is for the system with 500 tasks; it is 255 GBps for cloud and 84 GBps for the fog system, with ~170 GBps the aggregate network load. The network load consistently rises to reach 6,459 GBps (cloud), ~266 (fog), and 3362 GBps (aggregate) for the busiest workflow scheduling system. Note that the consistent increase in the network load is due to the equal load of each job as has been explained earlier. The relatively smaller values for fog systems is due to the smaller network latencies, implying a much lower time period for the fog tasks to travel over the fog-edge networks compared to the fog-cloud networks.

Fig. 8 plots the network workload in GBps for a cloud-fog-edge workflow scheduling system containing 5000 nodes each in the cloud, fog, and edge layers. An equal number of nodes are used for simplicity of explanation and it does not pose any limitations on the design of the systems. There are a total of 10 different systems for cloud and fog each, which have been modelled. These 10 x 3 systems relate to the different workloads on the systems in terms of the tasks being received by the system, beginning from 500 tasks, up to 5000 tasks per hour while the capacity all the systems is constant. The network load is depicted using blue, orange, and grey bars for cloud, fog and edge respectively. The edge values are relatively small and therefore their bars are not visible but the labels can be seen on the right of the orange bars. The fourth set of bars in the yellow colour represents the aggregate network load of the cloud-fog-edge systems. The minimum network load is for the system with 500 tasks; 255 GBps (cloud), 58 GBps (fog), 5 GBps (edge), and 106 (aggregate). Note that the network load for the fog system with 500 nodes was 84 in the cloud-fog system depicted in Fig. 7; the higher value in that case is due to the cloud system that had 7500 maximum nodes. Since we modelled systems with cloud scheduling first, the fog is scheduled after the 7500 cloud nodes and this increase the overall load of system and in turn the factor ψ, which is dependent on the steady-state probabilities and system utilisation. The network load consistently rises to a maximum of 4,118 GBps (cloud) and 1433 GBps (aggregate) for the busiest workflow scheduling system with 5000 nodes each in the cloud, fog, and edge. The reason for relatively smaller values for fog (and edge) systems has been explained in the previous paragraph.

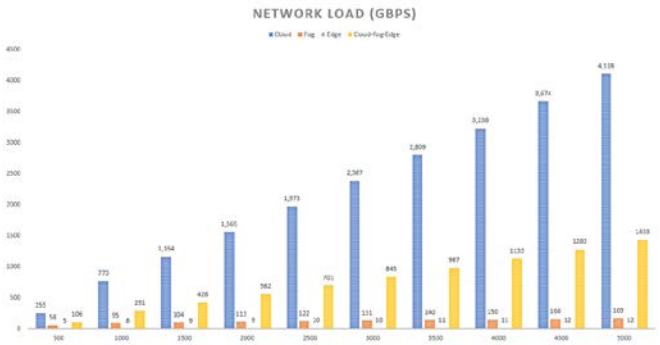


Fig. 8. Network Load in GBps (Cloud-Fog-Edge System with 5,000 Nodes Each).

D. Average Response Delay

The calculations of average Response Delay (RD) were explained earlier in Section III and Equation (9). We now in this section will analyse the system performance related to response delay of various cloud-fog, and cloud-fog-edge systems. We have modelled cloud-only and several other different configuration systems but we will limit our analysis to the cloud-fog and cloud-fog-edge systems for the sake of brevity.

Fig. 9 depicts the average response delay in milliseconds (ms) for a cloud-fog workflow scheduling system containing 7,500 nodes each in the cloud and fog. This system is similar to the one depicted in Fig. 7. There are a total of 15 different systems for cloud and fog each that have been modelled and their average response delays have been computed according to Equation (9). The cloud response delay is depicted using blue bars and the fog delay is depicted using orange bars. The third set of bars in the grey colour represents the aggregate

response delay of the cloud-fog systems. The minimum delay is for the system with 500 tasks; 404ms, 102ms, and 253ms for cloud, fog, and aggregate delays, respectively. The increase in the delay is consistent and non-linear, however, minimal. The minimal increase is because the system serves the jobs in parallel. Also, this minimal value and consistent increase is due to the equal load of each job as has been explained earlier. This increase in the delay depends on the system and task characteristics that can be tuned using the factor ψ and using tasks with some low or high-variance distributions.

Fig. 10 depicts the average response delay in milliseconds (ms) for a cloud-fog-edge workflow scheduling system containing 5000 nodes each in the cloud, fog, and edge. This system is similar to the one depicted in Fig. 8. There are a total of 10 different systems for cloud, fog, and edge, each, which have been modelled and their average response delays have been computed according to Equation (9). The cloud, fog, edge, and aggregate response delays are depicted using blue, orange, grey, and yellow bars. The minimum delay is for the system with 500 tasks; 404ms, 102ms, 100ms, and 202ms for cloud, fog, edge, and aggregate delays, respectively. The increase in the delay is consistent and non-linear, however, minimal. We have explained the reasons for this while explaining Fig. 9. Note that using a cloud-fog-edge system as opposed to cloud-only or cloud-fog system significantly decreases the aggregate delay bringing to half of it (from 652 to 326 ms). Obviously, increasing the relative number of edge nodes compared to cloud and fog can significantly bring down the aggregate delays.

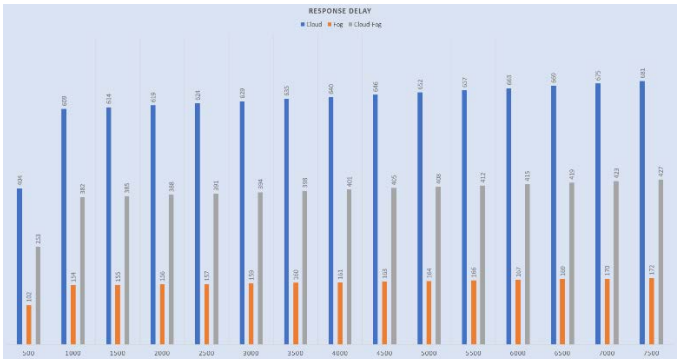


Fig. 9. Response Delay -- Milliseconds -- Cloud-Fog System with 7,500 Nodes Each.

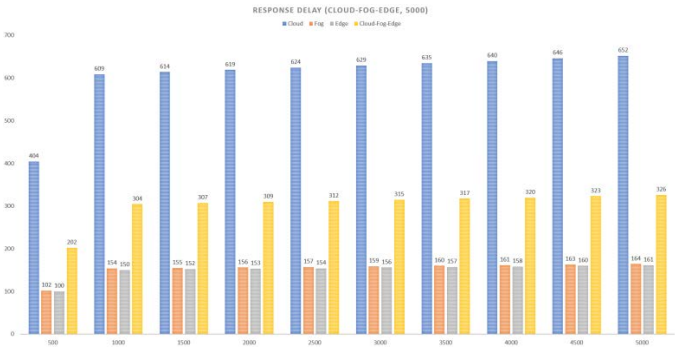


Fig. 10. Response Delay -- Milliseconds -- Cloud-Fog-Edge System with 5000 Nodes Each.

E. Energy Consumption

The network energy consumption per hour (NE) was defined and explained in Section III and calculated using Equation (10). We, in this section, analyse the system performance related to network energy consumption of various cloud-fog-edge systems. We have modelled cloud-only and cloud-fog, and several other different configuration systems, however, for the sake of brevity, we will limit our analysis in this section to the cloud-fog-edge systems.

Fig. 11 depicts the network energy consumption in MWh for a cloud-fog-edge workflow scheduling system containing 5000 nodes each in the cloud, fog, and edge. This system is similar to the one depicted in Fig. 10, however, it provides network energy consumption data. There are a total of 10 different systems for cloud, fog, and edge, each, which have been modelled and their energy consumption have been computed according to Equation (10). The cloud, fog, edge, and aggregate network energy consumption are depicted using blue, orange, grey, and yellow bars. The minimum energy consumption is for the system with 500 tasks; 496, 112, 10, and 206 MWh for cloud, fog, edge, and aggregate energy consumption, respectively. The increase in the consumption is consistent, non-linear, and reaches roughly 16 times (496 to 8006) as opposed to the 10 times increase in the number of nodes (500 to 5000). We have explained the reasons for this while explaining Fig. 9. Note that using a cloud-fog-edge system as opposed to cloud-only or cloud-fog system significantly decreases the aggregate energy consumption (from 8006 to 2786 MWh).

F. Energy Cost

The energy cost (EC) per hour was defined and explained in Section III and calculated using Equation (11). We here analyze the system performance related to network energy cost of various cloud-fog-edge systems. We have modelled cloud-only and cloud-fog, and several other different configuration systems; however, for the sake of brevity, we will limit our analysis in this section to the cloud-fog-edge systems.

Fig. 12 depicts the energy cost in GBP (x million) for a cloud-fog-edge workflow scheduling system containing 5000 nodes each in the cloud, fog, and edge. There are a total of 10 different systems for cloud, fog, and edge, each, which have been modelled and their energy cost have been computed according to Equation (11). The cloud, fog, edge, and aggregate monthly network energy consumption are depicted using blue, orange, grey, and yellow bars. The minimum monthly energy cost is for the system with 500 tasks; 62, 14, 1, and 26 million GBP for cloud, fog, edge, and aggregate monthly cost respectively. The increase in the cost is consistent, non-linear, and reaches roughly 16 times (62 to 1003) as opposed to the 10 times increase in the number of nodes (500 to 5000). Note that using a cloud-fog-edge system as opposed to cloud-only or cloud-fog system significantly decreases the aggregate energy consumption (from 1003 to 349 million GBP).

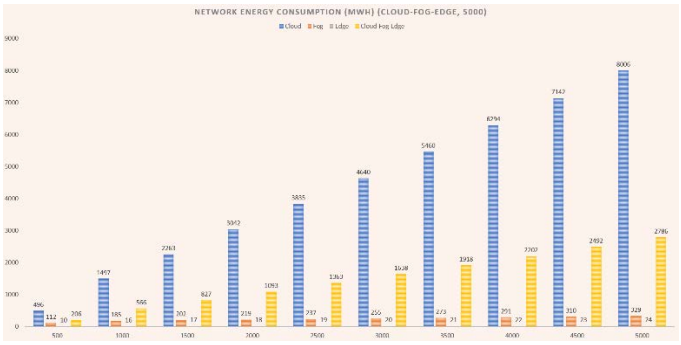


Fig. 11. Network Energy Consumption -- MWh -- Cloud-Fog-Edge System with 5000 Nodes Each.

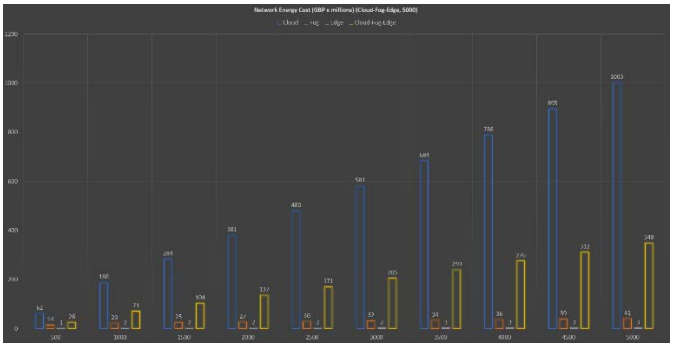


Fig. 12. Monthly Energy Cost – GBP x million -- Cloud-Fog-Edge System with 5000 Nodes Each.

V. Discussion, Utilization and Conclusion

Fog and edge computing has emerged as an important paradigm to address many challenges related to time-sensitive and real-time applications, high network loads, user privacy, security, and others. These developments offer huge potential; however, many efforts are needed to study and design applications and systems for these emerging computing paradigms.

This paper provided a detailed study of workflow scheduling and offloading of service-based applications. We developed different models of cloud, fog and edge systems and studied the scheduling of workflows using a range of system sizes and application intensities. Firstly, we developed several Markov models of cloud, fog, and edge systems and computed the steady-state probabilities for system utilization and stability. Secondly, using steady-state probabilities, we defined a range of system metrics to study the performance of workflow scheduling and offloading including, network load, response delay, energy consumption, and energy costs. An extensive investigation of application intensities and cloud, fog, and edge system sizes revealed that significant benefits can be accrued from the use of fog and edge computing in terms of low network loads, response times, energy consumption and costs.

The proposed workflow scheduling and offloading models can be utilized in practice to study a range of applications and derive several benefits. Firstly, different well-known standardized workflow can be plugged in our proposed workflow scheduling models to study their various performance behaviors including network load, average response delay, energy consumption, and energy cost, and this can be done for a range of cloud only, cloud-fog, and cloud-fog-edge systems. Some examples of standardized workflows include Montage workflow, SIPHIT workflow, epigenomics workflow, LIGO workflow, Cyber-Shake workflow, and more; see [23], for explanations and use cases of these workflows, and [19]–[22] for additional examples for practical utilization of our work. Let us take the epigenomics workflow as an example that captures the execution workflows related to the operations involved in genome sequences. Such a workflow can be embedded in our proposed Markov model by defining the execution workflows within the Markov chain and thereby we can study how that workflow will behave for cloud-only, cloud-fog, and cloud-fog-edge systems in terms of the network load, average response delay, energy consumption, and energy cost of the system.

The computational loads used by the different tasks modeled in this paper are the same. Similarly, the network loads in terms of the bytes sent around the network are also the same. However, this is not the limitation of the model. The equations developed in the models do use different computational and network loads and other parameters. The proposed models can capture the additional network load due to the task offloading or the different execution times of tasks in nodes due to the differences in their computational performance such as device speed and power that may also lead to higher energy consumption by the devices and networks. This is because the models define separately each of the tasks' computational and network loads, as well as computational and network characteristics of the devices and networks in the cloud, fog and edge layers. These could be easily changed based on various workflows to study their performance. However, the developed system is a Markov chain and therefore it does use exponential distribution to capture the arrival and departure rates.

VI. Future Work

The future work will focus on investigating variations in the cloud, fog, and edge system configurations, variations in application intensities, and variations in task sizes. Moreover, we plan to investigate system utilization and stability models and their effects on the computations of system characteristics including energy consumption and costs, response times, and network throughput.

Acknowledgment

The author gratefully acknowledges the approval and the support of this research from the Deanship of Scientific Research study by the grant no CIT-2019-1-10-F-8452, Northern Border University, Arar 91431, Kingdom of Saudi Arabia.

References

[1] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A Systematic Survey of Industrial Internet of Things Security: Requirements and Fog Computing Opportunities," IEEE Commun. Surv. Tutorials, vol. 22, no. 4, pp. 2489–2520, Oct. 2020, doi: 10.1109/COMST.2020.3011208.

[2] Y. Arfat et al., "Enabling Smarter Societies through Mobile Big Data Fogs and Clouds," in Procedia Computer Science, 2017, vol. 109, pp. 1128–1133, doi: 10.1016/j.procs.2017.05.439.

[3] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, "UbeHealth: A personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities," IEEE Access, vol. 6, pp. 32258–32285, 2018, doi: 10.1109/ACCESS.2018.2846609.

[4] A. Yousefpour et al., "All one needs to know about fog computing and related edge computing paradigms: A complete survey," J. Syst. Archit., vol. 98, pp. 289–330, 2019, doi: 10.1016/j.sysarc.2019.02.009.

[5] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa, and H. Song, "Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications," IEEE Access, vol. 4, pp. 6171–6180, 2016, doi: 10.1109/ACCESS.2016.2613278.

[6] Y. Han, X. Wang, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," 2019.

[7] W. Yu et al., "A Survey on the Edge Computing for the Internet of Things," IEEE Access, vol. 6, pp. 6900–6919, 2018, doi: 10.1109/ACCESS.2017.2778504.

[8] T. Mohammed, A. Albeshri, I. Katib, and R. Mehmood, "UbiPriSEQ—Deep reinforcement learning to manage privacy, security, energy, and QoS in 5G IoT hetnets," Appl. Sci., vol. 10, no. 20, 2020, doi: 10.3390/app10207120.

[9] N. Janbi, I. Katib, A. Albeshri, and R. Mehmood, "Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoE and 6G Environments," Sensors, vol. 20, no. 20, p. 5796, Oct. 2020, doi: 10.3390/s20205796.

[10] A. Marchisio et al., "Deep Learning for Edge Computing: Current Trends, Cross-Layer Optimizations, and Open Research Challenges," in Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI, Jul. 2019, vol. 2019-July, pp. 553–559, doi: 10.1109/ISVLSI.2019.00105.

[11] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," Proc. IEEE, vol. 107, no. 8, pp. 1738–1762, May 2019, doi: 10.1109/JPROC.2019.2918951.

[12] S. Bosaeed, I. Katib, and R. Mehmood, "A Fog-Augmented Machine Learning based SMS Spam Detection and Classification System," 2020, doi: 10.1109/FMEC49853.2020.9144833.

[13] J. C. Guevara, R. da S. Torres, and N. L. S. da Fonseca, "On the classification of fog computing applications: A machine learning perspective," J. Netw. Comput. Appl., vol. 159, p. 102596, Jun. 2020, doi: 10.1016/J.JNCA.2020.102596.

[14] G. Javadzadeh and A. M. Rahmani, "Fog Computing Applications in Smart Cities: A Systematic Survey," Wirel. Networks 2019 262, vol. 26, no. 2, pp. 1433–1457, Dec. 2019, doi: 10.1007/S11276-019-02208-Y.

[15] M. Abdel-Basset, R. Mohamed, M. Elhoseny, A. K. Bashir, A. Jolfaei, and N. Kumar, "Energy-Aware Marine Predators Algorithm for Task Scheduling in IoT-Based Fog Computing Applications," IEEE Trans. Ind. Informatics, vol. 17, no. 7, pp. 5068–5076, Jul. 2021, doi: 10.1109/TII.2020.3001067.

[16] S. N. Srirama, F. M. S. Dick, and M. Adhikari, "Akka framework based on the Actor model for executing distributed Fog Computing applications," Futur. Gener. Comput. Syst., vol. 117, pp. 439–452, Apr. 2021, doi: 10.1016/J.FUTURE.2020.12.011.

[17] M. Abdel-Basset, D. El-Shahat, M. Elhoseny, and H. Song, "Energy-Aware Metaheuristic Algorithm for Industrial-Internet-of-Things Task Scheduling Problems in Fog Computing Applications," IEEE Internet Things J., vol. 8, no. 16, pp. 12638–12649, Aug. 2021, doi: 10.1109/JIOT.2020.3012617.

[18] M. Goudarzi, H. Wu, M. Palaniswami, and R. Buyya, "An Application Placement Technique for Concurrent IoT Applications in Edge and Fog Computing Environments," IEEE Trans. Mob. Comput., vol. 20, no. 4, pp. 1298–1311, Apr. 2021, doi: 10.1109/TMC.2020.2967041.

[19] M. Farid, R. Latip, M. Hussin, and N. A. W. A. Hamid, "A Survey on QoS Requirements Based on Particle Swarm Optimization Scheduling Techniques for Workflow Scheduling in Cloud Computing," Symmetry 2020, Vol. 12, Page 551, vol. 12, no. 4, p. 551, Apr. 2020, doi: 10.3390/SYM12040551.

[20] M. Kumar, S. C. Sharma, A. Goel, and S. P. Singh, "A comprehensive survey for scheduling techniques in cloud computing," J. Netw. Comput. Appl., vol. 143, pp. 1–33, Oct. 2019, doi: 10.1016/J.JNCA.2019.06.006.

[21] I. Grosof, M. Harchol-Balter, and A. Scheller-Wolf, "The Finite-Skip Method for Multiserver Analysis," vol. 1, no. 1, Sep. 2021, Accessed: Dec. 21, 2021. [Online]. Available: https://arxiv.org/abs/2109.12663v1.

[22] M. Ala'anzy and M. Othman, "Load Balancing and Server Consolidation in Cloud Computing Environments: A Meta-Study," IEEE Access, vol. 7, pp. 141868–141887, 2019, doi: 10.1109/ACCESS.2019.2944420.

[23] M. Adhikari, T. Amgoth, and S. N. Srirama, "A Survey on Scheduling Strategies for Workflows in Cloud Environment and Emerging Trends," ACM Comput. Surv., vol. 52, no. 4, Aug. 2019, doi: 10.1145/3325097.

[24] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal Workload Allocation in Fog-Cloud Computing Toward Balanced Delay and Power Consumption," IEEE Internet Things J., vol. 3, no. 6, pp. 1171–1181, Dec. 2016, doi: 10.1109/JIOT.2016.2565516.

[25] R. Deng, R. Lu, C. Lai, and T. H. Luan, "Towards power consumption-delay tradeoff by workload allocation in cloud-fog computing," IEEE Int. Conf. Commun., vol. 2015-September, pp. 3909–3914, Sep. 2015, doi: 10.1109/ICC.2015.7248934.

[26] J. Oueis, E. C. Strinati, S. Sardellitti, and S. Barbarossa, "Small cell clustering for efficient distributed fog computing: A multi-user case," 2015 IEEE 82nd Veh. Technol. Conf. VTC Fall 2015 - Proc., Jan. 2016, doi: 10.1109/VTCFALL.2015.7391144.

[27] M. Aazam and E. N. Huh, "Fog computing micro datacenter based dynamic resource estimation and pricing model for IoT," Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA, vol. 2015-April, pp. 687–694, Apr. 2015, doi: 10.1109/AINA.2015.254.

[28] K. Habak, M. Ammar, E. W. Zegura, and K. A. Harras, "Workload management for dynamic mobile device clusters in edge femtoclouds," 2017 2nd ACM/IEEE Symp. Edge Comput. SEC 2017, Oct. 2017, doi: 10.1145/3132211.3134455.

[29] K. Intharawijitr, K. Iida, and H. Koga, "Analysis of fog model considering computing and communication latency in 5G cellular networks," 2016 IEEE Int. Conf. Pervasive Comput. Commun. Work. PerCom Work. 2016, Apr. 2016, doi: 10.1109/PERCOMW.2016.7457059.

[30] H. J. Jeong, I. Jeong, H. J. Lee, and S. M. Moon, "Computation offloading for machine learning web apps in the edge server environment," Proc. - Int. Conf. Distrib. Comput. Syst., vol. 2018-July, pp. 1492–1499, Jul. 2018, doi: 10.1109/ICDCS.2018.00154.

[31] T. Zhao, S. Zhou, X. Guo, and Z. Niu, "Tasks scheduling and resource allocation in heterogeneous cloud for delay-bounded mobile edge computing," IEEE Int. Conf. Commun., Jul. 2017, doi: 10.1109/ICC.2017.7996858.

[32] S. Agarwal, S. Yadav, and A. K. Yadav, "An Efficient Architecture and Algorithm for Resource Provisioning in Fog Computing," undefined, vol. 8, no. 1, pp. 48–61, Jan. 2016, doi: 10.5815/IJIEEB.2016.01.06.

[33] A. Yousefpour et al., "All one needs to know about fog computing and related edge computing paradigms: A complete survey," J. Syst. Archit., vol. 98, pp. 289–330, Sep. 2019, doi: 10.1016/J.SYSARC.2019.02.009.

[34] D. Zeng, L. Gu, S. Guo, Z. Cheng, and S. Yu, "Joint Optimization of Task Scheduling and Image Placement in Fog Computing Supported Software-Defined Embedded System," IEEE Trans. Comput., vol. 65, no. 12, pp. 3702–3712, Dec. 2016, doi: 10.1109/TC.2016.2536019.

[35] M. Chen and Y. Hao, "Task Offloading for Mobile Edge Computing in Software Defined Ultra-Dense Network," IEEE J. Sel. Areas Commun., vol. 36, no. 3, pp. 587–597, Mar. 2018, doi: 10.1109/JSAC.2018.2815360.

[36] T. Huang et al., "Energy-Efficient Computation Offloading for Multimedia Workflows in Mobile Cloud Computing," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 270, pp. 113–123, Nov. 2018, doi: 10.1007/978-3-030-12971-2_7.

[37] J. Wan, B. Chen, S. Wang, M. Xia, D. Li, and C. Liu, "Fog Computing for Energy-Aware Load Balancing and Scheduling in Smart Factory," IEEE Trans. Ind. Informatics, vol. 14, no. 10, pp. 4548–4556, Oct. 2018, doi: 10.1109/TII.2018.2818932.

[38] M. A. Benblidia, B. Brik, L. Merghem-Boulahia, and M. Esseghir, “Ranking fog nodes for tasks scheduling in fog-cloud environments: A fuzzy logic approach,” 2019 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2019, pp. 1451–1457, Jun. 2019, doi: 10.1109/IWCMC.2019.8766437.

[39] R. O. Aburukba, M. AliKarrar, T. Landolsi, and K. El-Fakih, “Scheduling Internet of Things requests to minimize latency in hybrid Fog–Cloud computing,” Futur. Gener. Comput. Syst., vol. 111, pp. 539–551, Oct. 2020, doi: 10.1016/J.FUTURE.2019.09.039.

[40] W. J. Stewart, Numerical Solution of Markov Chains (Probability: Pure & Applied). New York: CRC, 1991.

[41] S. Altowaijri, R. Mehmood, and J. Williams, “A Quantitative Model of Grid Systems Performance in Healthcare Organisations,” 2010 Int. Conf. Intell. Syst. Model. Simul., pp. 431–436, 2010, doi: 10.1109/ISMS.2010.84.

[42] R. Mehmood, R. Meriton, G. Graham, P. Hennelly, and M. Kumar, “Exploring the influence of big data on city transport operations: a Markovian approach,” Int. J. Oper. Prod. Manag., vol. 37, no. 1, pp. 75–104, 2017, doi: 10.1108/IJOPM-03-2015-0179.

[43] A. Andrae and T. Edler, “On Global Electricity Usage of Communication Technology: Trends to 2030,” Challenges, vol. 6, no. 1, pp. 117–157, Apr. 2015, doi: 10.3390/challe6010117.

[44] Selectra, “Energy bills: How much is the average electric bill?” https://selectra.co.uk/energy/guides/billing/electricity (accessed Nov. 29, 2021).

Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur

Shuzlina Abdul-Rahman[1], Sofianita Mutalib[4]
Research Initiative Group of Intelligent Systems
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

Nor Hamizah Zulkifley[2]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

Ismail Ibrahim[3]
Data Science Department
PETRONAS Digital Sdn Bhd
Kuala Lumpur, Malaysia

***Abstract*—House price is affected significantly by several factors and determining a reasonable house price involves a calculative process. This paper proposes advanced machine learning (ML) approaches for house price prediction. Two recent advanced ML algorithms, namely LightGBM and XGBoost were compared with two traditional approaches: multiple regression analysis and ridge regression. This study utilizes a secondary dataset called 'Property Listing in Kuala Lumpur', gathered from Kaggle and Google Map, containing 21984 observations with 11 variables, including a target variable. The performance of the ML models was evaluated using mean absolute error (MAE), root mean square error (RMSE), and adjusted r-squared value. The findings revealed that the house price prediction model based on XGBoost showed the highest performance by generating the lowest MAE and RMSE, and the closest adjusted r-squared value to one, consistently outperformed other ML models. A new dataset which consists of 1300 samples was deployed at the model deployment stage. It was found that the percentage of the variance between the actual and predicted price was relatively small, which indicated that this model is reliable and acceptable. This study can greatly assist in predicting future house prices and the establishment of real estate policies.**

***Keywords*—House price; house price prediction; machine learning; property; regression analysis**

I. Introduction

House is one of the most essential basic needs in human life, along with other basic needs such as food and water. Demand for houses has rapidly increased through the years as people's standard of living has improved. Even though there are people who make their house as an investment and asset, yet most people around the world buy a house to live in. Undoubtedly, the housing sector has a positive impact on a country's currency, which is an important scale for the national economy [1]. Homeowners will buy goods such as furniture and house appliances for their house and home builders or contractors will buy raw material to make houses to fulfil the demand for houses, which is an example of the economic wave effect created from new house supply. Meanwhile, consumers have the capital to make a large investment, and the construction industry is vibrant or otherwise can be seen through the high level of house supply or demand in a country. Nevertheless, house has become unaffordable as there is a significant price expansion in the housing market sector in many countries [2].

In Malaysia, buying a house is never an easy experience because this decision can cost a lot of money. According to the Department of Statistics Malaysia, the average price of a house in the 2nd quarter of 2018 was RM 408,774 compared to ten years ago which was RM 199,431 [3]. The average house price in Malaysia was obviously two times higher in 2018 compared to 2009. A wide variety of factors may affect house prices such as the facilities provided, type of houses, number of bedrooms and size of a house. These factors also vary depending on the location of the house, for example, there will be an obvious difference in house prices, in Singapore compared to Malaysia. Yet, we cannot simply say that the price of each house is similar throughout Malaysia as the prices of houses in Kuala Lumpur (urban area) is not the same as the price of houses in Perlis (rural area). Yop [3] in his report mentioned that the average house price in Kuala Lumpur in 2018 was RM 772,398, while the average house price in Perlis for the same year was just RM 177,945. It is suggested that the prediction of house prices could be more accurate if the prediction is considered based primarily on a specific region.

House prices can be predicted using machine (ML) algorithms including support vector regression, artificial neural network, and linear regression. The ML model that provides the best prediction results will be beneficial for researchers, home buyers, property investors, and house builders in terms of gaining a lot of knowledge and information of the house price values in the present sector. Additionally, this model can facilitate potential buyers to determine the characteristics of a house they prefer that adhere to their budget [4]. Prediction of house prices in Kuala Lumpur would be a significant research as Kuala Lumpur is the capital city of Malaysia that offers a range of facilities

including efficient public transportation, shopping malls, and many more compared to states in a rural area such as Perlis has yet to provide. All the facilities provided can encourage many home buyers, investors, and house builders to supply more houses in this area, at the same time the demand for houses in this specific region would increase. Nonetheless, the suitable model in predicting house prices in Kuala Lumpur remains unclear as there is limited study conducted in this region.

Research showed that ML algorithms have proven extremely useful for addressing many predictions and classification problems with broad application scope, including customer classification and segmentation [5], market analysis [6], [7], and education [8]. Unfortunately, ML is relatively limited and very far from being used in real estate applications mainly in the Malaysian sense. Evaluation of property prices and values is extremely critical for the real estate sector, the stock market, the economics and tax sector, as well as the scale of buyers' and sellers' wallets [9]. Although researchers are relatively aware of the existing current prediction model, consideration should be given to the current methodologies constrained by the scope of data of the current system in the real estate industry. Therefore, it is important to examine the correlation between house prices and housing attributes and identify significant variables that are essential to the use of the ML techniques in the real estate industry, involving pre-processing and exploration of the datasets obtained.

As various factors such as location and property demand could affect house prices, most parties involved including buyers and investors, housebuilders, and real estate market may want to know the exact attributes or the main factors affecting house prices to assist investors in making decision and to facilitate house builders in setting house prices [2], [10], [11]. Nevertheless, other characteristics including distance from local facilities and the physicality of a house might be overlooked. This may contribute to the creation of a house price model that does not reflect the actual condition of the housing sector. In comparison, houses with identical attributes may be priced at varying rates, while houses with different type of attributes can be priced at the same amount. Besides that, real-estate industry in Malaysia is far behind compared to that in the United States and United Kingdom in valuation [2], [12] as Malaysia is still using the traditional method to valuate a house. Developers or real-estate agencies will send valuers to each house, to appraise the house price resulting in wastage of time and cost, also the fluctuation of house prices. However, this issue and many other pressing industrial problems can be effectively addressed using big-data technologies that includes ML in this age of Industry 4.0.

This paper presents an exploration of ML algorithms for house price prediction by focusing on Kuala Lumpur housing data. The study aims to propose an advanced ML algorithm that can generate a promising model for house price prediction. To achieve this aim, four models namely multiple regression analysis, ridge regression, light gradient boosting machine (LightGBM), and XGBoost are used to learn about the relationship between the house attributes and house prices as well as to predict the house price. The remainder of this paper is organized as follows: Section II describes related works on prediction of house price in Malaysia and followed by housing price prediction models. Subsequently, in Section III and Section IV, the experimental designs and the data modeling are presented consecutively. Section V presents the model deployment while Section VI discusses the results and findings of the study. Finally in Section VII, the study is concluded.

II. Related Work

A. Prediction of House Price in Malaysia

Prior to 2019, there were several studies regarding the prediction of house prices in Malaysia. In 2018, a study has been conducted by Yap and Ng [13] to determine house affordability in Malaysia. This study found out that key factors affecting affordability of houses were income, price of a property, land cost, policy of supply and demand, and changes in the economy. Their study [13] offered detailed insights into exploring housing market in Malaysia; however, they only provided a descriptive analysis, which lacked in the predictive analysis towards the housing market in Malaysia. The study by [2] found out that there were several attributes that play a significant role in determining house prices in Petaling District. Another significant finding from this study was Puchong and Petaling Jaya was classified as less volatile housing markets compared to Sungai Buloh. A similar study conducted by Abdullahi et al. [14] using Multiple Regression Analysis and Hedonic Regression Analysis in explaining price variations in Malaysia found that one of the most influential attributes was the location of a house.

Apart from that, building area and building age were significant in variations of prices of houses in Malaysia. The locations mentioned in the study were based on states in Malaysia; hence location is the most dominant attribute compared to others (attributes). Meanwhile, this study focused on the location of a small region of Kuala Lumpur to predict the house price. Thus, location cannot be generalized as the main significant attribute in predicting house prices. Another similar study conducted examined the prediction of house prices in Petaling District Malaysia. The study by Chang et al. [9] used a different method called the Functional Relationship Model. The model was used to identify the impact of residential property attributes on house price in Petaling District. Several attributes were identified such as building size and bedroom numbers which had a significant effect in predicting the house prices. The study successfully developed a new predictive model and then applied on the Petaling District terrace-houses only, even though, there existed various types of houses in the area such as bungalows, condominiums, or apartments. Taking everything into account, the model was unable to reflect the whole housing market in Petaling District.

B. Housing Price Prediction Models

ML algorithms to develop house price prediction models have been actively researched and models are constructed by using algorithms such as random forest, decision tree, lasso, and linear regression [12], [15]. A study by Wu et al. [12] categorized models in analyzing the real-estate market into the

conventional valuation system and the advanced valuation system. The traditional valuation system includes the multiple regression method and stepwise regression method, while the advanced valuation is the hedonic pricing method, artificial neural network (ANN), and spatial analysis method. The choice of a model that needs to be used to predict house prices is quite critical as there is a variety of models available. One of the most widely used models in the real-estate field is the regression analysis namely the multiple linear regression, support vector regression, and hedonic regression analysis, which is used by several researchers including [9], [16]–[20]. In addition, several other machine learning models such as the gradient boosting model including Catboost, XGBoost and LightGBM, random forest, decision and artificial neural network have been used frequently in the study of real- estate [10], [11], [21]–[23]. In this study, four models namely multiple regression analysis, ridge regression, LightGBM, and XGBoost are used to learn about the relationship between house attributes and house prices as well as to predict house prices. The next subsections explain further these four models.

- Multiple Regression Analysis

Multiple Regression analysis is an extension of simple linear regression to predict the value of a variable based on the value of two or more other variables. It can determine the strength of the relationship between an outcome (the dependent variable) and several predictor variables as well as the importance of each of the predictors to the relationship [24] [25]. Four basic assumptions need to be fulfilled to use the multiple regression analysis model [26] as cited in [27]. The first assumption is the variable used in the model must be normally distributed. Multivariate data cleaning is also an important consideration in multiple regression.

The second assumption that needs to be met is the relationship between the dependent and independent variables which must be linear to estimate the variables accurately. Next, the assumption to use this model is no perfect multicollinearity between variables. The last assumption to be fulfilled is little or no auto correlation. There are various other assumptions for this model; however, these assumptions are among the easiest to deal with if needed. The prediction of house prices using multiple regression model is conducted by assigning the price of a house as a target variable or dependent variable, while other attributes are set as independent variables to determine the most significant variables by identifying the correlation coefficient of each attribute.

- Ridge Regression

Ridge regression is a technique used to assess the presence of collinearity in multiple regression data [28]. As multicollinearity happens, the estimates of the least-squares are unbiased, but their variances are large enough such that they can be far from the real value. Ridge regression provides a more credible performance by reducing the standard errors. There were several industries that have deployed ridge regression model as their solution. For example, in medical industry, this model was deployed in healthcare analysis system and blood-base tissue gene expression as well as in wind speed forecasting [29]–[31]. According to Manasa et al. [10], the ridge regression model is a regularization model that incorporates and optimizes an additional variable (tuning parameter) to resolve the effect of multiple variables in linear regression, typically referred to as ‘noise’ in a statistical sense. This model has been used by a lot of researchers in the real-estate previous studies including [10], [32]–[34]. In their study, [33] found that ridge regression is able to provide the lowest MSE value for the house prediction model compared to other models in that study including lasso and gradient boosting.

- Light Gradient Boosting Machine (LightGBM)

Initially launched in late 2017, the Light Gradient Boosting Machine or LightGBM has a stable release in November 2020. This model has been used by many researchers in their fields such as medical fields [35], and biochemistry [36], [37]. This model, however, is still rarely used in the real estate market, since there is only one academic study that has utilized this model [22]. In their research related to prediction of house prices, [22] combined this model with another prediction model, which were CatBoost and XGBoost model to forecast house rentals in China. This joint model managed to provide the smallest RMSE value of the house price prediction model.

- Extreme Gradient Boosting (XGBoost)

XGBoost is the most powerful algorithm for any regression or classification problem which stands for ‘extreme gradient booster’. The developers [38] of this model in their article describe XGBoost as a scalable tree boosting machine learning system. The framework of this model can be seen as an open-source kit. In a range of machine learning and data mining problems, the influence of this model has been widely recognized. In fact, many successful machine learning competitions utilize these styles of model. Besides, this model is often used in real-world development networks, including internet ads. According to [34], XGBoost algorithm addresses issues of the linear regression model. The XGBoost model can handle numerical as well as categorical variables very well. This model can automate different loss functions and offers many tuning choices for machine learning engineers to modify the model. This model had been used in many fields including the real estate and housing market sector [10], [32], [34], [39]. Most studies in the real estate industry that use the XGBoost model found out that this model is able to provide the lowest RMSE value for the prediction of house prices compared to other models [23], [34]. Table I summarizes past studies that work on similar domains from 2011 until 2020. As can be seen, the first four ML models appear to be the most commonly used by past researchers. The list of variables that associate with this domain can be referred in our paper [40]. As stated in the paper, the factors influencing house prices can be classified into three categories: location, structural and neighborhood condition.

TABLE I. SSUMMARY OF LITERATURES ON HOUSE PRICE PREDICTION USING ML MODELS

Previous study	MRA	RR	LGBM	XGB	RF	NN
[23]	✔			✔		
[21]			✔	✔	✔	
[10]		✔		✔		
[34], [39]	✔	✔		✔		
[16], [20]	✔					
[22]			✔	✔		
[15]	✔				✔	
[41], [42], [19]						✔
[43]	✔					
[32], [33]		✔		✔		
[14],[17], [44]	✔					

[a] MRA: Multiple Regression Analysis; RR: Ridge Regression; LGBM: Light Extreme Gradient Boosting; XGB: Extreme Gradient Boosting; RF: Random Forest; NN: Neural Networks

III. EXPERIMENTAL SETUP

A. Data Preparation

This study is based on a secondary dataset and retrieved from the Kaggle website (https://www.kaggle.com/dragonduck/property-listings-in-kuala-lumpur) and Google Maps. The dataset has originally scrapped from property listings in Kuala Lumpur, Malaysia in 2019. The original dataset contains only 8 variables including location, price, rooms, bathrooms, property type, size, and furnishing. Other than that, there were other four variables that also showed a significant impact to house price prediction derived from Geocoder and Googleplaces Python package used to retrieve locations through Google Maps. The variables were distance to shopping mall, distance to hospital, access of public transport, and distance to nearest school. Thus, the final dataset contains 53883 observations with 12 variables including one target variable. In this study, the target variable was the price (which is a continuous variable) in Ringgit Malaysia (MYR). Meanwhile, the independent variables were location (location of a house), bedroom (number of bedrooms available), bathroom (number of bathrooms available), car park, size (house lot size), furnishing status, property type, shopping mall (the nearest shopping mall to the house in KM), school (the nearest school to the house in KM), hospital (the nearest hospital to the house in KM) and public transport (the nearest LRT and MRT station to the house in KM). Table II shows the description of these variables.

TABLE II. DESCRIPTION OF VARIABLES

No.	Variable Name	Role	VariableType	Description
1.	Price	Target	Continuous	Price of house sold in the market
2.	Location	Input	Nominal	The location of a house for example: 1. Mont Kiara 2. City Centre 3. Bangsar 4. Desa Park City 5. Bukit Tunku (KennyHills)
3.	Room	Input	Discrete	Number of rooms in a house
4.	Bathroom	Input	Discrete	Number of bathrooms in a house
5.	Car Park	Input	Discrete	Number of car parks provided for a house
6.	Size	Input	Continuous	The built-up area of a house or the land area of a house
7.	Furnishing	Input	Nominal	Furnishing status of a house: 1. Fully furnished 2. Semi-furnished 3. Unfurnished
8.	Property Type	Input	Nominal	The type of condominium: 1. Corner 2. Duplex 3. End Lot 4. Intermediate 5. Penthouse 6. SOHO 7. Studio 8. Triplex
9.	Shopping Mall	Input	Continuous	The nearest shopping mall to the house in kilometer (KM)
10.	School	Input	Continuous	The nearest school to the house in kilometer (KM)
11.	Hospital	Input	Continuous	The nearest hospital to the house in kilometer (KM)
12.	LRT/MRT	Input	Continuous	The nearest LRT or MRT station to the house in kilometer (KM)

The data preparation stage included several tasks to create the final dataset. Tasks for data preparation were likely performed several times, and not in any specified order. Tasks included selection of table, record, and attribute, data cleaning, new attribute building, and data transformation for modeling tools were performed at this stage. For this study, the original dataset of a property listing was reduced through the data cleaning process stages and data transformation. Data cleaning can be referred to as a process of identifying and correcting errors in a dataset, for example removing missing values, and data transformation refers to a process of transforming a data to be more valuable towards a study. The invalid data including missing values and inconsistent data observed in this stage. A total of 21,995 rows containing missing values in which the majority of the missing values existed in more than two columns (good if we know the variables). All 21,955 rows with missing values were removed using Python. There were 31,899 rows left after removing missing values from the dataset.

B. Data Transformation

To examine whether the dataset is normally distributed or not, the study used three graphical methods such as histograms, QQ-plot, and boxplot. On the other hand, skewness and kurtosis were also used to assess normality of each variable. According to [45], a standard normal dataset has a kurtosis value between -3 to 3. If the kurtosis value is higher or lower than this value, it will be considered as a thin bell shape. Meanwhile, [46], [47] suggested that if the skewness value is between -2 to 2, the dataset can be roughly considered as normal. Based on Fig. 1, the skewness value is 5.03 and the kurtosis value is 62.75. It means that the dataset is not normally distributed. Hence, data transformation needs to be done to make the dataset conform to normality. The transformation of the dataset was executed by using log-transformation to normalize the dataset. The transformed data showed better value of skewness and kurtosis which were 0.7699 and 0.3670, respectively. These two values are depicted in Fig. 1 and Fig. 2. The skewness value (0.77) between -2 and 2 can be considered as normally distributed.

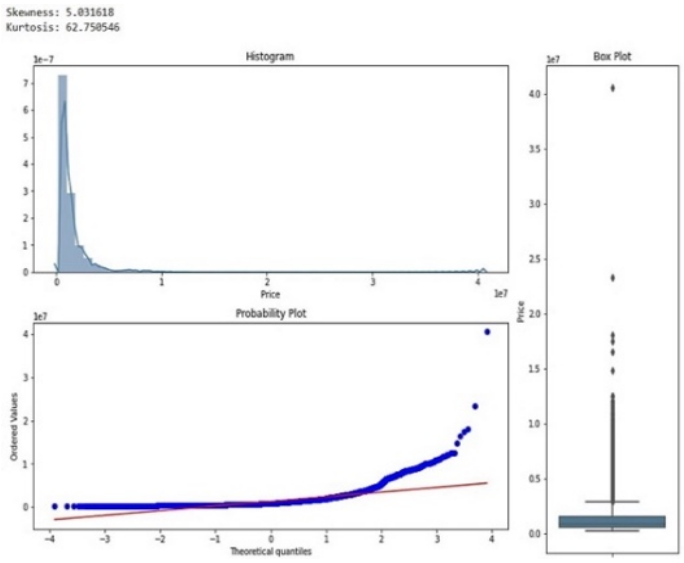


Fig. 1. Distribution of Price before Transformation.

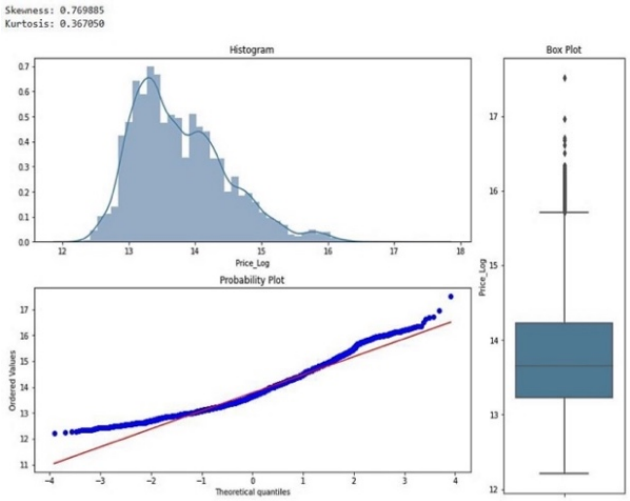


Fig. 2. Distribution of Price after Transformation.

C. Identification of Importance Features using Correlation Score

Pearson correlation matrix was utilized to check for the correlation value between independent and dependent variable. This test helped this study determine which attributes played an important role in valuating house prices. On the other hand, the correlation value was also used to detect whether there is multicollinearity in this dataset. According to Ratner (2009), multicollinearity values between 0 and 0.3 (0 and-0.3) imply a weak positive (negative) linear relationship via a shaky linear law, while values between 0.3 and 0.7 (0.3 and−0.7) via a fuzzy-firm linear rule suggest a moderate positive (negative) linear relationship and the values between 0.7 and 1.0 (-0.7 and -1) suggest a strong positive (negative) linear relationship. Fig. 3 illustrates the heatmap of the Pearson correlation coefficient matrix based on each variable. As depicted in Fig. 3, variable hospital which is the distance to the nearest hospital has a weak negative correlation which is -0.19. Meanwhile, other attributes show a moderate and strong correlation value. However, there are no attributes being dropped from this study, as this study utilizes machine learning model that can detect itself which attributes is good or not.

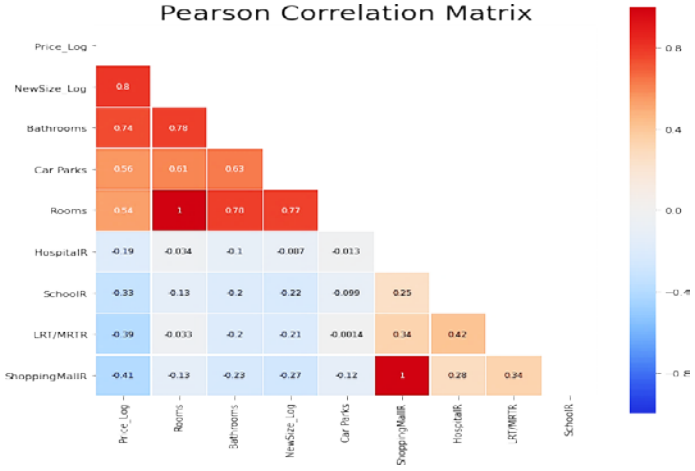


Fig. 3. Heatmap of Pearson Correlation Coefficient Matrix.

D. Data Representation using One Hot Encoding

One hot encoding is a technique used to express categorical features as numerical variables with more interpretable outcome values that will be easier to be understood by a machine learning model. For instance, the furnishing status is conveyed by a name that characterizes how furnished the space is (fully furnished, semi-furnished, and unfurnished). This technique transformed the input into 3 columns. Each column denotes a condition or status of furnishing. The column of a specific status of furnishing is 1, and those in the same row are set to 0.

IV. Data Modeling

In this study, predictive modeling was carried out using multiple regression analysis, ridge regression, LightGBM, and XGBoost. The dataset then was partitioned into two groups, training and testing sample. The training sample partition which consists of 70% of the dataset and 30% of the dataset was used in the testing sample. These four models were then compared in order to select the best model. This study utilized the hyper-parameter automated search module-GridSearchCV to search for the optimal parameter values to enhance the efficiency of each model [48]. Without losing generality, the RMSE value with optimal parameters can be generated in this module.

A. Implementation of Multiple Regression Analysis

The assumption of linear regression such as linear relationship, normal distribution of all variables, no perfect multicollinearity and little or no autocorrelation was checked before the model was used. Supposed that y represents the dependent variable, which is the prediction of house price, βk is the coefficient value of the linear function, and xk denotes each of the attributes used in this model such as the size of house and furnishing status. According to Manasa et al. [10], this model can be represented in a mathematical model as in equation (1):

$$y = \sum_{k=0}^{k} \beta_k x_k \quad (1)$$

where;

x_k, k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 denote eleven attributes for each house, y = the predicted price of a house in Kuala Lumpur. Several assumptions regarding linear regression were checked prior to the usage of this model. The first assumption of linear regression model is linearity. This model assumes that there is a linear relationship between the independent and dependent variables. Fig. 4 shows the graph of newsize_log against price_log. As can be seen, the graph shows a non-perfect linear relationship. However, it can still be said that the independent variables have a linear relationship towards dependent variable (price of house). The second assumption of linear regression is regarding the distribution of variables. This model assumes that all variables are normally distributed. This assumption has been checked by using the Q-Q plot as illustrated in Fig. 5.

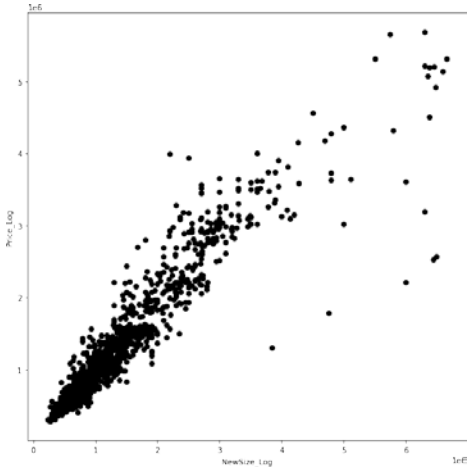


Fig. 4. Graph of Actual Price against Predicted Price.

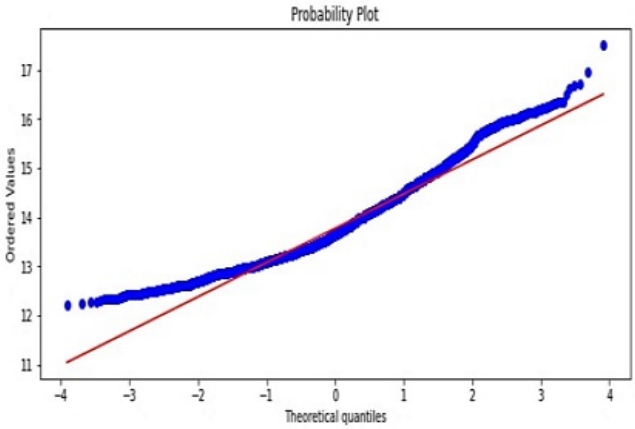


Fig. 5. The Q-Q Plot of All Variables.

Fig. 5 shows that not all the points lie perfectly on the red line which indicates that, all variables used in this study are not perfectly normal distributed. The third assumption for linear regression model is that there is no perfect multicollinearity from the variables. The multicollinearity can be examined by using the Pearson correlation score, as shown in Fig. 6.

As can be seen in Fig. 6, that there is no perfectly multicollinearity between each variable. Thus, all variables were used in this study. The last assumption for linear regression model is that there is little or no autocorrelation.

	Price	Rooms	Bathrooms	Car Parks	NewSize	ShoppingMallR	HospitalR	LRT/MRTR	SchoolR	Price_Log	NewSize_Log
Price	1.000000	0.521923	0.716527	0.539069	0.078465	-0.333602	-0.174326	-0.321983	-0.278551	0.922619	0.780736
Rooms	0.521923	1.000000	0.779465	0.605876	0.066019	-0.128391	-0.033920	-0.032961	-0.125751	0.537928	0.767445
Bathrooms	0.716527	0.779465	1.000000	0.625897	0.081031	-0.234781	-0.102386	-0.198127	-0.198296	0.742316	0.834190
Car Parks	0.539069	0.605976	0.625897	1.000000	0.067288	-0.118294	-0.013410	-0.001408	-0.099241	0.560038	0.634195
NewSize	0.078465	0.066019	0.081031	0.067288	1.000000	-0.021597	0.002629	-0.005584	-0.028538	0.073376	0.271348
ShoppingMallR	-0.333602	-0.128391	-0.234781	-0.118294	-0.021597	1.000000	0.279069	0.337276	0.250992	-0.406126	-0.269719
HospitalR	-0.174326	-0.033920	-0.102386	-0.013410	0.002629	0.279069	1.000000	0.421776	0.084124	-0.186169	-0.087223
LRT/MRTR	-0.321983	-0.032961	-0.198127	-0.001408	-0.005584	0.337276	0.421776	1.000000	0.264046	-0.390897	-0.208085
SchoolR	-0.278551	-0.125751	-0.198296	-0.099241	-0.028538	0.250992	0.084124	0.264046	1.000000	-0.326374	-0.222027
Price_Log	0.922619	0.537928	0.742316	0.560038	0.073376	-0.406126	-0.186169	-0.390897	-0.326374	1.000000	0.804256
NewSize_Log	0.780736	0.767445	0.834190	0.634195	0.271348	-0.269719	-0.087223	-0.208085	-0.222027	0.804256	1.000000

Fig. 6. The Correlation Score of Each Continuous Variable.

This last assumption was tested by using Durbin-Watson test. According to [49], there is no autocorrelation in the dataset if the Durbin-Watson test is in the range of 1.5 to 2.5. This indicates that there is no autocorrelation in this dataset. The multiple regression analysis model is being used as all the assumptions of the model have been fulfilled. The most important predictor variable analyzed by multiple regression analysis model was NewSize_Log followed by Rooms and Bathrooms which indicated the size of a house, the number of bedrooms and bathrooms. Analysis of the feature importance with Multiple Linear Regression model shows the rank of features (in sequence, highest to lowest): size of the house, number of rooms, number of bedrooms, public transport, shopping mall, carparks, school, and hospital.

B. Implementation of Ridge Regression

Equation (2) presents the equation of simple linear regression.

$$y = xb + e \quad (2)$$

where y represents the dependent variable, which is the prediction of house price, while x is the features of the matrix (number of bedrooms, location, etc.), b is the regression coefficients, and e is the residual errors. Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity that performs L2 regularization. On this basis, the variables were standardized by subtracting and dividing the respective factors by their standard deviations [10]. The rank of these five features is similar to the earlier model (i.e., Multiple Linear Regression). The first three features are NewSize_Log which is the size of a house followed by the number of bathrooms and the number of rooms. Next in the list are the nearest distance to public transport, the nearest distance to shopping mall, number of car parks, the nearest school to the house and the hospital.

C. Implementation of LightGBM

LightGBM is a Microsoft GBDT open-source algorithm. The histogram-based algorithm is used to accelerate the training process, reduces memory usage, and incorporates advanced network communication to optimize parallel learning known as the algorithm for the parallel voting tree. The leaf-wise method is used by LightGBM to find a leaf with the greatest gain in splitters. For the LightGBM algorithm, the Python LightGBM module was used to evaluate the house price dataset. The house price dataset, containing 12 attributes, was allocated randomly for 70% training and 30% testing.

D. Implementation of XGBoost

Extreme Gradient Boosting (XGBoost) is an enhanced gradient boosting machine using the tree ensemble boosting process. This process ends in the sum of the outputs from all the trees. The XGBoost algorithm used the XGBoost package in Python to evaluate house prices. The sample data allocation scheme used in this model is the same as the previous model which is LightGBM algorithm. Analysis of the feature importance with XGBoost model shows the rank of features (in sequence, highest to lowest): size of the house, the number of parking lots provided for a house, and the nearest distance to the public transport feature. The selection of the features is quite different compared to the previous two models (Multiple Linear Regression and Ridge Regression).

V. Model Deployment

Based on the analysis results, it was recommended to deploy the XGBoost model for the prediction of Kuala Lumpur house prices using new data. This model was being executed by removing the actual house price to generate the predicted house price. To perform this task, a new dataset consists of 1300 samples of Kuala Lumpur house price was collected and our data audit results show that there are no missing values in the sample. The predicted house price was then compared with the actual house price, and the percentage difference between these two values was calculated. Fig. 7 shows the sample of the dataset which had been executed using this model. As illustrated in Fig. 7, the percentage difference for actual and predicted log price is relatively small, which indicated that this model could provide very accurate results. The percentage difference is a bit high when the log price is being inversed to get the real price.

Fig. 8 shows a few rows that contain very high values in percentage difference between predicted and actual house prices. Based on these figures, a high value in percentage difference, which is above 60%, is coming from high-end locations, for example, KLCC and KL City. It can be concluded that a house in a high-end location is difficult to be priced as it might be below or above market value. A group of highest percentage difference value which is above 80% is usually because the price of a house is under market value and is located at KLCC which is a high-end location.

Location	Price	NewSize	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGB3	Difference	Percentage
KLCC	RM 3,843,000.00	1098	RM 1,302,591.00	-RM 2,540,409.00	98.7	14.079867	-1.082	7.4
City Centre	RM 6,000,000.00	3200	RM 2,211,152.00	-RM 3,788,848.00	92.3	14.609025	-0.998	6.6
KLCC	RM 4,756,000.00	1399	RM 1,781,539.00	-RM 2,974,461.00	91.0	14.392989	-0.982	6.6
KLCC	RM 6,441,600.00	2013	RM 2,526,425.00	-RM 3,915,175.00	87.3	14.742316	-0.936	6.2
KLCC	RM 6,500,000.00	2013	RM 2,571,892.00	-RM 3,928,108.00	86.6	14.760153	-0.927	6.1
KLCC	RM 6,300,000.00	2271	RM 3,194,441.00	-RM 3,105,559.00	65.4	14.976923	-0.679	4.4
KLCC	RM 585,000.00	750	RM 1,090,090.00	RM 505,090.00	60.3	13.9017725	0.622	4.6
KL City	RM 300,000.00	900	RM 557,978.00	RM 257,978.00	60.1	13.232077	0.621	4.8
Mont Kiara	RM 650,000.00	1598	RM 1,182,197.00	RM 532,197.00	58.1	13.982886	0.598	4.4
KLCC	RM 2,200,000.00	4520	RM 3,994,731.00	RM 1,794,731.00	57.9	15.200487	0.597	4.0
Bukit Jalil	RM 790,000.00	2500	RM 1,432,498.00	RM 642,498.00	57.8	14.174932	0.595	4.3
Mont Kiara	RM 1,500,000.00	1200	RM 833,534.00	-RM 666,466.00	57.1	13.633431	-0.588	4.2
Jalan Klang Lama (Old Klang Road)	RM 883,000.00	876	RM 493,809.00	-RM 389,191.00	56.5	13.109907	-0.581	4.3
KLCC	RM 635,000.00	913	RM 1,130,136.00	RM 495,136.00	56.1	13.93785	0.576	4.2
Cheras	RM 715,000.00	750	RM 404,165.00	-RM 310,835.00	55.5	12.909581	-0.570	4.3
Sentul	RM 275,000.00	973	RM 485,703.00	RM 210,703.00	55.4	13.093356	0.569	4.4

Fig. 7. Sample of Dataset Executed using XGBoost Model.

Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGB	Differen	Percentage
Mont Kiara	RM 1,850,000.00	RM 1,462,908.00	-RM 387,092.00	23.4	14.195938	-0.235	1.6
KLCC	RM 1,783,000.00	RM 1,278,865.00	-RM 504,135.00	32.9	14.061484	-0.332	2.3
Dutamas	RM 1,180,000.00	RM 1,085,255.00	-RM 94,745.00	8.4	13.897326	-0.084	0.6
Bukit Jalil	RM 900,000.00	RM 1,075,786.00	RM 175,786.00	17.8	13.888563	0.178	1.3
Mont Kiara	RM 3,100,000.00	RM 2,769,086.00	-RM 330,914.00	11.3	14.834028	-0.113	0.8
Desa ParkCity	RM 1,050,000.00	RM 1,394,852.00	RM 344,852.00	28.2	14.1483	0.284	2.0
Mont Kiara	RM 2,600,000.00	RM 2,240,366.00	-RM 359,634.00	14.9	14.62215	-0.149	1.0
Bukit Jalil	RM 978,000.00	RM 1,129,814.00	RM 151,814.00	14.4	13.937565	0.144	1.0
Mont Kiara	RM 1,630,000.00	RM 1,533,811.00	-RM 96,189.00	6.1	14.243267	-0.061	0.4
KLCC	RM 1,250,000.00	RM 1,205,892.00	-RM 44,108.00	3.6	14.002731	-0.036	0.3
Dutamas	RM 1,035,000.00	RM 1,218,741.00	RM 183,741.00	16.3	14.01333	0.163	1.2
KLCC	RM 1,000,000.00	RM 978,442.00	-RM 21,558.00	2.2	13.793718	-0.022	0.2
Mont Kiara	RM 2,480,000.00	RM 2,259,969.00	-RM 220,031.00	9.3	14.630862	-0.093	0.6
Mont Kiara	RM 2,480,000.00	RM 2,418,473.00	-RM 61,527.00	2.5	14.6986475	-0.025	0.2
Bukit Jalil	RM 1,060,000.00	RM 1,272,671.00	RM 212,671.00	18.2	14.056629	0.183	1.3
Bukit Jalil	RM 900,000.00	RM 1,056,281.00	RM 156,281.00	16.0	13.870266	0.160	1.2
KLCC	RM 1,850,000.00	RM 1,853,637.00	RM 3,637.00	0.2	14.432661	0.002	0.0
Sentul	RM 668,000.00	RM 958,553.00	RM 290,553.00	35.7	13.773182	0.361	2.7
KLCC	RM 2,998,000.00	RM 3,102,220.00	RM 104,220.00	3.4	14.947629	0.034	0.2
KLCC	RM 2,580,000.00	RM 2,814,522.00	RM 234,522.00	8.7	14.850304	0.087	0.6

Fig. 8. Sample of Dataset with High Value of Percentage Difference.

Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGB	Differen	Percentage
KLCC	RM6,500,000.00	RM 2,571,892.00	-RM3,928,108.00	86.6	14.760153	-0.927	6.1
KLCC	RM6,441,600.00	RM 2,526,425.00	-RM3,915,175.00	87.3	14.742316	-0.936	6.2
City Centre	RM6,000,000.00	RM 2,211,152.00	-RM3,788,848.00	92.3	14.609025	-0.998	6.6
KLCC	RM6,300,000.00	RM 3,194,441.00	-RM3,105,559.00	65.4	14.976923	-0.679	4.4
KLCC	RM4,756,000.00	RM 1,781,539.00	-RM2,974,461.00	91.0	14.392989	-0.982	6.6
KLCC	RM3,843,000.00	RM 1,302,591.00	-RM2,540,409.00	98.7	14.079867	-1.082	7.4
City Centre	RM5,999,000.00	RM 3,614,291.00	-RM2,384,709.00	49.6	15.100407	-0.507	3.3
Mont Kiara	RM5,000,000.00	RM 3,021,380.00	-RM1,978,620.00	49.3	14.921225	-0.504	3.3
KLCC	RM6,385,950.00	RM 4,507,802.00	-RM1,878,148.00	34.5	15.321321	-0.348	2.2
Bangsar	RM6,480,000.00	RM 4,920,957.00	-RM1,559,043.00	27.3	15.409014	-0.275	1.8
Ampang Hilir	RM5,800,000.00	RM 4,317,390.00	-RM1,482,610.00	29.3	15.278162	-0.295	1.9
Bangsar	RM5,115,000.00	RM 3,646,065.00	-RM1,468,935.00	33.5	15.109159	-0.339	2.2
Bangsar	RM6,600,000.00	RM 5,144,249.00	-RM1,455,751.00	24.8	15.45339	-0.249	1.6
Bangsar	RM6,660,000.00	RM 5,313,014.00	-RM1,346,986.00	22.5	15.48567	-0.226	1.4
Ampang Hilir	RM6,350,000.00	RM 5,081,078.00	-RM1,268,922.00	22.2	15.441034	-0.223	1.4
Bangsar	RM6,450,000.00	RM 5,203,855.00	-RM1,246,145.00	21.4	15.4649105	-0.215	1.4

Fig. 9. Sample of Dataset with High Value of Negative Percentage Difference.

Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGB	Differen	Percentage
KLCC	RM2,200,000.00	RM 3,994,731.00	RM1,794,731.00	57.9	15.200487	0.597	4.0
KLCC	RM2,500,000.00	RM 3,940,634.00	RM1,440,634.00	44.7	15.186852	0.455	3.0
KLCC	RM1,680,000.00	RM 2,700,143.00	RM1,020,143.00	46.6	14.808816	0.475	3.3
Mont Kiara	RM1,800,000.00	RM 2,797,306.00	RM 997,306.00	43.4	14.844168	0.441	3.0
Bukit Tunku (Kenny Hills)	RM2,300,000.00	RM 3,277,969.00	RM 977,969.00	35.1	15.002735	0.354	2.4
KLCC	RM1,500,000.00	RM 2,439,551.00	RM 939,551.00	47.7	14.707325	0.486	3.4
Mont Kiara	RM1,300,000.00	RM 2,209,378.00	RM 909,378.00	51.8	14.608222	0.530	3.7
KLCC	RM2,700,000.00	RM 3,566,958.00	RM 866,958.00	27.7	15.087224	0.278	1.9
KLCC	RM2,700,000.00	RM 3,522,552.00	RM 822,552.00	26.4	15.074697	0.266	1.8
KLCC	RM2,413,375.00	RM 3,179,384.00	RM 766,009.00	27.4	14.9721985	0.276	1.9
KLCC	RM2,357,250.00	RM 3,117,411.00	RM 760,161.00	27.8	14.952514	0.280	1.9
KLCC	RM2,700,000.00	RM 3,457,221.00	RM 757,221.00	24.6	15.055976	0.247	1.7
Bangsar	RM2,700,000.00	RM 3,449,481.00	RM 749,481.00	24.4	15.053735	0.245	1.6
Mont Kiara	RM2,200,000.00	RM 2,937,095.00	RM 737,095.00	28.7	14.892932	0.289	2.0
Bukit Tunku (Kenny Hills)	RM2,200,000.00	RM 2,936,515.00	RM 736,515.00	28.7	14.892735	0.289	2.0
KLCC	RM2,357,250.00	RM 3,083,831.00	RM 726,581.00	26.7	14.941684	0.269	1.8

Fig. 10. Sample of Dataset with High Value of Positive Percentage Difference.

Fig. 9 and 10 illustrate that the highest negative and positive difference in predicted and actual price is also coming from a high-end location in Kuala Lumpur, which is KLCC. A group of high positive difference which is colored in darker green and a group of high negative difference which is colored in darker red is also coming from similar locations which are KLCC, Bangsar, Ampang Hilir, and Bukit Tunku. These locations are considered as high-end locations in Kuala Lumpur, thus the process of pricing a house might be difficult, resulting in the pricing of below or above market value.

VI. Result and Discussion

A comparison of all four models was conducted to determine which model delivers the most accurate results. The performance metrics used were mean absolute error (MAE), root mean squared error (RMSE) value, and adjusted R squared. The model that recorded the smallest MAE, RMSE value and adjusted R squared value closest to 1 was chosen and then used in the model deployment phase. Model comparison and evaluation results are shown in Table III. The model evaluation results comparing four models which are multiple linear regression, ridge regression, LightGBM, and XGBoost show that the XGBoost model has slightly better performance with the lowest MAE and RMSE values and has adjusted R-squared closest to one which indicated a good fit model compared to multiple linear regression, ridge regression, and LightGBM model. The XGBoost model was chosen as the best predictive model.

The aim of this study is to provide an accurate machine learning model for forecasting condominium house prices in Kuala Lumpur. A machine learning model has been proposed to evaluate the relationship between a dependent variable (housing price) and a series of independent variables (attributes). Multiple linear regression, ridge regression, LightGBM, and XGBoost were used and measured against each other. To evaluate the performance of the model, statistical measures such as mean absolute error and root mean square error was also established. The coefficient of determination (R-squared) was also derived to determine how accurately the model predicted the outcome. The XGBoost model was used in the deployment phase as this model was able to accurately predict house prices in Kuala Lumpur, with the highest coefficient of determination (R-squared), which means this model is best-fitted to the dataset. Even though this model was able to predict house prices with the best coefficient of determination value, however, there were still a high percentage difference in predicted house prices and actual house prices in several rows which is less than 5% of entire dataset. Based on these findings, it can be concluded that several locations which can be categorized as high-end locations such as Mont Kiara and KLCC were difficult to be priced. The price of the house might be below or above market value.

The proposed XGBoost model is the first application of XGBoost to the study of the Kuala Lumpur housing market. The model used in this analysis was able to tackle the problems of the housing market in Kuala Lumpur as the XGBoost model has a better fitting and predictive abilities. The XGBoost model was able to generate results that were more consistent and justifiable than other models used for housing market data. The XGBoost model achieved better predictive ability, with the lowest mean absolute error (MAE) and root mean squared error (RMSE), and adjusted R-squared value closest to 1, which indicates the most accurate model. In addition, consistent model performance was found in the XGBoost model as XGBoost outperformed other models in the training and testing R-squared value. The proposed XGBoost model is, therefore, effective in predicting housing prices, which favor not only future house buyers but also investors and policymakers in the real estate industry. In other words, the proposed model will be used to estimate the selling price of the house and then equate it with the currently offered price to know the actual market conditions.

TABLE III. Model Comparison and Evaluation Result

Model	MAE	MSE	RMSE	R-Sq*	R-Sq**	AdjR*
XGBoost*	0.148	0.039	0.197	0.921	0.912	0.911
LightGBM	0.161	0.044	0.210	0.902	0.899	0.898
MLR	0.181	0.057	0.238	0.872	0.871	0.869
RR	0.195	0.064	0.252	0.855	0.855	0.853

R-Sq* for Training Data; R-Sq** for Testing Data;

VII. Conclusion

This paper demonstrated the use of the advanced ML models on house price prediction based on the Kuala Lumpur housing data. The two most recent ML models, namely LightGBM and XGBoost were implemented and compared with the traditional models, namely Multiple Linear Regression and Ridge Regression. The results showed that the XGBoost model was the most promising with 0.0387 for the MSE and was used in the deployment phase. This model accurately predicts house prices in Kuala Lumpur, with the highest coefficient of determination (R-squared). Future works can include more attributes such as the size of the house and proximity to amenities, which can significantly affect the accuracy of the predicted outputs. In addition, future research might consider other locations in their study as this study is only focused on locations in the Kuala Lumpur region. Future research might expand the area of this research, whether to conduct and include the whole nation in Malaysia. This study can significantly assist in predicting of future house prices and the establishment of real estate policies.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, and the Research Management Centre (RMC), Universiti Teknologi MARA, Malaysia, for the support throughout this research.

References

[1] A. Soy Temür, M. Akgün, and G. Temür, “Predicting housing sales in turkey using arima, lstm and hybrid models,” J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

[2] W. C. Choong, “Statistical Analysis Of Housing Prices In Petaling District Using Linear Functional Model.” UTAR, 2018.

[3] M.H. Yop, “Harga Rumah Mengikut Negeri,” 2018. https://www.data.gov.my/data/ms_MY/dataset/harga-rumah-mengikut-negeri.

[4] R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy, “Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java,” in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct. 2018, pp. 351–358, doi: 10.1109/ICACSIS.2017.8355058.

[5] S. Abdul-Rahman, N. F Kamal Arifin, M. Hanifiah and S. Mutalib, “Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier” International Journal of Advanced Computer Science and Applications(IJACSA), 12(9), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120950

[6] P. H. Damia Abd Samad, S. Mutalib, and S. Abdul-Rahman, “Analytics of stock market prices based on machine learning algorithms,” Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 2, pp. 1050–1058, 2019, doi: 10.11591/ijeecs.v16.i2.pp1050-1058.

[7] N. S. Mohd Shafiee and S. Mutalib, “Prediction of Mental Health Problems among Higher Education Student Using Machine Learning,” Int. J. Educ. Manag. Eng., vol. 10, no. 6, pp. 1–9, 2020, doi: 10.5815/ijeme.2020.06.01.

[8] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Md Ab Malik, Predictive Model of Graduate-On-Time Using Machine Learning Algorithms, vol. 1100, no. September. Springer Singapore, 2019.

[9] Y. F. Chang, W. C. Choong, S. Y. Looi, W. Y. Pan, and H. L. Goh, “Analysis of housing prices in Petaling District, Malaysia using functional relationship model,” Int. J. Hous. Mark. Anal., vol. 12, no. 5, pp. 884–905, Oct. 2019, doi: 10.1108/IJHMA-12-2018-0099.

[10] J. Manasa, R. Gupta, and N. S. Narahari, “Machine Learning based Predicting House Prices using Regression Techniques,” in 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Mar. 2020, pp. 624–630, doi: 10.1109/ICIMIA48430.2020.9074952.

[11] M. Thamarai and S. P. Malarvizhi, “House Price Prediction Modeling Using Machine Learning,” Int. J. Inf. Eng. Electron. Bus., vol. 12, no. 2, pp. 15–20, 2020, doi: 10.5815/ijieeb.2020.02.03.

[12] H. Wu et al., “Influence factors and regression model of urban housing prices based on internet open access data,” Sustain., vol. 10, no. 5, pp. 1–17, 2018, doi: 10.3390/su10051676.

[13] J. B. H. Yap and X. H. Ng, “Housing affordability in Malaysia: perception, price range, influencing factors and policies,” Int. J. Hous. Mark. Anal., vol. 11, no. 3, pp. 476–497, Jun. 2018, doi: 10.1108/IJHMA-08-2017-0069.

[14] A. Abdullahi, Usman, H. Usman & I. Ibrahim, "Determining house price for mass appraisal using multiple regression analysis modeling in Kaduna North, Nigeria," ATBU Journal of Environmental Technology, 11(1), 26-40, 2018.

[15] T. Mohd, S. Masrom, and N. Johari, “Machine learning housing price prediction in petaling jaya, Selangor, Malaysia,” Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 11, pp. 542–546, 2019, doi: 10.35940/ijrte.B1084.0982S1119.

[16] A. Jafari and R. Akhavian, “Driving forces for the US residential housing price: a predictive analysis,” Built Environ. Proj. Asset Manag., vol. 9, no. 4, pp. 515–529, Sep. 2019, doi: 10.1108/BEPAM-07-2018-0100.

[17] A. Nur, R. Ema, H. Taufiq, and W. Firdaus, “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 323–326, 2017, doi: 10.14569/ijacsa.2017.081042.

[18] J. H. Chen, C. F. Ong, L. Zheng, and S. C. Hsu, “Forecasting spatial dynamics of the housing market using Support Vector Machine,” Int. J. Strateg. Prop. Manag., vol. 21, no. 3, pp. 273–283, 2017, doi: 10.3846/1648715X.2016.1259190.

[19] Pai, P.F. and Wang, W.C., "Using machine learning models and actual transaction data for predicting real estate prices". Applied Sciences, 10(17), p.5832, 2020.

[20] T. D. Phan, “Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia,” in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Dec. 2018, pp. 35–42, doi: 10.1109/iCMLDE.2018.00017.

[21] Q. Truong, M. Nguyen, H. Dang, and B. Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” Procedia Comput. Sci., vol. 174, no. 2019, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[22] K. Zhang, L. Shen, and N. Liu, “House Rent Prediction Based on Joint Model,” in Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, Oct. 2019, pp. 507–511, doi: 10.1145/3373509.3373578.

[23] Y. Zhou, “Housing Sale Price Prediction Using Machine Learning Algorithms,” 2020, [Online]. Available: https://escholarship.org/uc/item/0th2s0ss.

[24] P. Katherina, "Data and Methodology", How to Write About Economics and Public Policy, pp 241-270, 2018.

[25] Watkins, M. W. A Step-by-Step Guide to Exploratory Factor Analysis with SPSS. Routledge, 2021

[26] Farmer, Antoinette Y., Antoinette Y. Farmer, and G. Lawrence Farmer. Research methods for social work: A problem-based approach. SAGE Publications, 2020.

[27] J. W. Osborne and E. Waters, “Four assumptions of multiple regression that researchers should always test,” Pract. Assessment, Res. Eval., vol. 8, no. 2, pp. 2002–2003, 2002.

[28] N. E. Jeremia, S. Nurrohmah and I. Fithriani, "Robust Ridge regression to solve a multicollinearity and outlier", J. Phys.: Conf. Ser. 2020

[29] N. Deepa et al., “An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier,” J. Supercomput., vol. 77, no. 2, pp. 1998–2017, Feb. 2021, doi: 10.1007/s11227-020-03347-2.

[30] W. Xu, X. Liu, F. Leng, and W. Li, “Blood-based multi-tissue gene expression inference with Bayesian ridge regression,” Bioinformatics, vol. 36, no. 12, pp. 3788–3794, Jun. 2020, doi: 10.1093/bioinformatics/btaa239.

[31] Y. Yang and Y. Yang, “Hybrid prediction method for wind speed combining ensemble empirical mode decomposition and bayesian ridge regression,” IEEE Access, vol. 8, pp. 71206–71218, 2020, doi: 10.1109/ACCESS.2020.2984020.

[32] C. Fan, Z. Cui, and X. Zhong, “House Prices Prediction with Machine Learning Algorithms,” in Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Feb. 2018, pp. 6–10, doi: 10.1145/3195106.3195133.

[33] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, “A hybrid regression technique for house prices prediction,” IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2017-December, no. August 2018, pp. 319–323, 2018, doi: 10.1109/IEEM.2017.8289904.

[34] L. Zhu and L. Li, “Comparison of Regression Models on House Value Prediction,” 2020.

[35] H. Zeng et al., “A LightGBM-Based EEG Analysis Method for Driver Mental States Classification,” Comput. Intell. Neurosci., vol. 2019, 2019, doi: 10.1155/2019/3761203.

[36] C. Chen, Q. Zhang, Q. Ma, and B. Yu, “LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion,” Chemom. Intell. Lab. Syst., vol. 191, no. June, pp. 54–64, 2019, doi: 10.1016/j.chemolab.2019.06.003.

[37] Y. Song et al., “Prediction of Double-High Biochemical Indicators Based on LightGBM and XGBoost,” in Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Jul. 2019, pp. 189–193, doi: 10.1145/3349341.3349400.

[38] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[39] J. Eronen, “Housing unit price prediction system.” 2018.

[40] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, “House price prediction using a machine learning model: A survey of literature,” Int. J. Mod. Educ. Comput. Sci., vol. 12, no. 6, pp. 46–54, 2020, doi: 10.5815/ijmecs.2020.06.04.

[41] J. J. Wang et al., “Predicting House Price with a Memristor-Based Artificial Neural Network,” IEEE Access, vol. 6, pp. 16523–16528, 2018, doi: 10.1109/ACCESS.2018.2814065.

[42] M. F. Mukhlishin, R. Saputra, and A. Wibowo, “Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor,” in 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Nov. 2017, pp. 171–176, doi: 10.1109/ICICOS.2017.8276357.

[43] A. Varma, A. Sarma, S. Doshi, and R. Nair, “House Price Prediction Using Machine Learning and Neural Networks,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Apr. 2018, pp. 1936–1939, doi: 10.1109/ICICCT.2018.8473231.

[44] R. Reed, “The relationship between house prices and demographic variables,” Int. J. Hous. Mark. Anal., vol. 9, no. 4, pp. 520–537, Oct. 2016, doi: 10.1108/IJHMA-02-2016-0013.

[45] A. Kallner, “Formulas,” in Laboratory Statistics, A. B. T.-L. S. (Second E. Kallner, Ed. Elsevier, 2018, pp. 1–140.

[46] J. F. Hair, W. C. Black, B. J. Babin and R. E. Anderson (2019). Multivariate data analysis, 2019.

[47] Byrne, B.M. (2016). Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Third Edition (3rd ed.). Routledge. https://doi.org/10.4324/9781315757421, 2016.

[48] G. S. K. Ranjan, Verma, A. K. and R. Sudha. "K-nearest neighbors and grid search cv based real time fault monitoring system for industries." In 2019 IEEE 5th international conference for convergence in technology (I2CT), pp. 1-5. IEEE, 2019.

[49] J. Macaluso, “Testing Linear Regression Assumptions in Python.,” 2018. https://jeffmacaluso.github.io/post/LinearRegressionAssumptions/.

Prediction of Tourist Visit in Taman Negara Pahang, Malaysia using Regression Models

Sofianita Mutalib*, Athila Hasya Razali, Siti Nur Kamaliah Kamarudin, Shamimi A Halim, Shuzlina Abdul-Rahman
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, 40450 Shah Alam,
Selangor, Malaysia

***Abstract*—Tourism is among the significant source of income to Malaysia and Taman Negara Pahang is one of the Malaysia's tourism spots and the heritage of Malaysia in achieving the Sustainable Development Goals (SDG). It has attracted many international and local tourists for its richness in flora and fauna. Currently, the information of tourists' visits is not properly analyzed. This study integrates the internal and public information to analyze the visits. The regression models used are multiple linear regression, support vector regression, and decision tree regression to predict the tourism demand for Taman Negara, Malaysia and the best model was deployed. Predictive analytics can support the decision-making process for tourism destinations management. When the management gets a head-up of the demand in the future, they can choose a strategic planning and be more aware about the factors influencing tourism demand, such as the tourists' web search engine behaviors for accommodation, facilities, and attractions. The factors affecting the tourism demand are determined as the first objective. The role of independent variable was set to the total number of visitors, subsequently being set as the target variable in the modeling process. A total of 30 models were generated by tuning the cross-validation parameters. This study concluded that the best model is the multiple linear regression due to lower root mean square error (RSME) value.**

Keywords*—Regression models; SDG; Taman Negara Pahang; tourist analytics

I. Introduction

Attractions of tourism destinations produce economic values as it impacts the number of tourist arrivals in Malaysia. As a result, tourism has become one of the highest revenue industries after automobiles and oil (REF). Today, one of the largest service sectors is the tourism industry where the industry became one of the highest in terms of revenue after automobiles and oil. This significant growth is a result of the efforts undertaken by the Ministry of Tourism where the planning and execution policy underlined by the government spearhead the success of the tourism industry. It is the long-term aim of the government to make Malaysia as one of the most popular tourism destinations. The success of the tourism industry is defined by the demand and supply which can be measured by tourists arrivals and receipts [1].

One of the famous tourism destinations is Taman Negara Malaysia, also known as Taman Negara National Park. This natural park protects a diverse flora and fauna, renowned for its nature trails, and adventure activities hence making it a valuable tourism source [2]. Encompassing an area of 4,343 km2, Taman Negara National Park straddles three states of Malaysia; Taman Negara Pahang, Taman Negara Kelantan, and Taman Negara Terengganu; in which Taman Negara Pahang takes up around 57% of the total national park area [3]. There are two different main entrances for Taman Negara Pahang namely, Kuala Tahan and Sungai Relau. There are many activities to do in Taman Negara such as jungle trekking, hiking, and fishing. Panoramic scene and captivating places such as waterfall cascades and canopy walkway attract many people to visit the Taman Negara National Park. Therefore, the service industry needs to be concerned about visitors' management of the tourism destination. In 2013, a research study was made about Ecotourism in Taman Negara National Park: the issues and challenges [4]. One of the issues that they found is the lack of visitor management especially on overcrowding problem and excess visitors during certain period of time. Moreover, based on the news [5], it was pointed out early on those statistics shows of slightly higher number of visitors to Taman Negara in January to February. The lack of proper service management on popular places at the park lead to overcrowding problem, mainly due to the inability to optimize staff's workload/working hours. In addition, overcrowding of tourists lead to the loss of authenticity and implies a significant risk to the destination's future attractiveness, especially towards vulnerable destinations such as the Taman Negara National Park.

The attractiveness of Taman Negara Malaysia as a tourism destination has been studied by Universiti Putra Malaysia [6]. The study evaluated that there are total of thirteen attractions in Taman Negara, namely oral history, local culture and lifestyle, flora, fauna, building architecture, nature trails, shopping opportunity, canopy walkway, caves, stream, fishing, mountain, and adventure activities [6]. The attractions of tourism destinations produce economic values as it also gives an impact to the number of tourist arrivals in Malaysia. These attractions have been characterized as demand structures. Thus, demand studies are needed for decision making support of tourism destination management.

Therefore, tourism demand forecasting is vital and will benefit the nation's tourism industry greatly. The study of regression techniques helps to forecast future and seasonal demands for tourism growth, management, and planning purposes. Regression analysis is a collection of statistical methods for estimating relationships between a dependent variable and one or more independent variables which can be used to determine the strength of a relationship between

*Corresponding Author.

variables and to predict how they will interact in the future [7]. This study was made to determine the factors affecting the tourism demand, to make a comparison study of regression techniques, and develop an analytical dashboard based on the best regression model that helps in forecasting tourism demands components, incorporating the applicable criteria in the following sub-sections.

A. Factors Affecting Tourism Visit

In demand studies, factors affecting tourism demand was determined as the independent variables that might affect the target variable, where the target variable is the total number of visitors. The comparison study was made to find the best model in predicting the target variable. Tourism industry has been a huge contribution factor to the economy, even though there exist many factors affecting tourism demand in Malaysia. Several studies have shown that the economic variables play an important role as the key economic factors. Based on [8] and [9], the key economic factors are exchange rate, income, consumer price index (CPI), and population of the country. The studies found a strong relationship between these key economic factors and the volume of tourists. However, the study also showed that there is a negative correlation between the exchange rate and the number of tourists, the higher the exchange rate, the lower the volume of tourists' arrivals. The exchange rate is related to the depreciation of Ringgit Malaysia (RM) that affects the cost of living in Malaysia.

As many researchers used economic factors as the contribution to the demands, some researchers [10-12] observed the tourist's web search behavior by using Google Trends data. Specific keywords were identified for web scraping related to tourism activities, such as "skiing", "skiing in sweden" or "sweden skiing" or "ski sweden" [10]. Meanwhile, [12] made use of these keywords: "destination", "destination + guide", "destination + travel guides", "destination + tickets", "destination + weather", depending on the tourism destination, strategy, ticket price, scenic spots, weather, and accommodation, among many other factors. At the end, 50 initial keywords related to the decision-making process were selected [11]. Due to the time gaps between web search activity and tourist arrivals, this new approach is truly relevant. Web-based data sources, such as search engine traffic, often have a natural relationship with tourism demand. Because of strong interest in certain tourism destinations, potential tourists for instance, browse websites extensively before visiting these destinations.

In other study, climate change is an additional factor to know the relationship with tourism demand. Research by [13] was conducted regarding the dimension of climate change in Malaysia based on tourists' perception. Generally, Malaysia has an equatorial climate. Extreme weather and seasonality are commonly related with climate change in Malaysia. Temperature, rainfall, and, to a smaller extent, wind are all examples of extreme weather factors. Seasonality, on the other hand, is always linked to the dry and wet (monsoon) seasons. The main variables for this factor are the average temperature and the average precipitation. The weather in Malaysia is hot and humid all year with Malaysia's average daily temperature ranges from 21°C to 32°C. Precipitation is the measure of the falling water from the sky which is the rainfall. The findings of the study revealed that tourists had sufficient knowledge of climate change, which influence their travel decisions [13].

B. Prediction in Tourist Domain

Many past research studies have been conducted in order to predict the tourism demand. A comparative study made by [14] to forecast the tourism demand in Turkey using data mining techniques based on regression modelling. The techniques used include multiple linear regression (MLR), multilayer perceptron (MLP) regression, and support vector regression (SVR). The author in [14] used monthly data points for their study, unlike previous studies which usually uses yearly or quarterly data. The author in [14] decided to choose these regression models because of the nonlinear mapping capabilities. In addition, the conventional methods like these models are more efficient to use for data that are likely to pattern the trends, seasonality, and cyclicality. SVR originated from a machine learning model hence a support-vector machine (SVM) can work for regression tasks and is suggested in order to forecast tourism demand. Unlike most conventional neural network model, SVR applies the theory of structural risk minimization which based on the idea of empirical risk minimization, to minimize the upper limit of the generalization error, instead of minimizing the error in training [15-17].

A research on tourists visit in the province of West Sumatra was done using MLR and Artificial Neural Network (ANN) [25] with inflation rates and Rupiah exchange rates. The results show an impressive accuracy within 96 to 99 percent. Other researchers from Indonesia made use of seven independent variables including the characteristics of foreign tourists (sex, age, occupation/profession, length of stay, nationality, purpose of visit, and accommodation) to identify the effect on total expenditure [26]. They also found that American and European tourists contributed to the largest average of the total expenditure for vacation purpose. Regression tree was used by [27] to segment the tourist length of stay in Barbados, with socio-demographic profile of the tourist, trip-related characteristics, distance, and economic conditions in the source country. Another study on tourism to model the revenue from international tourism using the foreign trade balance of the country shows the positive correlation in Azerbaijan example [28]. The result also showed that tourism will increase the country's foreign trade turnover. More advanced methods were explored in tourism domain in China [29] with social evaluation index as the attributes and hybrid methods of back propagation and fuzzy as the model.

Another angle on predicting tourism trend is by looking at the flow of tourists' movement in a specified area, as studied by [30-31]. Usage of user-generated content assisted in narrowing down specific criteria to forecast tourism demand apart from projecting possible point-of-interest for tourists [30]. By creating trajectory graphs on past data, the study by [30] yield a better result than traditional machine-learning based algorithms for forecasting the next tourist movement, which is useful for predicting tourism demand in certain areas. While the study by [31] focuses more on statistical-based techniques, they applied statistical method with BPNN model (SMBPNN) on variously collected data such as historical tourism flow, weather, and temperature. Their hybrid model which combines statistical technique with neural network

suggested an improvement of forecast as compared to stand-alone neural network models [31].

The methodology used by other researcher suggested using various techniques and at the same time, incorporating a multi-dimensional dataset. Hence for this study, an integrated dataset were created from various sources that will be described later in this paper. The suitability of the techniques with the available real-world dataset was also considered; hence this study will be focusing on application of regression techniques.

In the rest of the paper, we first provide some necessary background of proposed modelling methods in Section II, and the data source and evaluation are discussed in Section III. The experimental results are demonstrated and discussed in Section IV. Finally, the conclusion and future works are explained in Section V.

II. Modelling Methods for Regression

Regression analysis is used for predicting real values, for instance, to forecast the daily sales of the business which makes the number of sales as the target or dependent variable. To determine the relationship between the variables of interest, the data collected will be trained using a regression model. In order to determine the optimal model, Goodness-of-fit measurements, such as the square of the correlation coefficient (r^2 or squared correlation), was used to examine the scatter of data points around the fitted value. The number denoted the percentage of variance in one variable that can be explained by the other. The higher the r^2 score, the more precise the prediction. However, the number does not tell how precise the forecasts were in the dependent variable's units.

A. Multiple Linear Regression

Multiple Linear Regression (MLR) aims to model the linear relationship between the independent (explanatory) variables and dependent (response) variables, whereas simple linear regression only has a single input to estimate the value of the coefficients used in data representation. MLR model helps to predict an outcome based on multiple explanatory variables provided with details [18]. The representation of multiple linear regression will be like the following:

$$Y = a_0 + a_1x_1 + \ldots + a_nx_n \quad (1)$$

In which Y refers to the dependent variables, x_1, …, x_n represent independent variables, a_1, …, a_n as regression coefficients, and a_0 is y-intercept (a constant term).

By measuring slope and regression coefficients, it can be represented in the form of mathematical equations in multiple linear regression. Using the regression coefficient formula, the intensity and direction of the relationship between the two variables can be calculated [19]. Hence when comparing with simple linear regression, MLR perform better with less error rate. Moreover, multiple regression can be implemented in linear and non-linear modelling. Multiple regression is based on the statement that there is a linear relationship between both dependent and independent variables, where no assumption was made for major correlation between the independent variables [18-20].

B. Support Vector Regression

Support-vector regression (SVR) comes from a machine learning model namely the support-vector machine (SVM). SVM can work for regression tasks and is suggested in order to forecast tourism demand. SVM is a class of linear algorithms that can be used for classification, regression, density estimation, novelty detection, and other applications. SVM uses classification techniques to build a predictive model where its algorithm's main purpose is to find a hyperplane in an N-dimensional space that distinctly classify the data points. Separating two classes of data points may lead to many possible hyperplanes. The hyperplane equation is reflected in (2) and in Fig. 1:

$$\vec{w} \cdot \vec{x} + b = 0 \quad (2)$$

where w is a weight vector, x is input vector and b is bias. SVM searches for the hyperplane with the largest margin in separating the circle objects and square objects with minimum classification error.

An optimal network structure can be achieved by SVR on the basis of the theory of structural risk minimization which the margin of the hyperplane will be maximized [15, 21]. The main differences between Least Square Support Vector Machine (LSSVM) and SVM is that LSSVM includes the equality of constraints instead of the inequalities, and it is based on the least squares cost function [22]. SVR has been successfully implemented to solve forecasting problems in many fields, such as financial time series (stock index and exchange rate) forecasting, engineering and software (production values and reliability) forecasting, atmospheric science forecasting, electric load forecasting and commodity demand forecasting [16].

The SVR model has been successfully applied to forecast tourist arrivals too. Empirical research has shown that the choice of the parameters in an SVR model significantly influences the accuracy of forecasting. SVR solves the problems of estimation, classification, and nonlinearity via its loss function [17]. Tourism data often exhibit nonlinear characteristics, with SVR widely used in the tourism demand forecast. Low speed, however, is the key drawback of SVR in the training process [23], due to various hyperparameters setting in the model. Some inappropriate SVR parameters allow for the occurrences of overfitting or underfitting problem.

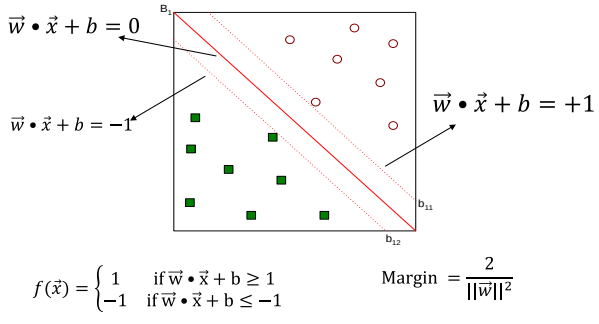


Fig. 1. The Hyperplanes for Two Classes of Objects [21].

C. Decision Tree Regression

Decision tree algorithm that has been used for classification or regression predictive modelling problems is called Classification and Regression Trees (CART). Decision Tree is one of the classifiers in supervised learning algorithm with a tree-like structure. It consists of root, interior, and leaf nodes in which the outcomes are represented at leaf node. CART is relatively straight-forward for prediction making. The algorithm works through multiple iterations until the tree is able to predict a proper value for the data point. Among the benefits of using CART algorithm is that it is easy to understand, less data cleaning process, non-linearity does not affect the output of the model, and the number of hyper-parameters to be tuned is almost null [21, 24]. The drawback is that it may have an overfitting problem, but which can be solved using the Random Forest algorithm.

The split attribute in the tree is chosen based on the standard deviation for the independent variable and dependent variable (outcome), and the formula for the standard deviation based on an attribute *x* as shown in equation (3):

$$S(x) = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \quad (3)$$

where, S is standard deviation, x_i is the value of the attribute, $\overline{x}$ is the mean value of x and n is the record of x.

The standard deviation reduction is based on the decrease in standard deviation after a dataset is split on an attribute. Construction of a decision tree is basically finding an attribute that have the highest standard deviation reduction (SDR), where according to the calculation in equation (3) as the most homogeneous branch. In other words, the standard deviation of the target will be compared to different standard deviation of each independent variables in the dataset.

III. Data Source And Model Evaluation

A. Data Preparation

The data collection is obtained from different sources which are opened to both public and private use. This research study used monthly data points with observation period from year 2012 until 2018. The first dataset collected was from Google Trends, which is a search trends feature that displays the frequency with which a certain search term is entered into Google's search engine in relation to the site's total search volume over time. There are six keywords or search term used for this study: 1) "Taman Negara", 2) "taman negara accommodation", 3) "taman negara resort", 4) "taman negara canopy walkway", 5) "Cuti Cuti Malaysia", and 6) "visit Malaysia".

The second dataset was retrieved from the Federal Reserve Bank of St. Louis (FRED) where the dataset recorded monthly currency exchange rate with units of Malaysian Ringgit to One U.S. Dollar. The third dataset was extracted from Visual Crossing website that provides historical weather public data. To generate the worldwide weather observation database, the website processes millions of hourly weather observations from thousands of observation stations. However, the dataset extracted from the website is in weekly frequency. Hence, the need to calculate the monthly frequency of average attributes were done using the formula of average in Microsoft Excel. The attributes extracted from the source are Location, Date and Time, Maximum Temperature (degC), Minimum Temperature (degC), Temperature (degC), Heat Index (degC), Chance Precipitation (%), Precipitation (mm), Wind Speed (kph), Wind Direction, Wind Gust (kph), Visibility (km), Cloud Cover, Relative Humidity, and Conditions. However, for this research process, only two attributes were selected: 1) Temperature (degC); 2) Precipitation (mm).

The last and most important dataset is the total visitors' arrival to Taman Negara Pahang, Malaysia. This public dataset was found in a Malaysia Open Data Portal at data.gov.my. However, the data is in a yearly term basis and only covers data from the year 2012 until 2018. Thus, an additional data obtained from Jabatan Perlindungan Hidupan Liar dan Taman Negara (PERHILITAN) Pahang was used to identify the monthly number of visitors to Taman Negara Pahang. Since the data collected for this research were huge and obtained from different sources hence making it unstructured, data transformation process needed to be done. Data preprocessing is the process of changing the variety of raw dataset into one dataset suitable to be used in software such as RapidMiner. For this study, the regression modelling was done with the help of data analytics software tool, RapidMiner. Fig. 2 shows the conceptual framework for the regression modelling that illustrated the source of data, independent variables, and dependent variables. The variables for the observation period (year 2012-2018) were Date, Currency Exchange, Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort, Taman Negara canopy walkway, Average Temperature, Average Precipitation and Total of visitors.

B. Model Evaluation

For this study, a comparison was made between three predictive models, namely Multiple Linear Regression (MLR), Support-Vector Regression (SVR), and Decision Tree Regression (DTR), as shown in Fig. 3. Among the important tasks conducted were data preparation from four data sources: data preprocessing, training and testing data split, modeling with three algorithms, and finally evaluation and deployment. The model development in this research study used RapidMiner's default values. The comparison was measured based on the models' performances by manipulating the sampling type and number of folds in cross validation.

The RMSE (Root Mean Squared Error) is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. While squared_correlation is a relative measure of fit, RMSE is an absolute measure of fit. Thus, the lower the values of RMSE, the better it is. The formula for Root Mean Square Error (RSME) is as in (4),

$$\sqrt{\sum_{i=b}^{n}(X_{actual} - X_{model})^2} \quad (4)$$

where, x_{actual} is an observed value and x_{model} is the predicted value.

Source	Independent Variables		Dependent Variables
FRED	Currency Exchange		
Google Trends	Keywords: Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort, Taman Negara canopy walkway,	Regression Model	Total of visitors
Visual Crossing	Weather: Average Temperature, Average Precipitation		

[a] FRED: https://fred.stlouisfed.org/categories/15

[b] Visual Crossing: https://www.visualcrossing.com/

Fig. 2. Conceptual Framework of Tourist Visit Regression Model.

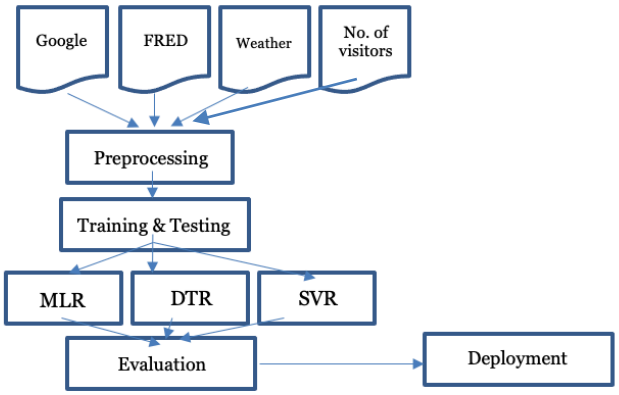


Fig. 3. The Overview of Processes Involved in the Study.

Another useful measure to determine the optimal model is goodness-of-fit measurements, such as the square of the correlation coefficient (r^2 or squared_correlation). This measure is used to examine the scattered-locations of data points around the fitted value. The number denotes the percentage of variance in one variable that can be explained by the other. The higher the r^2 score, the more precise the prediction. It does not, however, tell how precise the forecasts are in the dependent variable's units.

IV. Result and Discussion

This section presents the results and provides a discussion based on the outlined framework given in Fig. 2 and Fig. 3.

A. Data Description

The final dataset was constructed based on the following data sources:

- The Google Trends dataset, in Fig. 4 recorded 504 data which were generated from monthly frequency of year 2012 until year 2018 for six related keywords, as mentioned in section 3 (12 months/year × 7 years × 6 keywords = 504 data).

	A	B	C	D	E
1		Category: All categories	Category: All categories	Category: All categories	Category: All categories
2	Month	Taman Negara: (Malaysia)	taman negara accommodation: (M	taman negara resort: (Malaysia)	taman negara canopy walkway: (Mala
3	2012-01	39	0	39	0
4	2012-02	64	0	20	0
5	2012-03	64	0	26	0
6	2012-04	45	0	35	0
7	2012-05	59	0	75	0
8	2012-06	57	0	42	0
9	2012-07	50	0	16	0
10	2012-08	44	69	60	0
11	2012-09	49	0	16	56
12	2012-10	45	59	15	0
13	2012-11	38	0	7	0
14	2012-12	37	0	35	0
15	2013-01	48	57	35	0
16	2013-02	44	0	16	0
17	2013-03	32	0	32	0
18	2013-04	48	0	19	0
19	2013-05	43	0	18	0
20	2013-06	37	0	38	0
21	2013-07	41	0	30	0
22	2013-08	39	51	32	0

Fig. 4. Google Trend Hits.

- Second dataset, retrieved from FRED (Federal Reserve bank of St. Louis) - the dataset recorded monthly currency exchange rate with units of Malaysian Ringgit to One U.S. Dollar, for each month from 2012, as in Fig. 5.

10	Frequency: Monthly	
11	observation_date	EXMAUS
12	2012-01-01	3.1092
13	2012-02-01	3.0220
14	2012-03-01	3.0444
15	2012-04-01	3.0586
16	2012-05-01	3.0978
17	2012-06-01	3.1783
18	2012-07-01	3.1653
19	2012-08-01	3.1153
20	2012-09-01	3.0758
21	2012-10-01	3.0524
22	2012-11-01	3.0555
23	2012-12-01	3.0537
24	2013-01-01	3.0407
25	2013-02-01	3.0964
26	2013-03-01	3.1074
27	2013-04-01	3.0480
28	2013-05-01	3.0188
29	2013-06-01	3.1433

Fig. 5. Daily Currency Exchange.

- Third dataset, extracted from Visual Crossing website which provides historical weather public data. To generate the worldwide weather observation database, the website processes millions of hourly weather observations from thousands of observation stations. However, the dataset extracted from the website is in weekly frequency.

- Fourth dataset consists of the number of visitors in Taman Negara National Park, including nearby places and Gunung Tahan from Malaysia Open Data Portal by yearly and PERHILITAN Pahang, by monthly, starting from January till December, as in Table I. The highest number of visitors found was in March.

Based on the gathered information, the final dataset consists of 11 variables or attributes with 84 rows of monthly data points for the observation period (year 2012-2018). The variables/columns are observation date (month), Currency exchange, the six keywords: Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort and Taman Negara canopy walkway, Average Temperature, Average Precipitation and Total Visitors. The Fig. 6 shows the snippets of 10 rows from 2012 as the sample of the real dataset used in the study.

TABLE I. THE VISITORS OF TAMAN NEGARA

No	Month	Nearby places		Gunung Tahan		Total
1	January	991	3	58	-	**1052**
2	February	1347	6	50	-	**1403**
3	Mac	1850	46	55	18	**1969**
4	April	823	28	119	-	**970**
5	May	1260	7	106	12	**1385**
6	June	1296	23	157	35	**1511**
7	July	855	19	12	2	**888**
8	August	1479	40	112	-	**1631**
9	September	1387	20	125	-	**1532**
10	October	1361	13	188	-	**1562**
11	November	1080	-	5	-	**1085**
12	December	574	-	-	-	**574**
		Total				**15562**

Open Data: https://www.data.gov.my/

observation_da	Currency exhange	visit malaysia: (	Cuti Cuti Malaysia (	Taman Negara:	taman negara acco	taman negar	taman negara c	Average Tempe	Average Preci	Total Visitors
1/1/2012	3.1092	31	39	39	0	39	0	25.44194	15.59032	6722
1/2/2012	3.022	31	36	64	0	20	0	26.44483	2.327586	8642
1/3/2012	3.0444	29	26	64	0	26	0	26.47097	7.383871	11523
1/4/2012	3.0586	36	30	45	0	35	0	26.92	12.03	5761
1/5/2012	3.0978	32	37	59	0	75	0	27.31613	11.03226	8642
1/6/2012	3.1783	31	33	57	0	42	0	27.68667	5.383333	9602
1/7/2012	3.1653	38	30	50	0	16	0	27.12903	3.383871	5761
1/8/2012	3.1153	35	27	44	69	60	0	26.9871	9.070968	9602
1/9/2012	3.0758	44	37	49	0	16	56	27.03333	5.286667	9602
1/10/2012	3.0524	33	41	45	59	15	0	26.47097	5.829032	9602

Fig. 6. Sample of Merged Dataset.

Meanwhile Fig. 7 to 10 show statistical measures of selected attribute. In Fig. 6, the currency exchange is represented in real value, while for the six keywords (Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort and Taman Negara canopy walkway) are represented as the count of the words mentioned every month. The count of word "Taman Negara accommodation" being mentioned were found to be more as compared to other words. The total number of word count in the dataset is as shown in Fig. 10, with "Visit Malaysia" being the highest count and "Taman Negara canopy walkway" as the lowest count.

Name	Type	Missing	Statistics			Filter (11 / 11 attributes):
Currency exhange	Real	0	Min 3.019	Max 4.457	Average 3.697	
visit malaysia: (Worldwide)	Integer	0	Min 25	Max 85	Average 44.202	
Cuti Cuti Malaysia (Worldwide)	Integer	0	Min 16	Max 57	Average 30.048	
Taman Negara: (Malaysia)	Integer	0	Min 29	Max 67	Average 44.083	
taman negara accommodation: ...	Integer	0	Min 0	Max 100	Average 11.464	
taman negara resort: (Malaysia)	Integer	0	Min 5	Max 75	Average 24.024	
taman negara canopy walkway: ...	Integer	0	Min 0	Max 75	Average 6.786	

Fig. 7. Statistical Measures for Currency Exchange Rate and Count of the Keywords Searched in Google.

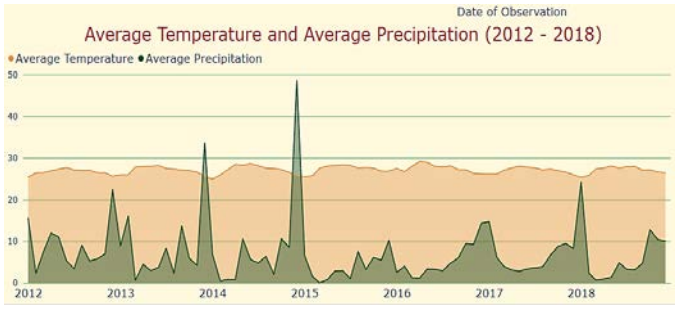


Fig. 8. The Graph for Average Temperature and Average Precipitation from 2012 Till 2018.

In Fig. 8, the average temperature and average precipitation are presented from 2012 till 2018, with gathered information from Visual Crossing. The average temperature is between 25 and 30 degrees Celsius. Here, the precipitation shows the highest value in 2015 and the lowest are shown to be present in every year, due to the dry season that occurs in the respective years.

Fig. 9 illustrates the graph of total visitors in every month to the Taman Negara from 2012 till 2018. The total number of visitors continuously changed over time.

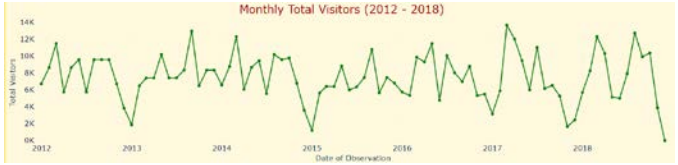


Fig. 9. The Graph for Monthly Total Visitors from 2012 Till 2018.

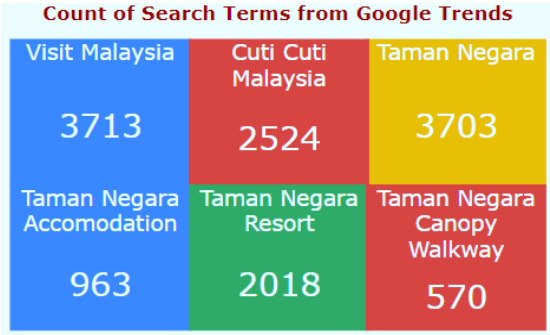


Fig. 10. Count of Words related to Taman Negara.

A total of 30 experiments were performed by tuning five different numbers of folds with two types of sampling in three predictive models (5×2×3 = 30). The comparison table shows the overall performance recorded during the experiment for the modelling.

B. Multiple Linear Regression Results

The best model, multiple linear regression, resulted with the RMSE values of 2545.977 and the values of r^2 is 0.276 which is the highest value with linear sampling among four models, as shown in Table II. Thus, linear regression equation for the best model is as the following:

*Predicted (Total Visitors) = - 631.866 * Currency exchange*
*+ 7.229 * visit malaysia: (Worldwide)*
*- 9.240 * Cuti Cuti Malaysia (Worldwide)*
*+ 113.991 * Taman Negara: (Malaysia)*
*- 7.465 * taman negara accommodation: (Malaysia)*
*- 2.319 * taman negara resort: (Malaysia)*
*+ 5.474 * taman negara canopy walkway: (Malaysia)*
*+ 1062.211 * Average Temperature*
*- 20.455 * Average Precipitation*

TABLE II. LINEAR REGRESSION MODELLING

Sampling method	**RMSE**	r^2
Linear Sampling (k = 10)	2624.972	0.301
Linear Sampling (k = 8)	2545.977	0.276
Shuffled Sampling (k = 10)	2570.164	0.227
Shuffled Sampling (k = 4)	2591.315	0.162

C. Decision Tree Regression Results and Rules

Modeling of decision tree resulted in a set of rules represented in a tree-like structure. Each node corresponds to a splitting rule for a single attribute. Fig. 11 shows the extracted tree and the conditions of each predicted total of tourist visit. Table III shows the result of RMSE and r^2 for the decision tree regression, in which the RMSE is higher as compared to multiple linear regression. As can be seen, average temperature and the count of the keywords appeared to be among the important variables.

```
Average Temperature > 26.984
|   Currency exhange > 4.371: 12859.000 (count=2)
|   Currency exhange ≤ 4.371
|   |   Taman Negara: (Malaysia) > 49.500
|   |   |   taman negara accommodation: (Malaysia) > 38: 6295.000 (count=2)
|   |   |   taman negara accommodation: (Malaysia) ≤ 38
|   |   |   |   visit malaysia: (Worldwide) > 40
|   |   |   |   |   Average Temperature > 27.417
|   |   |   |   |   |   Currency exhange > 4.070: 11782.500 (count=2)
|   |   |   |   |   |   Currency exhange ≤ 4.070
|   |   |   |   |   |   |   Currency exhange > 3.408: 8386.500 (count=2)
|   |   |   |   |   |   |   Currency exhange ≤ 3.408: 9907.000 (count=2)
|   |   |   |   |   Average Temperature ≤ 27.417
|   |   |   |   |   |   taman negara canopy walkway: (Malaysia) > 16: 12665.500 (count=2)
|   |   |   |   |   |   taman negara canopy walkway: (Malaysia) ≤ 16: 11138.000 (count=2)
|   |   |   |   visit malaysia: (Worldwide) ≤ 40
|   |   |   |   |   visit malaysia: (Worldwide) > 35: 6377.000 (count=2)
|   |   |   |   |   visit malaysia: (Worldwide) ≤ 35: 9122.000 (count=2)
|   |   Taman Negara: (Malaysia) ≤ 49.500
|   |   |   visit malaysia: (Worldwide) > 49.500
|   |   |   |   taman negara resort: (Malaysia) > 32.500: 9070.500 (count=2)
|   |   |   |   taman negara resort: (Malaysia) ≤ 32.500
|   |   |   |   |   Taman Negara: (Malaysia) > 44.500: 6309.500 (count=2)
|   |   |   |   |   Taman Negara: (Malaysia) ≤ 44.500
```

Fig. 11. Rules Extracted from the Decision Tree Regression.

TABLE III. DECISION TREE REGRESSION MODELLING RESULTS

Sampling Method	**RMSE**	r^2
Linear Sampling (k = 4)	3278.578	0.07
Linear Sampling (k = 9)	3474.772	0.104
Shuffled Sampling (k = 4)	3424.049	0.096
Shuffled Sampling (k = 9)	3570.849	0.058

Based on the lowest RSME, the best SVR is when linear sampling done with k = 4, at 3278.578, though the r^2 is the lowest among the other experiments.

D. Support Vector Regression Results

The Support Vector Regression produces an average result between Multiple Linear Regression and Decision Tree Regression. Table IV displays the best result for linear sampling and shuffled sampling. Based on the lowest RSME, the best SVR is at linear sampling with k = 9, at 2727.532 and the r^2 is the second best among the experiments conducted.

TABLE IV. SUPPORT VECTOR REGRESSION MODELLING RESULTS

Sampling Method	**RMSE**	r^2
Linear Sampling (k = 9)	2727.532	0.286
Linear Sampling (k = 10)	2749.607	0.306
Shuffled Sampling (k = 9)	2741.599	0.234
Shuffled Sampling (k = 10)	2741.599	0.234

E. Best Model Deployment

Selection of best model is measured by the lowest RMSE value and the highest value of the squared correlation between the predicted and the actual values. The experiment proved that the decision tree model displayed a weak performance for this research study as it produced a greater error as compared to other model at any parameters tuning. Between MLR and SVR, the error produced in MLR is almost the lowest, but the squared correlation value is lesser than the SVR model. Additionally, the cross-validation with number of folds, k = 8 for the linear sampling type indicated the best model is MLR which outperforms SVR and Decision Tree. The assumption is made that the number of folds is depending on the number of instances in the dataset and the sampling type is based on the problem model, which in this study the input is the linear problem to find the relationship between these variables.

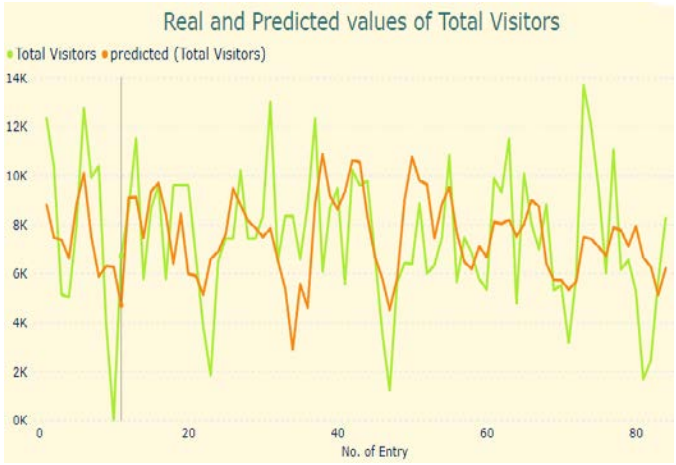


Fig. 12. The Graph of Real (Actual) Value and Predicted Value of Total Visitors.

Fig. 13. The Graph of Prediction until 2030.

Fig. 12 represents the real (actual) value of visitors and the predicted value of visitors by using multiple linear regression. The advantage of using multiple linear regressions is that when given values in decimal point, the results can be easily interpreted by the decision maker. The graph also visualizes clearly the gap between actual and predicted value of the visitors. Subsequently, Fig. 13 shows the predicted total number of visitors until 2030 with some simulated values for each variable in the best MLR model. In future, this research can be extended by providing the estimated values for each relevant variable, and the predicted total number of visitors can then be stipulated to the tourist management.

V. Conclusion

This paper presented the implementation of regression models, namely Multiple Linear Regression (MLR), Support Vector Regression (SVR) and Decision Tree Regression (DTR). A set of variables were constructed based on the selected keywords, the currency exchange and the weather variables for predicting the number of total visitors. The experiments conducted had indicated that these regression algorithms were able to predict the total number of visitors to Taman Negara National Park. The results for the experiments after tuning of parameters demonstrated an improved accuracy of the models since it can control the complexity, which indirectly prevented from overfitting of the model. In this study, the linear problem (input) discovered was to find the relationship between the factors affecting the demand for the total number of visitors to Taman Negara National Park.

Multiple Linear Regression model with linear sampling type and 8-fold cross validation approach appeared to be the best model. The experiments showed that the best parameters setting was based on the instances of the dataset itself. Consequently, some suggestions for future works to improve the quality of the research study were identified. Firstly, the use of advanced visualization tools to work with real-time data to the dashboard can be applied. Next, more data ought to be collected to produce a better performance of predictive models, such as different keywords and any other related campaigns. Lastly, the use of hybrid machine learning and optimization algorithms can be considered to optimize the parameter tuning for better accuracy. The developed model is useful to the tourism management, for predicting the number of visitors to Taman Negara National Park, Malaysia. The tourism management as the user can improve their operations by making a strategic decision making based on the predicted outcomes. If the tourism destination can operate more smoothly, the visitors can reap the benefits from the meaningful experience they received when visiting the national park. Other study can also be performed such as the length of stay and recommended activities based on the tourists' profiles.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for all the support given.

References

[1] G. Ghani, N. Mohamad, and M. I. Ariffin, "Malaysia's Tourism Demand: A Gravity Model Approach (Permintaan Pelancongan Di Malaysia: Pendekatan Model Graviti)," vol. 6, pp. 39-50, 03/01 2018.

[2] A. Shuib, "Tourism in Taman Negara Malaysia Its Contribution as Perceived by Residents of Ulu Tembeling," Akademika, vol. 47, 1995.

[3] "National Park (Taman Negara) of Peninsular Malaysia." https://whc.unesco.org/en/tentativelists/5927/ (accessed 13 July 2021, 2021).

[4] Z. Samdin, Y. A. Aziz, and M. R. Yacob, "Ecotourism in Taman Negara National Park: issues and challenges," 2013.

[5] H. Reduan, "New Straits Times: Protecting Taman Negara," 22 June 2017 2017. [Online]. Available: https://www.nst.com.my/opinion/columnists/2017/06/251183/protecting-taman-negara.

[6] A. Aziz, S. Nur, M. Jamaludin, N. Idris, and M. Mariapan, "The Attractiveness Of Taman Negara National Park, Malaysia As Perceived By Local Visitors," vol. 33, pp. 1-13, 12/15 2018.

[7] "Corporate Finance Institute: Regression Analysis." https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/ (accessed 26 November 2021.

[8] M. Hanafiah and M. Harun, "Tourism Demand in Malaysia: A cross-sectional pool time-series analysis," International Journal of Trade, Economics and Finance, vol. 1, pp. 80-83, 01/01 2010, doi: 10.7763/IJTEF.2010.V1.15.

[9] S. S. A. Kosnan, N. Ismail, and S. Kaliappan, "Determinants of international tourism in malaysia: Evidence from gravity model," Jurnal Ekonomi Malaysia, vol. 47, pp. 131-138, 01/01 2013.

[10] W. Höpken, T. Eberle, M. Fuchs, and M. Lexhagen, "Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden," Information Technology and Tourism, Article vol. 21, no. 1, pp. 45-62, 2019, doi: 10.1007/s40558-018-0129-4.

[11] K. Li, W. Lu, C. Liang, and B. Wang, "Intelligence in tourism management: A hybrid FOA-BP method on daily tourism demand forecasting with web search data," Mathematics, Article vol. 7, no. 6, 2019, Art no. 531, doi: 10.3390/MATH7060531.

[12] U. Claude, "Predicting tourism demands by google trends: A hidden markov models based study," Journal of System and Management Sciences, Article vol. 10, no. 1, pp. 106-120, 2020, doi: 10.33168/JSMS.2020.0108.

[13] A. H. B. Pengiran Bagul, Developing Climate Change Dimensions In Malaysia Through Tourists' Perception. 2013.

[14] S. Cankurt and A. Subasi, "Tourism demand modelling and forecasting using data mining techniques in multivariate time series: A case study in Turkey," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 24, 01/01 2015, doi: 10.3906/elk-1311-134.

[15] C. Zhong-jian, L. Sheng, and Z. Xiao-bin, "Tourism demand forecasting by support vector regression and genetic algorithm," in 2009 2nd IEEE International Conference on Computer Science and Information Technology, 8-11 Aug. 2009 2009, pp. 144-146, doi: 10.1109/ICCSIT.2009.5234447.

[16] W.-C. Hong, Y. Dong, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic genetic algorithms for tourism demand forecasting," Applied Soft Computing, vol. 11, no. 2, pp. 1881-1890, 2011/03/01/ 2011, doi: https://doi.org/10.1016/j.asoc.2010.06.003.

[17] Z.-y. Mei, H. Qiu, C. Feng, and Y. Cheng, "Research on a forecasting model of tourism traffic volume in theme parks in China,"

Transportation Safety and Environment, vol. 1, no. 2, pp. 135-144, 2019, doi: 10.1093/tse/tdz011.

[18] A. Hayes. "Multiple Linear Regression (MLR) definition." Investopedia. https://www.investopedia.com/terms/m/mlr.asp (accessed 26 November 2021.

[19] E. Sreehari and S. Srivastava, "Prediction of Climate Variable using Multiple Linear Regression," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 14-15 Dec. 2018 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777452.

[20] M.S. Hilmi, S. Mutalib, S.R. Sharif and S.N.K. Kamarudin, "Forecasting electricity demand from daily log sheet with correlated variables," ESTEEM Academic Journal 16, 31-41.

[21] P-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to Data Mining, 2nd Edition," 2019.

[22] M.Z. Shahrel, S. Mutalib, S. Abdul-Rahman, "PriceCop-Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform," International Journal of Information Engineering & Electronic Business, vol. 13 (1), 2021, pp. 1-14, doi: 10.5815/ijieeb.2021.01.01.

[23] F.-C. Yuan, "Intelligent forecasting of inbound tourist arrivals by social networking analysis," Physica A: Statistical Mechanics and its Applications, vol. 558, p. 124944, 2020/11/15/ 2020, doi: https://doi.org/10.1016/j.physa.2020.124944.

[24] N.A.M. Salim, Y.B. Wah, C. Reeves et al. Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Sci Rep 11, 939 2021, https://doi.org/10.1038/s41598-020-79193-2.

[25] R. Sovia, M. Yanto, Yuhandri, "Prediction Tourist Visits With Multiple Linear Regressions in Artificial Neural Networks," Turkish Journal of Computer and Mathematics Education, vol. 12 No. 3, pp. 1492-1501, 2021.

[26] N. Agustina, C.D. Puspita, D. Arifatin, R. Yordani, "Application of Logistic Regression to Determine The Quality of Foreign Tourists to Indonesia," In Journal of Physics: Conference Series 2021 Mar 1 (Vol. 1863, No. 1, p. 012029). IOP Publishing.

[27] M. Jackman and S. Naitram, "Segmenting tourists by length of stay using regression tree models", Journal of Hospitality and Tourism Insights, Vol. ahead-of-print No. ahead-of-print, 2021 https://doi.org/10.1108/JHTI-03-2021-0084.

[28] Nurkhodzha Akbulaev, Gulnar Mirzayeva, "Analysis of a paired regression model of the impact of income from international tourism on the foreign trade balance," African Journal of Hospitality, Tourism and Leisure, Volume 9(1), pp 1-13, 2020.

[29] R. Zhang, "Exploration of Social Benefits for Tourism Performing Arts Industrialization in Culture-tourism Integration based on Deep Learning and Artificial Intelligence Technology," Frontiers in Psychology. 2021;12:37.

[30] S. Moghtasedi, C. I. Muntean, F. M. Nardini, R. Grossi and A. Marino, "High-Quality Prediction of Tourist Movements using Temporal Trajectories in Graphs," In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 348-352, 2020.

[31] F. Qian, C. Han and M. Haiyan, "Intelligent model system for tourism flow prediction: a study of Xi'an Museum," In Proceedings of the 2016 International Conference on Intelligent Information Processing, pp. 1-7, 2016.

Non-functional Requirements (NFR) Identification Method using FR Characters based on ISO/IEC 25023

Nurbojatmiko*, Eko K. Budiardjo, Wahyu C. Wibowo
Faculty of Computer Science, Universitas Indonesia
Kampus Baru UI Depok, Pondok Cina, Beji
Depok, West Java, Indonesia16424

***Abstract*—The researches show that software quality depends on Functional Requirements (FR) and Non-Functional Requirements (NFR). The developers identify NFR attributes by interviewing stakeholders. The difficulty in identifying NFR attributes makes quality requirements often ignored. The basic concept of software quality measurements is the quality measuring of the software product. During product-based quality measurement, the potential of software development process repetition will occur. Factors measuring software product quality are not suitable for NFR identification. These differences result in the software development process repeating itself and additional costs. This research proposes easy NFR attributes identification using FR characters. The NFR and FR tightly relations are obtained by extending the NFR measurement at ISO/IEC25023 to programming coding level, then generalizing to get the FR character. The generalization uses the Grounded Theory method. The result is the NFR attributes identification method using FR character based on ISO/IEC 25023. The analyst or programmer can identify the NFR attributes from the FR using the FR character in the requirements stage. This research produces an NFR Identification Method that has been validated by experimenting with several programmers and experts. Tests on programmers identify NFR using the FR character method. The test is to see the level of similarity of the resulting NFR. The result of the test shows level similarity upper 75%.**

***Keywords*—Non functional requirements; FR character; ISO/IEC 25023; NFR identification**

I. Introduction

Software development success requires quality measurement results of the software product. The Software quality measurements base ISO/IEC 25023 is a complete software product quality measurements [1]. The quality classification of the software in ISO/IEC 9126 [2] updated ISO/IEC 25023 [1] is the NFR attributes classification. NFR attributes identification awareness determines the quality of software products. NFR Identification affects the resulting software product [3], [4]. Failure of NFR attributes identification may be repeatable in the software development process [5], [6]. The results of the NFR attributes identification determine the success of the software product [7], [8]. This paper shows that awareness for the NFR attributes identification is essential. The problem of improper NFR identification causes dissatisfaction with product quality, resulting in repetition of the software development process and increasing costs.

The broad meaning of quality towards software products has prompted several studies to classify NFR. Several researchers have solved the problem of NFR identification using the NFR Classification [5] [9]. Problems arise again in determining an unambiguous NFR classification [10]. Developers based on the Agile method need NFR identification quickly and precisely. The problem software developments are NFR attributes identification suitable with software quality measurements. This research uses ISO/IEC 25023 as a basis so that the process of quality identification and measurement has the same reference. How to NFR attributes identification use ISO/IEC 25023?.

This research aims to develop ISO/IEC 25023 for NFR attributes identification. This Research in ISO/IEC 25023 extended the measurement function to the programming coding level. The generalization programming code to get FR character use grounded theory. The research is NFR attributes and FR character tightly relations. The result of this research is the NFR attributes identification method using FR character. This research on the NFR identification process uses FR characters based on ISO/IEC 25023. NFR attributes identification testing uses the FR character on several programmers. The result of the test is to determine the level of similarity of the NFR attributes obtained above 75%. This research method is open coding stage, axial data stage, selective coding stage, forming theory stage, and memoing. Writing systematic of this paper is abstract, introduction, related works, research method (the generalization use grounded theory), discuss, conclusion, acknowledgments, and reference. The grounded theory method consists of stages open coding, axial coding, selective coding, and forming theory. Forming of theory determines NFR attributes identification formulation method and NFR attributes method testing.

II. Related Work

Yusop identifies NFR attributes used resulting qualitative research with an interview from 5 developers. So, the qualitative research result is NFR attribute classification [11]. Sharma, problem-solving the NFR attributes identification uses automatization detection. The Algoritma automation

*Corresponding Author

detection uses NFR attributes categorization and classification based on semantic patterns [6]. Singh use ISO/IEC 9126 standard for NFR attributes categorization dan classification [12]. Li uses the quality specification of domain and subdomain for NFR attributes classification [13]. Chung classifies NFR attributes from several studies literature such as RADC (Rome Air Development Center), Sommerville, Mac Calls, Matsumoto, Grady, and others. Some studies classify NFR based on software products [5]. Kaur develops integration of the NFR attributes and formal reference model for the NFR classification [14]. NFR attributes classification uses quantitative NFR with questioners for minimizing ambiguity [15].

Chung presents a goal-object pattern framework. That framework uses a model-driven by way of UML metamodel extension. The framework is capturing and reusing FR and NFR knowledge of the small-scale application [16]. Singh identifies NFR attributes using NFR attributes classification with different thematic roles based on ISO 9126 quality factors [12]. Kassab makes metamodel by tracing NFR attributes and FR relation based on strong interdependencies [17]. Farid identifies NFR attributes, its using risk-driven algorithms to the prioritization scheme [18]. Liu develops automatization to detect conflict of the NFR attributes evaluation. NFR classification approach uses ontology realizing with metamodel based on cause-effect and inferences knowledge [19].

The determination of the NFR classification so as not to be confused needs to be standardized. ISO/IEC issued standards for the quality classification and measurement of product quality in ISO/IEC9126 [2]. ISO/IEC9126 updated by ISO/IEC 25023 [1]. The NFR classification according to ISO / IEC 25023 is functionality suitability, performance efficiency, compatibility, usability, reliability, security, maintenance, and portability [1].

Several researchers made improvements to the method stages of the Agile Method to be able to identify NFR. Lawrence Chung uses approaches of modeling and techniques to explain software requirements.[20]. Farid, Agile processes use a risk priority approach with the Non-functional Requirements Planning (NORPLAN) method [18]. Domah uses the NERV methodology to obtain NFR artifacts on the User Story Card, the NAI (NERV Agility Index) score, and so on [21].

III. Research Method

This research generalization process of NFR attributes classification uses the Delon-style Grounded Theory method (Fig. 1) [22]. The first stage is open coding. The open coding stage is collecting and identifying data. The data is from the results of the SLR [23], updating papers, and ISO/IEC 25023 files. The results in this stage are NFR attributes measurement identification and classification (Table I) based on ISO / IEC 25023 standards. The Second stage is axial data. This stage is marking or tagging the measurement functions with programming codes. The result axial coding stage is the classification of NFR attribute measurements and the extension of the measurement function to code programming characteristics (Table II). The third stage is selective coding. The stage is the selecting and comparing of data programming code characteristics, it is the generalization process. The process of generalization results NFR attributes and FR characters relationship (Table III). The fourth is forming theory. This stage is the formation theory to NFR attributes and FR characters relation. The forming theory stage is formulation to NFR identification method (Fig. 2) and testing result (Fig. 3). The Fifth stage is memoing. The memoing stage record of this research to a paper journal.

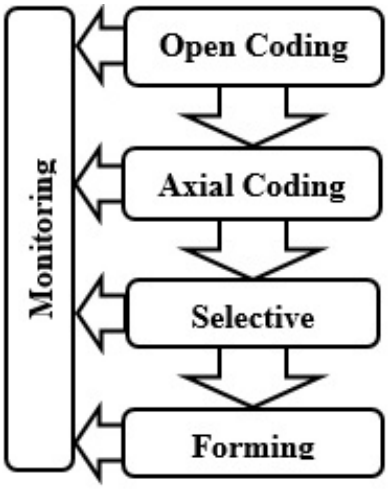


Fig. 1. Delon-style Grounded Theory [22].

IV. Results

A. Open Coding

Referring to SLA takes the condition of the established attributes NFR. That refers standard of software quality. This research uses ISO/IEC 25023 as an international standard product quality software. The measuring factors for software product quality in ISO/IEC 25023 are NFR attributes [24].

ISO/IEC 25023 is a software product quality standard. The elements of quality measures determine measurement functions. Software quality measurement in this way can determine the quantification, characteristics, and sub-characteristics of quality. System and software product quality programs explain that quality measures follow quality characteristics in evaluating internally or externally.

The result of quality measurement identification has eight NFR attributes. The eight NFR attributes in ISO/IEC 25023 are functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, portability. Each attribute is classified and specified into several measurement functions. The ISO/IEC 25023 file on page 10 is an example of measuring performance on Time behavior with specifications for measuring mean response time, response time adequacy, mean turnaround time, and turnaround time. The measurement function has measures the quality of the software product and characteristics of quality. The result of quality measurement identification is the example in Table I.

TABLE I. NFR ATTRIBUTE MEASUREMENT FUNCTION BASED ON ISO/IEC25023 (EXAMPLE: FUNCTIONAL SUITABILITY AND PART OF PERFORMANCE EFFICIENCY)

Specifications	Sub-specifications	Measurement Function	
		Measurements	**Measurement Characteristics**
NFR1. Functional Suitability			
NFR1.1. F. Completeness	NFR1.1.1. F. Completeness	What proportion of the specified functions has been implemented?	1) All FR has boolean value (yes/no=1/0) 2) The Unit of Number FR 3) Number of FR codes that have been made 4) Number of FR encodings that work 5) The number of FR accordance with the function of the actor 6) The number of FR has conformity to the system (average for point (5))
NFR1.2. F. Correctness	NFR1.2.1. F. Correctness	What proportion of functions provides the correct results?	
NFR1.3. F. Appropriateness (FA)	NFR1.3.1. FA of usage objective	What proportion of the functions required by the user provides appropriate outcomes to achieve a specific usage objective?	
	NFR1.3.2. FA of system	What proportion of the functions required by the user to achieve their objectives provides appropriate outcomes?	
NFR2. Performance Efficiency			
NFR2.1. Time Behaviour	NFR2.1.1. Mean response time	How long is the mean time taken by the system to respond to a user task or system task?	1) Unit of time is ms (millisecond) 2) Response time is the time of the page to page 3) Average of response time
	NFR2.1.2. Response Time Adequate	How well does the system response time meet the specified target?	
	NFR2.1.3. Mean turnaround time	What is the meantime taken for the completion of a job or an asynchronous process?	1)Unit of Time is ms (millisecond) 2) Turnaround time is the time of process to one task or process to asynchronous until the finish 3) Average of turnaround time
	NFR2.1.4. Turnaround Time	How well does the turnaround time meet the specified targets?	
	NFR2.1.5. Mean Throughput Time	What is the mean number of jobs completed per unit time?	1) This unit is the number of data transfers per time (number of data every ms) 2) Throughput time is the time needed to start transferring some data to completion to the destination 3) Average throughput time
NFR2.2. Resource Utilization	NFR2.2.1. Mean processor utilization	How much processor time is used to execute a given set of tasks compared to the operation time?	1) The unit of time, ms (millisecond) 2) Processor time is time used to execute a given set of tasks 3) Operation time to perform the tasks 4) Time required for 1 task to perform arithmetic OR logic functions 5) Mean processor time
	NFR2.2.2. Mean memory utilization	How much memory is used to execute a given set of tasks compared to the available memory?	1) Unit size of memory (byte) 2) Size of memory used to perform series of task 3) Average for point (2) 4) Size of memory available
	NFR2.2.3. Mean I/O device utilization	How much of the I/O device's busy time is used to perform a given set of tasks compared to the I/O operation time?	1) Unit of time (ms) 2) The time of I/O device used to perform a series of task 3) Time of I/O operation
	NFR2.2.4. Bandwidth utilization	What proportion of the available bandwidth is utilized to perform a given set of tasks?	1) Unit of bandwidth (bps/bits per second) 2) Size of data (byte) carried to perform a series of tasks per time (second)
NFR2.3. Capacity	NFR2.3.1. Transaction processing capacity	How many transactions can be processed per unit time?	1) Unit number of transaction processes for per time (second) 2) Average of point (1) 3) Transactions related to record in the database (Create, Read, Update, Delete)
	NFR2.3.2. User access capacity	How many users can access the system simultaneously at a certain time?	1) Maximum number of users at the same time 2) Unit number of user
	NFR2.3.3. User access increase adequacy	How many users can be added successfully per unit time?	1) Unit number of users per time 2) Acceleration of user growth

TABLE II. EXTENDING FROM MEASUREMENT FUNCTION TO PROGRAMMING CODE (EXAMPLE: PERFORMANCE EFFICIENCY OF THE TIME BEHAVIOR SPECIFICATION)

ID_NFR	Measurement Function		
	Measurements	**Measurement Characteristics**	**Programming Code**
NFR2. Performance Efficiency			
NFR2.1.1.	How long the meantime took by the system to respond to a user task or system task is?	1) Unit of time is ms (millisecond) 2) Response time is the time of the page to page 3) Average of response time	Link, submit, download, upload, back, next
NFR2.1.2.	How well does the system response time meet the specified target?		
NFR2.1.3.	What is the meantime taken for the completion of a job or an asynchronous process?	1)Unit of Time is ms (millisecond) 2) Turnaround time is a time of process to one task or process to asynchronous until the finish 3) Average of turnaround time	Proses CRUD in the database (Create, Read, Update, Delete)
NFR2.1.4.	How well does the turnaround time meet the specified targets?		
NFR2.1.5.	What is the mean number of jobs completed per unit time?	1) This unit is the number of data transfers per time (number of data every ms) 2) Throughput time is the time needed to start transferring some data to completion to the destination 3) Average throughput time	Sum of data per unit time for the transfer process to or from the database (CRUD)
NFR2.2.1.	How much processor time is used to execute a given set of tasks compared to the operation time?	1) The unit of time, ms (millisecond) 2) Processor time is time used to execute a given set of tasks 3) Operation time to perform the tasks 4) Time required for 1 task to perform arithmetic OR logic functions 5) Mean processor time	1) Process of logic and arithmetic functions 2) Process of mathematics operation (+, -, /, *) 3) Process of logic (<, >, = =, <=, >=) 4) Process of condition functions (IF, FOR, While, Switch - Case, Do – While)
NFR2.2.2.	How much memory is used to execute a given set of tasks compared to the available memory?	1) Unit size of memory (byte) 2) Size of memory used to perform series of task 3) Average for point (2) 4) Size of memory available	Functions of variable OR declaration
NFR2.2.3.	How much of the I/O device's busy time is used to perform a given set of tasks compared to the I/O operation time?	1) Unit of time (ms) 2) The time of I/O device used to perform a series of task 3) Time of I/O operation	Time used to I/O operation (example, process to print, download(curl_setopt), upload(fungsi; input type="file") and scan functional view to report, tranfering data network)
NFR2.2.4.	What proportion of the available bandwidth is utilized to perform a given set of tasks?	1) Unit of bandwidth (bps/bits per second) 2) Size of data (byte) carried to perform a series of tasks per time (second)	Transmission (variable array)
NFR2.3.1.	How many transactions can be processed per unit time?	1) Unit number of transaction processes for per time (second) 2) Average of point (1) 3) Transactions related to record in the database (Create, Read, Update, Delete)	The Number of CRUD function process in the database
NFR2.3.2.	How many users can access the system simultaneously at a certain time?	1) Maximum number of users at the same time 2) Unit number of user	The active user name function
NFR2.3.3.	How many users can be added successfully per unit time?	1) Unit number of users per time 2) Acceleration of user growth	

B. Axial Coding

This stage performs the preparation of NFR measurements and decreases from measurement function to programming coding. The composition of the software quality measurement is the arrangement of the NFR attributes in Table I. This research extends from measurement functions to programming code. This measurement results in a more precise measurement. That NFR measurement determines FR and NFR relations. The fragment of an ISO / IEC 25023 file of Time Behavior Measures in the Measurement Function column contains the measurement algorithm for that quality. For example, the measurement algorithm for response time adequacy on the performance efficiency attribute says Ai = the time it takes the system to respond. The time needed by the system to respond has the meaning of time when the user clicks on a function until the function appears. Simplifies, it is the span of the time from one page to the next page. The time has a unit of time in ms (millisecond). The time-span from one page to another on the measurement translates in the code programming, namely link, submit, download, upload, back, next in the code programming column with the id number NFR2.1.1 in Table II. The result of this stage extends from the measurement function to the programming code, as shown in Table II.

TABLE III. FUNCTION CHARACTER OF NFR ATTRIBUTE BASED ON ISO/IEC 25023

No	NFR Attributes	Function Characters
1	**Functional Suitability**	The function runs according to the actor
2	**Performance Efficiency**	CRUD (Create, Read, Update, Delete)/ database Process
3	**Compatibility**	Files Transfer
4	**Usability**	Make it easier for users and reduce human error
5	**Reliability Measures**	Failure/error detection in the system
6	**Security**	Security at the time of transfer, data, access rights, and control
7	**Maintainability**	Module-related functions (reusability, log, analysis, modification, testability)
8	**Portability**	Configuring apps; Functions to adapt to other environments/applications/software (installations, products)

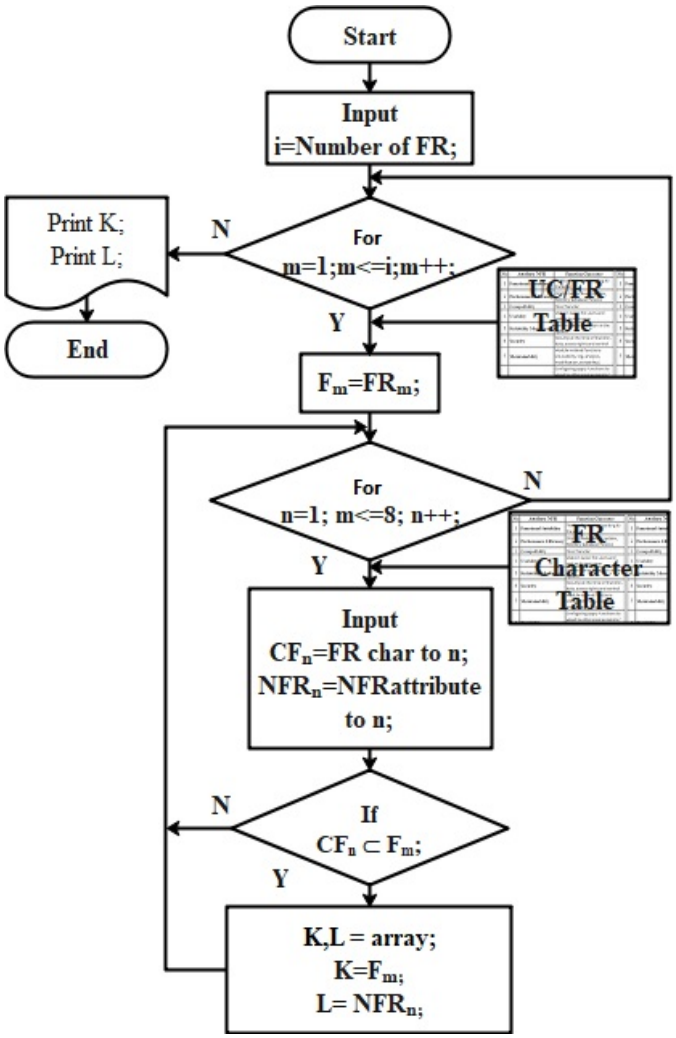


Fig. 2. Pseudo Code NFR Identification use FR Character Method.

C. Selecting Coding

The research takes as an example the NFR attributes of performance efficiency (Table II) with the behavior of time, resource utilization, and capacity specifications. The table contains the ID_NFR, measurements, measurement characteristics, and programming code columns.

Comparison of 2 is the process of CRUD with the process of arithmetic and logic in line 4 on ID_NFR 2.1.4. The CRUD process follows some arithmetic and logical process more than others. The arithmetic and logic processes for measuring processor speed followed by the CRUD process significantly affect processes that are not followed by the CRUD process because there is a process for bringing extensive data from the database. The arithmetic and logic processes that need to be measured are those that the CRUD process follows.

Comparison of 3 is the process of CRUD with rows 5,6,7 and 9, on ID_NFR 2.2.1, ID_NFR 2.2.2, ID_NFR 2.2.3, ID_NFR 2.2.5. These rows have a profound effect on the process the CRUD process follows. That it can ignore another process because the process represents that the CRUD process follows. Performance efficiency measurement for FR concludes that all FRs have a function for the CRUD process. The data comparison results show in Table III, namely that FR's characters have NFR attributes.

D. Forming Theory

The generalization results in Table III show that each FR character has an NFR attribute. NFR is highly dependent on the FR. The determination of the NFR attribute is from the character content in the FR. NFR has a relationship with FR. Research shows that NFR has tightly coupled with FR. FR has more than one character means it has more than one NFR attribute. Each NFR attribute can be on multiple FRs. The relationship between FR and NFR has cardinality, many to many, meaning that FR has more than one NFR, and conversely, NFR has more than one FR. Identification of quality can be known early based on the FR obtained.

The FR character is the result of the generalization from the measurement function, as shown in Table III. NFR attributes can be detected quickly and accurately against FR by an analyst at the requirements stage. Early identification of NFR attributes can monitor the quality of functions during the development process.

1) NFR attributes identification method formulation: The requirements stage is the identification process of FR and NFR. The determination of the NFR attribute at that stage is after the FR determination. Determination of NFR attributes using the FR character (Table III) against the FR. Each FR will derive NFR attributes based on the characters it contains. Further research formulates a method for the identification of NFR attributes using the FR character. The FR from Use Case Specification (UC Spec) or the FR table is the input. Each FR determines the content of FR characters in Table III to obtain NFR attributes. The result is that FR is related to NFR attributes. Fig. 2 is the NFR identification method algorithm using the FR character.

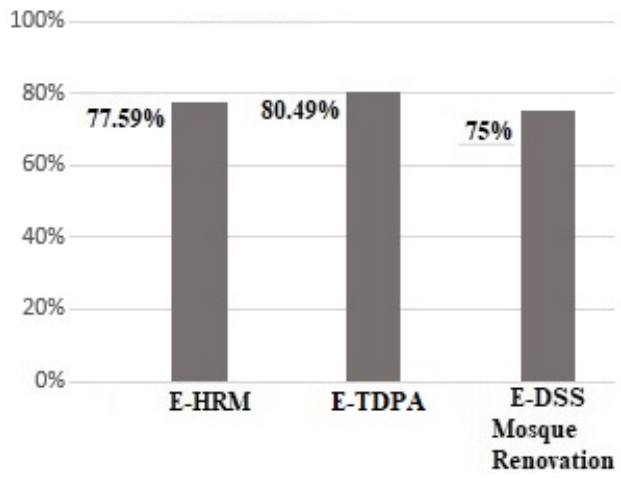


Fig. 3. The Result of the Similarity Level of NFR Attributes Identification.

2) NFR attributes identification method testing: Testing the NFR attribute identification method to determine the level of similarity of the programmers to determine the NFR attributes to their FR. The testing used the samples of 3 web applications based on PHP and MySQL, namely e-HRM, electronic Tax Dispute Power of Attorney (e_TDPA), Electronic DSS Renovation of Indonesian Mosques (e-DSS Mosques Renovation). The study identified FR in all three applications according to its SRS. The research identified the FR of the SRS documents in each application, then arranged them in the FR table. The programmer uses the FR character table guide to fill in the NFR attributes contained in each FR. The research uses the samples of 5 programmers. The test results show the similarity level of NFR attributes above 75% (Fig. 3). These results indicate that the programmers have a high common perception of NFR attributes.

V. Discussion

Pratama determined software quality using ISO/IEC 25010. He assessed software quality of 8 attributes using Black-Box testing, stress testing, and questioner of the 100 respondents. And then leveling of the 8 attributes used AHP [25]. This research for software quality using ISO/IEC 25023 standard of software quality product. That standard measure to suitable between software product quality with requirements. ISO/IEC 25010 is the family of 250n standards. ISO/IEC 25010 is measurements to software quality model and ISO/IEC 25023 is measurements to software product quality. Yusop classified NFR attributes based on qualitative research results [11]. Sharma classified NFR attributes only performance, availability, and security is not using an international standard, Li used quality specifications into domain and subdomain [13], and Singh used ISO/IEC 9126 [6], [26]. Kaur classified NFR attributes with NFR attributes and formal reference relations. The Formal References are domain knowledge, customer requirements, specification, programming platform, and machine [14]. This research NFR attributes categorization and classification use ISO/IEC 25023 that is updating from ISO/IEC 9126. Then, the research develops NFR attributes categorization and classification to get NFR attributes and FR characters integration.

Chung identified the requirements to use the objective-object pattern of FR and NFR relations. The knowledge patterns have format from the experience of several applications samples [16]. Farid used a risk-driven algorithm for NFR attributes identification [18]. Liu detected NFR attribute conflict using ontology realization. This research NFR attributes identification used the FR characters approach. Kassab used an understanding of NFR attributes and FR relationships for the detection of NFR attributes [17]. This research uses NFR attributes and FR relation with FR characters approach. Research identifies NFR attributes based on the character from FR. Each FR has characters. FR character comes from NFR attributes categorization and classification based on ISO/IEC 25023 extend. NFR attribute identification and software quality measurements have the same based on ISO/IEC 25023. The result of software quality measurements suite requirements.

VI. Conclusion

ISO/IEC 25023 is a standard for measuring software product quality. This research succeeded in constructing the FR character from ISO/IEC25023 for the NFR Identification Method. The FR character to identify the NFR attribute of each FR that has the characters. FR character and measurement of quality software products have the same basis, namely ISO/IEC 25023. Identification of NFR attributes using FR character will produce NFR attributes following the desired quality software product.

FR character is a bridge that connects FR with NFR quality or attributes. NFR and FR are relations that both have tightly coupled. Stakeholders are very helpful in determining NFR attributes without having to interpret the type of software quality. Programmers can control the quality of built-in functions while coding these functions. Product software quality is in line with the same NFR attributes based on ISO/IEC 25023. Quality control of software before it becomes a product avoids repeating the software development process and adding costs. Identifying the right NFR attributes determines a quality software product. The future research monitors software quality using FR characters in the Scrum software development method.

Acknowledgment

The funder for this research was from the PUTI Grant the Universitas Indonesia with the number NKB-559/UN2.RST/HKP.05.00/2020.

References

[1] ISO/IEC25023, “Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of system and software product quality,” Int. Stand. ISO / IEC 25023, vol. 2016, no. First edition, pp. 1–54, 2016.

[2] ISO/IEC9126, “International Standard ISO / IEC 9126 Software Engineering — Product quality ISO 9126 - Content,” Int. Stand. ISO/IEC 9126, Softw. Eng. Qual., pp. 1–8, 2001.

[3] H. Kaur and A. Sharma, “A Measure for Modelling Non-Functional Requirements Using Extended Use Case,” Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, pp. 1101–1105, 2016.

[4] A. Alwadi, A. Nahhas, S. Bosse, N. Jamous, and K. Turowski, "Toward a performance requirements model for the early design phase of IT systems," Proc. - 2018 6th Int. Conf. Enterp. Syst. ES 2018, pp. 9–16, 2018.

[5] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, NON-FUNCTIONAL REQUIREMENTS IN SOFTWARE ENGINEERING, 1st ed. New York: SPRINGER SCIENCE+BUSINESS MEDIA, LLC, New York, 2000.

[6] V. S. Sharma, R. R. Ramnani, and S. Sengupta, "A Framework for Identifying and Analyzing Non-functional Requirements from Text Categories and Subject Descriptors," TwinPeaks, Hyderabad, India, June 1, 2014, vol. 14, pp. 1–8, 2014.

[7] M. Iqbal, "NFR Modeling Approaches," Proc. - 1st ACIS Int. Symp. Softw. Netw. Eng. SSNE 2011, pp. 109–114, 2011.

[8] M. Ahmad, N. Belloir, and J. Bruel, "Modeling and verification of Functional and Non-Functional Requirements of ambient Self-Adaptive Systems," J. Syst. Softw., vol. 107, pp. 50–70, 2015.

[9] R. Gomes and N. Bencomo, "On Modeling and Satisfaction of Non-Functional Requirements using Cloud Computing," IEEE, pp. 1–6, 2013.

[10] I. M. S. Raharja and D. O. Siahaan, "Classification of Non-Functional Requirements Using Fuzzy Similarity KNN Based on ISO / IEC 25010," Int. Conf. Inf. Commun. Technol. Syst., vol. 12, pp. 264–269, 2019.

[11] N. Yusop, D. Zowghi, and D. Lowe, "The impacts of non-functional requirements in web system projects," Proc. Eur. Mediterr. Conf. Inf. Syst. EMCIS 2006, 2006.

[12] P. Singh, D. Singh, and A. Sharma, "Rule-Based System for Automated Classification of Non-Functional Requirements from Requirement Specifications," Intl. Conf. Adv. Comput. Commun. Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India Rule-Based, pp. 620–626, 2016.

[13] F. Li et al., "Non-functional Requirements as Qualities , with a Spice of Ontology," IEEE RE 2014, Karlskrona, Sweden, pp. 293–302, 2014.

[14] H. Kaur and A. Sharma, "Use of Reference Model for Formal Specification of NFR," India Int. Conf. Inf. Process. IICIP 2016 - Proc., pp. 1–3, 2017.

[15] K. I. Gómez Sotelo, C. Baron, P. Esteban, C. Y. A. G. Estrada, and L. de J. Laredo Velázquez, "How to Find Non-Functional Requirements in System Developments," IFAC-PapersOnLine, vol. 51, no. 11, pp. 1573–1578, 2018.

[16] L. Chung and S. Supakkul, "Capturing and Reusing Functional and Non-Functional Requirements Knowledge: A Goal-Object Pattern Approach," Proc. 2006 IEEE Int. Conf. Inf. Reuse Integr. IRI-2006, pp. 539–544, 2006.

[17] M. Kassab, O. Ormandjieva, and M. Daneva, "A Metamodel for Tracing Non-Functional Requirements," World Congr. Comput. Sci. Inf. Eng., pp. 687–694, 2009.

[18] W. M. Farid and F. J. Mitropoulos, "NORPLAN : Non-functional Requirements Planning for Agile Processes," IEEE, 2013.

[19] C. Liu, "CDNFRE : Con fl ict detector in non-functional requirement evolution based on ontologies," Comput. Stand. Interfaces, vol. 47, pp. 62–76, 2016.

[20] L. Chung, M. Noguera, and M. L. Rodríguez, "An Agile Requirements Elicitation Approach based on NFRs and Business Process Models for Micro- Businesses," Profes , TORRE CANNE (BR), Italy. June 20 - 22, 2011, vol. 11, pp. 50–56, 2011.

[21] D. Domah and F. J. Mitropoulos, "The NERV Methodology : A Lightweight Process for Addressing Non-functional Requirements in Agile Software Development," Proc. IEEE SoutheastCon 2015, April 9 12, 2015 Fort Lauderdale, Florida, 2015.

[22] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Ramani, and N. Elmqvist, "Integrating Visual Analytics Support for Grounded Theory Practice in Qualitative Text Analysis," Comput. Graph. Forum, vol. 36, no. 3, pp. 201–212, 2017.

[23] Nurbojatmiko, E. K. Budiardjo, and W. C. Wibowo, "SLR on identification & classification of non-functional requirements attributes, and its representation in functional requirements," ACM Int. Conf. Proceeding Ser., no. December, pp. 151–157, 2018.

[24] O. Gordieiev, V. Kharchenko, and M. Fusani, "Software Quality Standards and Models Evolution: Greenness and Reliability Issues," ICTERI 2015, CCIS 594-Communications Comput. Inf. Sci., vol. 594, pp. 38–55, 2016.

[25] A. A. Pratama and A. B. Mutiara, "Software Quality Analysis for Halodoc Application using ISO 25010:2011," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 8, pp. 383–392, 2021.

[26] P. Singh, D. Singh, and A. Sharma, "Rule-Based System for Automated Classification of Non-Functional Requirements from Requirement Specifications," Intl. Conf. Adv. Comput. Commun. Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, pp. 620–626, 2016.

Blockchain-oriented Inter-organizational Collaboration between Healthcare Providers to Handle the COVID-19 Process

Ilyass El Kassmi, Zahi Jarir
Computer Science Engineering Laboratory
Faculty of Sciences, Cadi Ayyad University
Marrakech, Morocco

***Abstract*—Collaborative business activities have aroused great interest from organizations because of the benefits they offer. However, sharing data, services, and resources and exposing them to external use can prevent organizations involved in collaboration from being engaged. Therefore, the need for advanced mechanisms to ensure trust between the different parties involved is paramount. In this context, blockchain and smart contracts are promising solutions for performing choreography processes. However, the seamless integration of these technologies as non-functional requirements in the design and implementation phases of inter-organizational collaborative activities is a challenging task, as reported in the literature. Consequently, we aim in the proposed approach to extend the modeling and implementation of the choreography lifecycle based on service-oriented processes. This is fulfilled by integrating blockchain transactions and smart contract calls, to allow collaboration and interoperability between different entities while guaranteeing trust and auditability. Moreover, to conduct this extension efficiently we use a BPMN choreography diagram combined with Finite State Automata to ensure a meticulous modeling which targets the processes' internal interactions. Hyperledger Fabric is used as a permissioned blockchain for proof-of-concept implementation. A use case of COVID-19 collaborative processes is used to experiment our approach, which aims to guarantee a fluid collaboration between healthcare providers and epidemiological entities at a national scale in Morocco.**

Keywords*—*Blockchain; inter-organizational collaboration; choreography; permissioned blockchain; business process management; COVID-19

I. Introduction

Collaborative business activities have aroused great interest from organizations because of the benefits they offer. This success is due to the technological advances that the interconnected digital world has known. However, sharing data, services, and resources, and exposing them for external use can prevent organizations involved in the collaboration from being engaged. Therefore, the need for advanced mechanisms to ensure trust between the different parties involved is paramount. In this context, blockchain and smart contracts are promising solutions for performing choreography processes. They aim to create effective and secure infrastructures to ease coordination between interacting organizations, ensuring greater achievement of common goals, and avoiding data redundancy caused by the multitude of exchanged messages in conventional schemes.

However, the seamless integration of these technologies as non-functional requirements (NFRs) [1] in the design and implementation phases of inter-organizational collaborative activities is a challenging task, as reported in the literature. This is the case for the inter-organizational healthcare domain, as an example, which is characterized by its complexity both in terms of the management of sensitive data and the heterogeneity of the activities involved in the collaboration process. Blockchain can be introduced to overcome the problem of lack of secure links connecting healthcare organizations including and not limited to public healthcare centers, private clinical centers, ministry regulation entities, specialized centers (dental, laboratory, imagery) and insurance organisms. Moreover, the privacy property characterizing the blockchain provides protection either for exchanged data through security policies or for healthcare professionals' identities as participants in the network.

Blockchain technology has gained attention over the last few years. It has extended its foundation to cover, in addition to the commonly known implementations of digital currencies, the integration of decentralized applications (DApps) [2] in a broad range of industries. This technology aims to overcome the reliance on a centralized authority to certify information's integrity and ownership, and provides a platform for decentralized business logic ensuring transparency and immutability of exchanged data between the distributed interacting parties. Advances in this technology have encouraged organizations to gain confidence in sharing information with other entities without requiring an intermediate trust third party [3]. In order to support the control and exchange of digital assets in the distributed ledger, for common blockchain infrastructures, smart contracts [4] are introduced for this purpose. These contracts are defined as executable programs that allow manipulation of data and functions within the blockchain. They aim to facilitate, verify, or negotiate a contract agreement under a set of conditions to which users agree. When these conditions are met, the agreement terms are automatically carried out.

In collaborative business processes [5][6], the integration of smart contracts aims to create an effective and secure infrastructure dedicated to facilitate collaboration between

interacting organizations and achieve common goals. Furthermore, the tamper-proof property and the exhaustivity provided by the logging characteristic of the blockchain allow the interacting organizations to track who, what, and when transactions are performed and shared to the ledger, according to the agreed security policies. This property guarantees auditability and thus, develops trust between participants without the need for an intermediate trust authority.

In previous contributions, we proposed approaches to tackle the modeling of both functional requirements (FRs) and NFRs [7] and their integration as part of composite services in order to produce an optimal composition, while guaranteeing a prior validation of user properties using model checking verification. Currently, we aim to enhance this approach by providing the ability to handle collaborative processes, ensuring that each process can be executed independently unlike one-time execution in the context of orchestrated service composition. The second important contribution of the proposed approach is the integration of blockchain transactions and smart contract execution as a potential implementation to support the collaboration of business processes based on a service-oriented architecture (SOA). The proposed solution is supported by a proof-of-concept applied to the healthcare domain to meet the emerging collaboration requirements raised by the coronavirus disease 2019 (COVID-19) pandemic.

Our current contribution aims to propose an approach that has the advantages of extending the modeling and implementation phases of the choreography lifecycle by integrating blockchain transactions and smart contract executions, to allow collaboration and interoperability between different entities while guaranteeing trust and auditability. As a case study, this approach was applied to COVID-19 collaborative processes integrating healthcare providers and epidemiological entities at a national scale in Morocco. The proposed solution has the following advantages:

- Proposing a modeling process based on two phases: first using the Business Process Modeling Notation (BPMN) choreography diagram to design the interactions between collaborative business processes, then using Finite State Automata (FSA) to design the internal interactions of the corresponding service compositions.
- Integrating the blockchain technology to tackle common challenges of collaborative business processes such as trust and auditing.
- Providing a support for transactions and smart contract executions, adapted to the context of collaborative business process as service compositions.
- Taking advantage of the security properties and policies granted by the integrated blockchain solution in order to enhance inter-organizational collaboration.
- Proposing a proof-of-concept based on a real-world problematic of inter-organizational collaboration among healthcare entities during the COVID-19 pandemic.

As stated, the proposed approach is based on SOA, consequently, it supports a service-based integration that can be adapted to different blockchain platforms. In this contribution, we implemented a proof-of-concept based on the Hyperledger Fabric architecture [8]. This technology improves the security measures by integrating identity management, access control over assets and resources, and offering the ability to create and use special sub-networks restricted to users with granted permissions. Although we highlight some security improvements offered by the selected technology, our research scope does not focus on vulnerabilities and blockchain-related security threats. The focus is preliminarily oriented to the aspects of modeling and integrating the blockchain as a NFR, to tackle the challenges of inter-organizational business process collaboration.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of the blockchain technology concepts. In Section 3 we expose some interesting contributions tackling business process collaboration challenges using blockchain and SOA. In Section 4, we present a novel approach for handling the integration of permissioned blockchain technology to manage decentralized business processes. Section 5 presents a COVID-19 related case study to demonstrate the behavior of this approach. Finally, in Section 6, we summarize the suggested approach and highlight future works and upcoming perspectives.

II. A Brief Blockchain Overview

Blockchain is a decentralized ledger that records the provenance of digital assets, using a set of protocols and cryptography technologies to securely store data on the network [9]. It can be defined as a sequence of blocks that store all exchanged transactions in a peer-to-peer network [10]. Each block contains, in addition to the stored data, the hash of the previous block and the timestamp of its production. Thus, all the peers store a complete and exact copy of the database, ensuring immutability, decentralization, anonymity and auditability, which promotes trust between participants without the need for a central authority. These cited characteristics, among other remarkable features, have opened the opportunity for diverse research fields around the blockchain, and provided the foundations for a number of application domains, beyond cryptocurrency, such as non-fungible tokens (NFTs) [11], Internet-of-Things (IoT) implementations [12-14], decentralized finance platforms (DeFi) [15], blockchain gaming, and DApps [2].

According to the literature and recent advances in blockchain-oriented technologies, different architectures have been proposed that vary according to the implementation purpose. The most popular architecture is the public blockchain, also known as permissionless blockchain [16], it is a publicly accessible network that does not require special roles for connecting or sending transactions. Thus, any person can access the ledger and interact with it (i.e., read and write data). A public blockchain ensures full transparency of the ledger while maintaining user anonymity. Additionally, no particular participant has total control over the data, as block validation is achieved through consensus mechanisms implying all mining nodes. In contrast, private blockchains are restrictive blockchain architectures that operate in closed networks [17]. They are mostly adopted in enterprises, organizations, and applications that require handling sensitive data, with the aim of using this

technology for internal usage. Moreover, private blockchain is often considered a more centralized architecture than public blockchain because the maintainability of the network is commonly ensured by a single authority. Hyperledger Fabric [18], Sawtooth [19], and Corda [20] are examples of private blockchains. Finally, the permissioned blockchain constitutes a mixed architecture that combines both public and private blockchains. They support numerous customization options, such as identity management, allowing only certain activities to be performed on the network.

Each blockchain platform uses a specific consensus mechanism to achieve the necessary agreement for storing data on a distributed ledger. A consensus mechanism is a procedure through which all the peers of the blockchain network reach a common agreement about the present state of the distributed ledger [21]. A large variety of consensus protocols have been described in the survey [22], such as Proof-of-Work (PoW) [23], Proof-of-Stake (PoS) [24], Proof-of-Authority (PoA) [25], and practical Byzantine fault tolerance (PBFT) [26]. PoW is the most popular consensus mechanism allowing miners to compete to solve the puzzle in order to validate a block and append it to the ledger. Bitcoin [27] and Ethereum [28] are the most popular implementations of public blockchains that use PoW consensus.

Provided that the robustness of the consensus protocols is guaranteed, smart contracts are deployed on the distributed ledger to manipulate data and functions using an analogy similar to classes/objects in object-oriented programming (OOP). Smart contracts can then be composed of attributes and methods, allowing better code structuration and business-oriented modeling of assets and their relationships. This allows the design and development of DApps in a fine-grained manner [4]. They can be used to automate the execution of an agreement so that all participants can be certain of the outcome, without involving any third party. They are also used to automate a workflow or trigger an action when specific conditions are met.

To support business process management (BPM) using blockchain technology, Mendling et al. proposed an interesting survey describing the main challenges and opportunities for this emerging technology [29]. They concluded by presenting seven future research directions based on the conducted research. Among these research directions, two are linked to our proposed approach: (1) the investigation of methods for analyzing and engineering business processes based on blockchain technology, and (2) the definition of appropriate methods for blockchain evolution and adaptation. Another interesting systematic literature review was presented by Garcia-Garcia et al., which aimed to identify opportunities and gaps in the area of collaborative business processes using blockchain technology [30]. They provided different comparative tables to illustrate the contribution scope of each research study. Among the 34 research papers studied, most of the contributions are oriented to public blockchain architecture using Ethereum, while only 15% use Hyperledger blockchain solutions. On the other hand, only 37% integrated process modeling and 45% used or extended BPMN, against 3% using languages based on state machines. The authors concluded that there is a rapid and growing interest in scientific communities and software industries to explore opportunities provided by blockchain technology to improve the management of collaborative business processes in a decentralized manner. Similarly, Xu et al. conducted a comparative analysis of consensus mechanisms [31], and presented an approach incorporating smart contracts and PBFT consensus to address challenges regarding time, prejudice, and trust of process executions in the context of collaborative business processes and IoT. They concluded by highlighting the advantages of the voting mechanism designed to decrease the delay caused by prominence consensuses such as PoW.

III. Related Work

We previously stated that our current approach consists of modeling collaborative processes, building appropriate service-oriented systems, and integrating the blockchain as a service to guarantee decentralization for desired interactions through transactions or smart contract executions. Therefore, in this section, we present some interesting approaches and implementations that address the modeling and integration of blockchain technology for collaborative business processes.

Modeling blockchain-oriented collaborative processes constitutes an open challenge because of the variety of blockchain platforms offering multiple possibilities and usages according to the desired blockchain architectures and consensus mechanisms. Most of the proposed approaches tackling the process choreography using blockchain technology are implementing permissionless blockchains, particularly Ethereum, which represents the first Bitcoin's alternative for public blockchains. Weber et. al proposed one of the first approaches to address the lack-of-trust problem in collaborative business processes based on the blockchain [X]. They aim to map the business processes, designed using BPMN choreography modeling, onto a peer-to-peer execution infrastructure that stores transactions in the blockchain. Smart contracts can then be used as a direct implementation of the mediator process control logic. Authors presented the built modules such as the translator allowing to parse and convert BPMN choreography diagrams files to Solidity scripts, and the Triggers that connect the participants' internal processes with the blockchain. The authors highlighted some of the limitation of their approach based on the presented permissionless platform, which consists on the lack of privacy once the organizations public keys are known, which constitutes opportunities for competitors to inspect and track the data on the ledger. Another approach presented by Garcia-Bañuelos et al. consists on presenting another aspect surrounding the use of blockchain in business process executions, which is the optimization [Y]. They introduced a two-phase modeling based on BPMN processes translated into Petri Net models, then reduced to be converted into Solidity smart contracts. The purpose consists on optimizing costs by reducing gas consumption for overall collaborative business processes. On the other hand, a model-driven methodology for choreography-based systems adapted to the Ethereum blockchain was proposed by Corradini et al. [34]. They implemented all process phases from modeling to execution using a framework that takes BPMN choreography models as input and provides the appropriate Solidity smart contracts [35] adapted to Ethereum. To provide more specifications and capabilities associated with blockchain technologies while dealing with choreography

processes, the authors in [36] assessed the capabilities and limitations of current choreography modeling approaches, and then proposed new language concepts adapted to BPMN choreography processes with their appropriate operational semantics. Di Ciccio et al. proposed another interesting study [37] that presented a comparison between two model-driven approaches, 'Lorikeet' [38] and 'Caterpillar' [39], to deal with the design and implementation of blockchain-based collaborative process execution. Both approaches are focused on BPMN-style process models and implement the Ethereum blockchain.

On the other hand, Pourheidari et al. proposed a study investigating the applicability of the execution of a real-world untrusted business process on a permissioned blockchain [40]. They highlighted the advantages of using the permissioned blockchain for collaborative business processes, and then proposed an order processing model based on the BPMN standard, considering the particularities of Hyperledger Fabric as a blockchain platform and Hyperledger Composer as a framework. This study focused on implementing blockchain to support an existing business process, taking advantage of the modeling capabilities offered by Hyperledger Fabric in addition to the use of identity management to handle permissions. According to the cited contribution, each process is executed independently and cannot be automated. In contrast, our proposed approach is based on a service-oriented architecture that automates the execution of interacting services through a service composition, by integrating blockchain transactions as services. Authors Autili et al. presented an approach to address the problem of trust in service choreography [41]. They proposed blockchain technologies to support decentralized and peer-to-peer collaboration in a trustworthy manner. A model using a BPMN2 choreography diagram is presented to illustrate the peer-to-peer communication between interacting parties. The proposed solution allows trust-based coordination between participants and the verification of the correctness of exchanged messages by providing the ability to validate transactions and manage permissions. The interactions with the blockchain are however weakly expressed during the modeling phase. In comparison, we use in our approach a second phase of modeling based on automata to define, in a fine-grained manner, each interaction with the blockchain to fulfill decentralization requirements.

The following contributions highlight approaches of tackling the interactions with blockchain technology based on SOAs. An interesting contribution [42] proposed a taxonomy based on an analysis of the Blockchain-as-a-Service (BaaS) market, according to service characteristics, the support of different distributed ledger technology (DLT) protocols and consensus mechanisms, and related pricing models for service provisioning. This study highlights the predominance of large IT service providers in the BaaS market, offering a variety of services to match the requirements of corporate customers, who prefer permissioned distributed ledger solutions, and prioritize quality characteristics such as performance, ease of use, and availability. In the same context, Daming et al. proposed a systematic review of recent research studies covering BaaS models [43]. They aimed to categorize the applied scenarios, trends, evaluated quality of service (QoS) factors, new challenges, and open directions on BaaS models in IoT management. According to the conducted study, the BaaS models are mainly applied to the network layer to manage the IoT environment more than other layers. In addition, security and privacy are two important factors for evaluating existing BaaS models in cloud-edge IoT environments. In accordance with these results, Zheng et al. proposed a platform for BaaS called NutBaaS [44]. This platform is based on a four-layer model and provides blockchain services over cloud computing environments, such as network deployment and system monitoring, smart contract analysis, and testing. It also provides some key built-in features, such as identity-chain profile management and smart contracts' security-related vulnerability detection. The proposed solution is technically complementary and based on Hyperledger Fabric. Comparably, our proposed approach uses specific services to interact with the ledger through simple transactions or smart contract execution. However, prior to this integration, we provide a fine-grained modeling process to describe the overall interactions of these blockchain-oriented services, to meet the existing service composition scheme and accomplish the associated business process. Another contribution presents a unified blockchain as a service platform (uBaaS) [45] to support both the design and deployment of blockchain-based applications. This solution combines a deployment as a service, aiming to avoid lock-in to specific cloud platforms, with a design pattern-as-a-service that aims to apply design patterns for data management and smart contract design to address the common scalability and security issues.

Finally, Table I presents a summary of the common aspects of the cited contributions. The covered compared characteristics are as follows: (1) the modeling used, (2) the type of blockchain architecture, (3) the platform used, (4) the implemented consensus mechanism, (5) the adaptation ability to more complex use cases, (6) the approach evolution to cover other blockchain architectures and platforms, and (7) the security attributes and additional behavioral NFRs covered when applicable.

TABLE I. A COMPARATIVE TABLE DESCRIBING CHARACTERISTICS OF CITED CONTRIBUTIONS

	[32]	[33]	[34]	[40]	[41]	[43]	[45]	Our proposed approach
Modeling	BPMN2 Choreography Diagram	BPMN Process Petri Net	BPMN2 Choreography Diagram	BPMN	BPMN2 Choreography Diagram	N/A	BPMN	- BPMN2 Choreography Diagram - Finite State Automata
Architecture	Public Permissionless	Public Permissionless	Public Permissionless	Permissioned	Permissioned	Permissioned	Public Permissionless	Permissioned
Platform	Ethereum	Ethereum	Ethereum	Hyperledger	Hyperledger	Hyperledger	Ethereum	Hyperledger
Consensus	Proof-of-Work	Proof-of-Work	Proof-of-Work	N/A	N/A	Kafka	Proof-of-Work	RAFT
Adaptation	Yes	Yes	Yes	N/A	Yes	Yes	Yes	Yes
Evolution	Yes	No	N/A	N/A	N/A	Yes	N/A	Yes
Security & Behavioral NFRs	N/A	N/A	- Role Management	- Privacy - Access Control	- Transaction Validation - Permission Management	- Profile Management - Smart Contract vulnerability detection	- Permission Management	- Identity & Role Management - Permission Management - Privacy - Custom Security NFRs

We notice in the comparative table that most of the presented works used either BPMN process diagram or BPMN2 choreography diagram to model the case study workflow. However, the designed modeling lacks expressiveness regarding the interaction between the current system and the blockchain. In contrast, we introduce automata-based modeling to describe in a fine-grained manner each interaction between the existing services fulfilling the business requirements and the desired blockchain transactions. Additionally, most of the presented contributions are oriented into a unique blockchain platform, which constitutes a strength when considering the advanced features characterizing the platform, but also a limitation when dealing with the implementation of heterogeneous blockchain architectures that might be implied in the collaborative inter-organizational business processes. Our proposed approach tackles this challenge by adopting a service-oriented architecture in which blockchain interactions are considered as services executed to fulfill predefined behavioral NFRs. Another concern that most cited papers has pointed to is security. Privacy and identity management are key security attributes that are integrated differently, depending on the adopted blockchain platform. For instance, Hyperledger Fabric integrates a built-in permission management system that aims to manage participant roles in accordance to business requirements. However, in Ethereum-based implementations, the permission module can be developed and integrated separately to control access permissions through smart contracts. In our approach, we have a limitless choice of security attributes to integrate. We can either use the built-in security policies provided by the chosen blockchain platform, such as access control lists and channels, and on the other hand, our service-based approach also allows us to define any desired security process as a behavioral NFR during the modeling phase, which, consequently, calls appropriate concrete services and inject them into the generated service composition.

According to the parameters listed in Table I, our contribution covers the overall phases allowing the production of a reliable business process collaboration with the integration of blockchain interactions. The proposed approach combines the benefits of two powerful technologies: SOA and blockchain, and provides an advanced modeling process that combines them with an exhaustive definition of their interactions.

IV. PROPOSED APPROACH

Inter-organizational collaboration implies interactions between heterogeneous components to achieve external processes controlled by foreign entities. This raises numerous challenges depending on (1) the number of interacting organizations, (2) the complexity of published processes, and (3) the agreed policies to secure access and manage permissions. To overcome these challenges, model-driven methods can be helpful to rigorously describe the overall interactions using modeling languages, propose a fine-grained overview abstracted from any technical implementation, and allow reliable interoperability with the ability to customize code execution according to the runtime environment.

In this study, we propose an approach for designing and implementing collaborative business processes based on SOA. Additionally, we aim to introduce the ability to integrate blockchain transactions as a way to guarantee a higher level of decentralization while ensuring reliable collaboration. The proposed approach is supported by a proof-of-concept proposed to fulfill the emerging need raised by the COVID-19 pandemic, which consists of ensuring a wide collaboration between healthcare organizations. The proposed solution will allow healthcare actors to interact and exchange key medical information in accordance with the Moroccan governmental information systems and policies.

In previous contributions, we focused our research on modeling and generating business-oriented service compositions capable of fulfilling both FRs and NFRs [46]. The proposed solution integrates a modeling module that allows the design of interactions between desired services with the ability

to inject behavioral and quality-oriented NFRs, and then generates the optimal appropriate service composition. In the current approach, we aim to extend the scope of handling orchestrated composite services to allow their interactions with external business processes and, therefore, promote inter-organizational collaboration. Thus, we describe a business process as a complex business workflow that can be achieved by executing a service composition to fulfill certain predefined functional and non-functional requirements. A comprehensive modeling of these complex business processes is performed using the FSA. BPMN choreography diagram modeling is used from a wider perspective to define the interactions between multiple inter-organizational business processes. Moreover, we enhance our approach by introducing the ability to inject specific services, allowing interactions with the blockchain through transactions or smart contract executions. These blockchain interactions are covered from the modeling to the execution phases.

A. Modeling Collaborative Business Processes based on Automata

We previously stated that, in our preceding contributions [7, 46], we proposed approaches tackling the service composition challenges and integrating a modeling module based on FSA, allowing a graphical description of the desired FRs and NFRs. The generated composition is an orchestration of concrete services that fulfill the requirements of abstract services designed in the modeling phase. The context of collaborative business processes is slightly different, as it consists of a choreography of different processes achieved on the need, discontinued over time, and not reliant on an orchestrator. For instance, a patient can perform the investigation operation on day *d*, and two days after he completes the PCR laboratory test on *d+2*. In the service composition context, the process is completely executed on the fly to fulfill all the desired requirements at once. In this contribution, we combine both concepts, allowing an initial modeling of the desired collaborative system using the BPMN choreography diagram. Subsequently, we design each process in a fine-grained manner using FSA modeling. Fig. 1 illustrates the proposed modeling approach with a graphical description of each phase.

The first phase, surrounded by orange color, illustrates a BPMN choreography diagram, which defines the interactions between different organizations' processes to fulfill inter-organizational collaboration. As the choreography diagram lacks modeling expressiveness adapted to a more fine-grained level to address internal interactions, we complement it using the automata-based modeling. The automata formalization used was introduced in our previous contributions [7, 46]. It allows a service-oriented modeling of all business processes, and then offers the ability to integrate behavioral NFRs, when needed, to concisely express complex requirements. We used this modeling formalization to design (1) the desired FRs and (2) both measurable and behavioral NFRs. We take advantage of the automata expressiveness to implement complex security NFRs into the desired service composition. An example of implementing complex security attributes into automata modeling is the integration of the Shamir multi-cloud sharing algorithm in a previous study [47]. In our current approach, we adopt this modeling method to address the challenges of designing blockchain interactions. Thus, a blockchain interaction is integrated as a service injected into an automatically orchestrated service composition. The Abstract Functional Automaton (AFA) is introduced to design the desired FRs of a specific business process, in order to allow its translation to a ready-to-use service composition that fulfills these requirements. These FRs are designed in the form of abstract services that constitute automata states, while transitions define the desired interactions between these abstract services. The formalization is inherited from the inputs/outputs/pre-conditions/results scheme (IOPR), allowing fine-grained modeling capabilities to describe complex case studies. From our perspective, the NFRs integrated into the service composition can either be measurable, such as quality-of-service attributes, or behavioral (e.g., security attributes). In this study, we aim to integrate blockchain transactions as behavioral NFRs. This integration can be introduced into AFAs to grant decentralization and immutability properties to any business process. The formalization used provides the ability to use behavioral scopes to graphically design the desired blockchain interactions. Behavioral scopes are graphical automata scopes that specify the subset of states associated with the integrated behavioral NFR. These are illustrated using dashed lines surrounding the target abstract service states. The arrow direction in the scope defines whether the concrete services fulfilling the behavioral NFRs should be implemented before or after scoped abstract services. In the illustrated example, we use a post-execution scope associated with the abstract service (AS_2), aiming to integrate a blockchain transaction as a service. The concrete representation of this integration is illustrated in the third phase, which defines the interactions between the concrete services and the integrated blockchain service.

B. The Proposed Blockchain based Service-Oriented Architecture

In our contribution, we aim to implement a service-oriented architecture that allows fluid interaction with the blockchain to ensure process collaboration using smart contracts. For the implementation, we opted for a permissioned blockchain because of the specificities of the introduced case study. It allows us to restrict interactions to only authorized participants due to the criticality of exchanged information in the context of healthcare business process collaboration. In this perception, despite the heterogeneity of the internal information systems used by collaborating organizations, we are able to provide on-the-fly invocation of blockchain transactions or smart contract execution to meet the requirements described in the initial modeling.

On a large scale, the diversity of providers' roles and the variety of functionally similar services published make business process collaboration a challenging task. The integration of API gateways allows the unification of service calls by using common URI suffixes for functionally similar services. This facilitates service lookup and consumption, especially when the number of organizations is important.

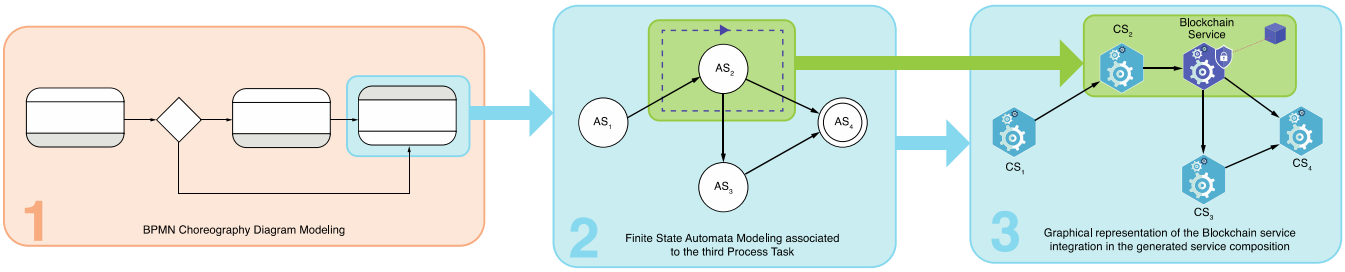


Fig. 1. Graphical Illustration of the Two Phases Modeling Process based on Choreography Diagram and Finite State Automata.

In our opinion, in addition to fulfillment of the primary requirements for a reliable business process collaboration, the privacy, identity management, and consensus mechanisms are crucial factors to consider when choosing a blockchain platform. In the proposed contribution, and to align with the context of the introduced case study, we integrate the Hyperledger blockchain as part of the proposed SOA, aspiring to improve collaboration between healthcare organizations, especially during the pandemic and post-pandemic periods. Hyperledger Fabric is a pillar platform for distributed ledger solutions that aims for democratization and standardization of blockchain for the business landscape, by allowing organizations to build custom blockchain apps to address their business requirements in a secure decentralized way [48]. It ensures a high level of confidentiality, resiliency, flexibility, and scalability. Hyperledger Fabric gained popularity through its plug-and-play approach, using the container technology to run its components to meet the desired business requirements and level of complexity. One of the essential components is the chaincode, which constitutes Hyperledger's native smart contract, defining the rules and operations for specific business processes. Additionally, Hyperledger comes with the ability to choose a consensus or even use a custom consensus implementation to define relations between members of the chain. Another significant improvement provided by Hyperledger Fabric is the support for channel creation [49]. This feature constitutes an additional security layer that allows a group of participants to create separate ledgers of transactions, especially for networks where some participants might be competitors and not want every transaction they make to be known to every participant. These previously cited features constitute the basis of our choice, which makes Hyperledger Fabric the most appropriate for the proposed blockchain proof-of-concept, especially when applied to a healthcare project.

The main components defining business logic in a blockchain solution are assets and transactions. The assets provide the ability to model business network logic in an object-oriented similar way. This feature allows designers to define business logic exhaustively using abstraction, inheritance, dependencies, and relationships. On the other hand, transactions constitute a structured and behavioral business-related description that allows the manipulation of assets through functions. These two components are explored to describe the business logic behind the implementation of collaborative processes.

The illustrated example presented in Fig. 2 describes a workflow showing the benefits of combining SOAs with blockchain technology. The blockchain is integrated as a service to constitute a complementary ledger while maintaining the organization's internal information system functioning and updated continuously. The API gateway intercepts the user requests and executes the appropriate services. The first service endpoint called allows updating the internal database to keep the local information system up-to-date. The second endpoint executes the appropriate blockchain transaction. The ledger is then updated, and events are triggered when applicable. The end-user receives a response describing whether all the steps were successfully performed. On the other hand, Fig. 3 illustrates the interactions of the integrated blockchain service with the modeled service composition. We graphically describe a simplified workflow that represents the interactions with the adopted Hyperledger platform. It shows the usage of the provided software development kit (SDK) to automatically launch the endorsement process, and then to validate the transactions by the orderer once the consensus is reached. Hyperledger Fabric achieves consensus through its ordering service. This service establishes the total order of transactions submitted to the network. In previous versions, Hyperledger Fabric used Kafka to achieve consensus while guaranteeing crash fault tolerance (CFT). As Kafka is commonly used to manage messaging queues, it is known to not provide a Byzantine fault-tolerant (BFT) consensus. Thus, the system cannot be prevented from reaching an agreement in the case of malicious or faulty nodes. In the latest versions, Hyperledger Fabric relies on RAFT [50], which is a distributed consensus algorithm for managing replicated logs, following a "leader and follower" model, where a leader node is elected for each channel. Leaders' decisions are then replicated by followers. This allows different organizations to provide nodes that contribute to the formation of a distributed ordering service.

In the proposed approach, we provide a model based on a choreography diagram with FSA to design the desired business processes, including the ability to perform blockchain transactions. The automata-based modeling of each process is translated into a composite service handled by the orchestrator to fulfill the business process requirements. A blockchain transaction is wrapped into an atomic service that is injected into the generated composition according to the NFR scope defined in the abstract automaton. As presented in our previous study [7], the orchestrator performs a lookup into the NFR registry to select the matching service fulfilling the behavioral NFR defined in the scope. In the current contribution, the behavioral NFR consists of a smart contract execution, which constitutes a blockchain transaction in the Hyperledger context.

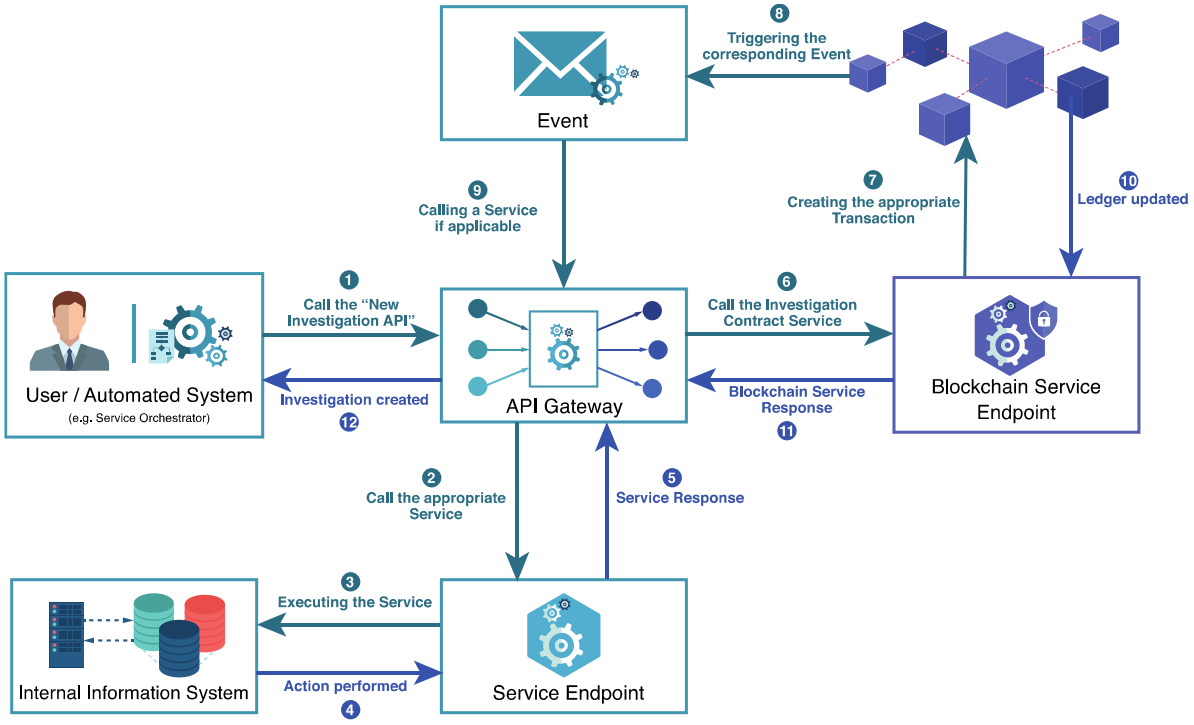


Fig. 2. The Workflow of a Blockchain as a Service Integration into the Service Composition.

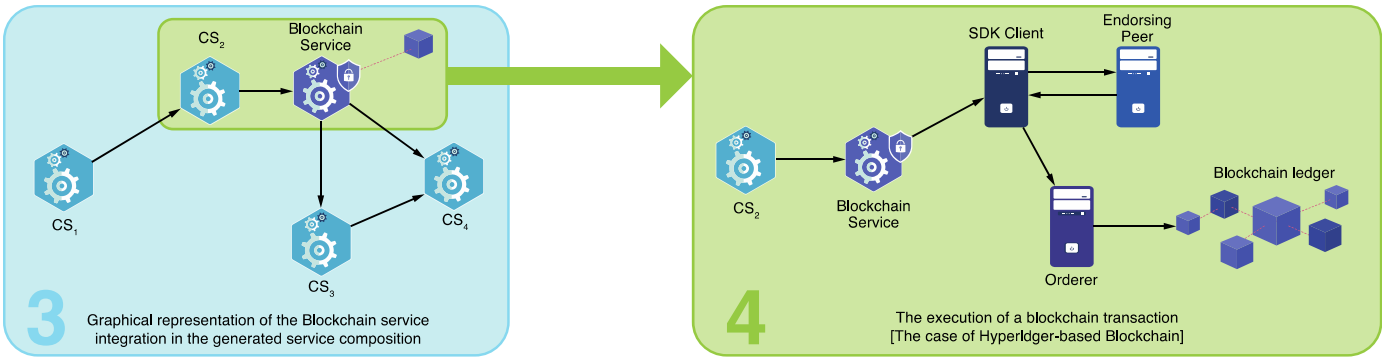


Fig. 3. The Blockchain Service Interactions with the Service Composition and the Ledger according on the Integrated Behavioral NFR Scope.

Finally, the advantage of using a service-oriented architecture is that it provides interoperability with the used blockchain platforms. The provided blockchain-as-a-service concept is not limited to Hyperledger technology, which is used in our proof-of-concept to meet the requirements of the considered case study. Subsequently, the approach is viable as long as the implemented blockchain platforms allow the automation of executed transactions using services, such as the Solidity-based smart contracts implemented by Ethereum blockchain.

V. The Proposed Implementation

In this contribution, we aim to introduce a prototype to overcome the collaboration issues between healthcare entities during the hospitalization process for affected patients, with the ability to extend the current model to cover vaccination processes in an upcoming phase of the project. Our proposed proof-of-concept is based on previously highlighted technologies by combining the composition of services with the blockchain. To fulfill this implementation, we opt for Hyperledger Composer, which is an open-source development toolset based on Hyperledger Fabric. Hyperledger Composer provides advanced abilities for modeling business networks and ensures the easy integration of blockchain applications with existing systems [51]. The main components of blockchain-related business network applications are the chaincode, assets, transactions, participants, events, and access-control rules. Chaincode is a program installed and instantiated onto a blockchain network to enable interactions with the ledger. It runs in a separate process from peers and is responsible for the initialization and management of the ledger state through transactions submitted by applications [52]. It provides also the ability of creating real-world decentralized applications, modeling business processes using assets and relationships, updating and exchanging information in the ledger through transactions, managing participants and their associated roles, and triggering specific business functions or services through events. The next subsections briefly define each of these

components with real-world examples inspired by the COVID-19 case study.

A. Case Study Overview

The world has experienced a remarkable health crisis, which has damaged the economy and negatively impacted healthcare systems. Most healthcare organizations are not logistically and structurally ready to deal with unpredictable large-scale healthcare-related challenges, which require urgent mobilization of resources to prevent additional losses. In the context of the Moroccan government, the healthcare ministry was very vigilant by publishing a new web-based platform allowing epidemiological centers to follow-up the cases of suspected and affected COVID-19 patients. In addition, they deployed a new mobile application allowing to notify users of any suspected physical interaction with affected patients using a geolocation tracking history. For both cited platforms, we noticed a lack of collaboration between healthcare providers, constituting the first and main destination for suspected and affected patients during the pandemic. These providers are required to maintain a continuous coordination with their respective regional epidemiological centers in order to process and investigate daily collected data, and then forward them to the national epidemiological organization within the healthcare ministry.

The collaboration between healthcare organizations in the cited context implies a good understanding of their interactions and their common processes. Briefly, the central epidemiological organization within the healthcare ministry collaborates closely with the regional epidemiological centers in each of the 12 regions of the kingdom. Similarly, each of these regional centers collaborates with all healthcare providers in their respective regions. Thus, in order to support the heterogeneity of methods and optimize interactions guaranteeing an exhaustive exchange of information, we introduce a complete process that allows the modeling and integration of collaborative business processes, with the extended ability to integrate the blockchain on demand. To present a realistic prototype, we aim to cover the overall healthcare processes and dependencies related to COVID-19, such as investigation, laboratory tests, and hospitalization, etc. Additionally, to allow an exhaustive collaboration and enhance inter-organizational interactions, we involve specialized healthcare entities, such as medical imagery centers and medical analysis laboratories. Each entity might contribute to one or multiple phases of the overall COVID-19 process, depending on the policies defined through our modeling process.

Among many interesting contributions integrating blockchain solutions to tackle collaborative challenges in the healthcare domain, especially during the COVID-19 pandemic, J. Xu et al. proposed a privacy-preserving scheme for fine-grained access control adapted to large-scale health data based on blockchain [53]. This platform, named Healthchain, is based on two blockchains to allow the users to upload IoT data and read doctors' diagnoses; on the other hand, it allows doctors to read users' uploaded data and upload appropriate diagnoses. The proposed blockchain-based platform prevents data from being tampered or denied; however, at any time, users can revoke doctors to preserve data privacy when needed. In contrast to the cited solution, and in the first phase of the project, our contribution focuses on business-oriented inter-organizational collaboration, and thus does not integrate patient interactions with the ledger. We are currently exploring the combination of heterogeneous blockchain architectures in order to connect a permissionless ledger for patients' interactions with the permissioned architecture presented in our current approach. Another solution was proposed by Alsamhi et al., aiming to integrate blockchain and multi-robot collaboration to combat the COVID-19 pandemic [54]. The proposed conceptual framework can increase the intelligence, decentralization, and autonomous operations of connected multi-robot collaboration in a blockchain network. The proposed architecture integrates Ethereum as a permissionless blockchain platform. The proposed solution presents the scope of use of each consensus algorithm in addition to shard techniques to maintain connectivity between collaborating robots, avoid collisions, and thus improve real-time. In the same direction, the authors in [55] proposed a blockchain-based system using Ethereum smart contracts to generate statistical information based on reported data related to the number of new cases, deaths, and recovered cases obtained from trusted sources. They presented a comparative analysis of the security and cost incurred by the stakeholders to prove the feasibility of integrating the proposed solution to ensure data integrity, security, transparency, and data traceability among stakeholders. Finally, the authors in [56] conducted a scoping review to identify relevant studies by searching 11 bibliographic databases. They conducted backward and forward reference list checking of the included studies and relevant reviews. According to their study based on 10 use cases of blockchain to tackle COVID-19 challenges, the most prominent use cases were contact tracing and immunity passports. Public blockchain technology was the most commonly used type in the included studies. Out of these 10 studies that identified the platform used, nine studies used Ethereum to run the blockchain, and Solidity was the most prominent programming language used to develop smart contracts. Although Ethereum continues to be the most popular and used blockchain platform in various industries and research fields, it is still susceptible to privacy leakage as public keys are made transparent to members of the network. This constitutes a key factor impacting the choice of blockchain solution and migration to permissioned blockchains, especially for addressing business process collaboration.

In the proposed implementation, we aim to tackle the business process collaboration challenge by implementing comprehensive well-designed composite services, fulfilling all business process requirements, and providing the ability to perform transactions over the blockchain. As illustrated in Fig. 4, we allow healthcare organizations to collaborate through well-defined processes, described using exhaustive modeling. The validated automata models are translated into composite services, integrating blockchain interactions and further behavioral NFRs when necessary. The concrete business services in blue color describe the services executed to perform a specific business need, whereas the concrete blockchain services in purple specify the services executed to perform an action on the blockchain, such as obtaining data from the ledger, or executing transactions to update the ledger's world state.

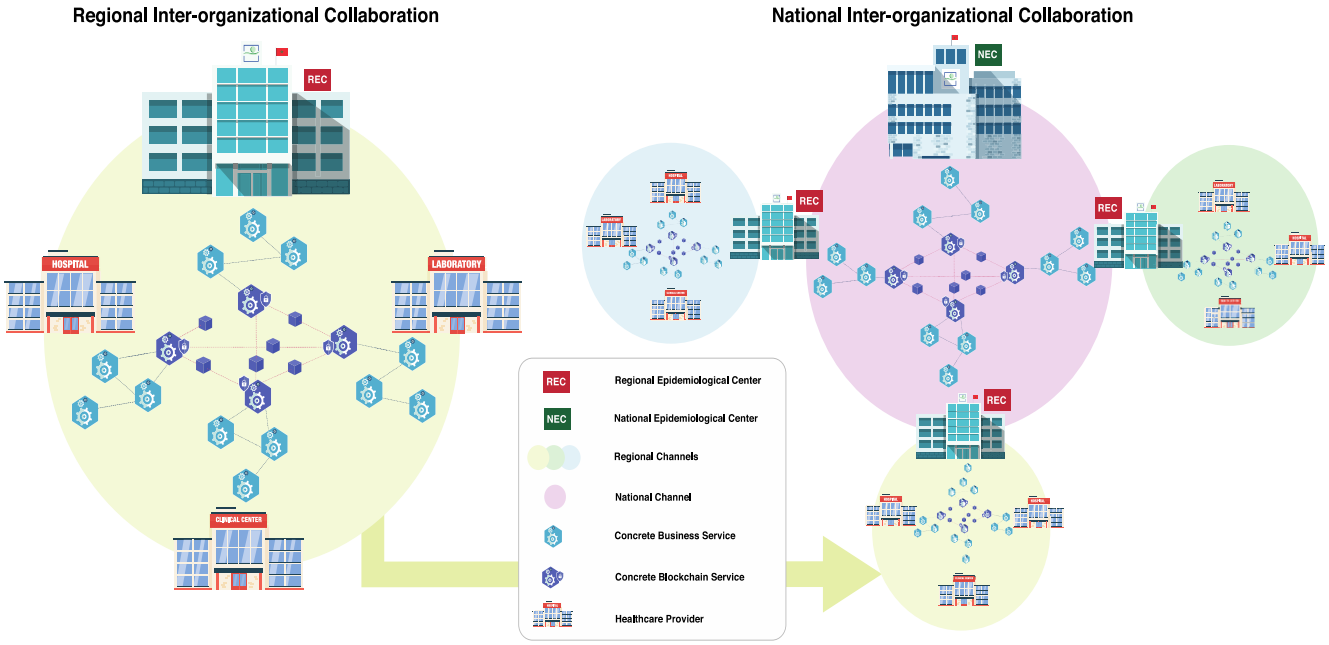


Fig. 4. The Proposed Implementation of Regional and National Inter-Organizational Collaborations based on the Blockchain.

To align the proposed architecture with the policies and hierarchical specificities of the healthcare case study, we integrate the channel feature, allowing the creation of sub-networks to ensure data privacy in accordance with each regional perimeter. This segmentation allows each regional epidemiological center (REC) to play a key role in federating and managing all organizations and participants falling under its regional administrative perimeter. Thus, coordinators of RECs will be in charge of managing the roles and permissions, and supervising the continuous conformance to healthcare ministry policies. Consequently, each REC coordinator interacts with two different ledgers: (1) the main ledger through the national channel, and (2) the regional ledger associated with the regional channel he manages. Examples of regional channels are depicted using blue, yellow, and green circles, whereas the pink circle defines the national channel.

B. COVID-19 Workflow Modeling

The initial phase in the proposed modeling process consists of defining the collaboration workflow through a choreography diagram. In Fig. 5, we present an illustrative model covering the main processes related to the COVID-19 workflow. The participant roles in this example continuously collaborate with RECs. These centers constitute, for each region in the kingdom, the main entities responsible for processing and analyzing collected data, before forwarding the daily COVID-19 key performance indicators (KPIs) to the national epidemiological center (NEC). Additionally, these RECs constitute the decentralized representatives of the healthcare ministry in the region, forwarding all epidemiological-related decisions to healthcare providers, and supervising their execution. The illustrated modeling depicts all phases of a COVID-19 patient, taking into account all cases, depending on the PCR results for both diagnosis and control laboratory tests. The main phases in the choreographic scheme can be shared by different actors; for example, a hospitalization can either be processed by a healthcare public provider or private provider. Similarly, laboratory tests can be performed by either a healthcare provider or a private laboratory center. These processes are examples reproducing real scenarios, summarized and abstracted into one choreography diagram. This diagram is shared by the 12 regions, all replicating the same unified process with the same common roles. The organization's roles participating in the process example are briefly described below, and defined more in deep in the following subsections.

- The regional epidemiological center, which is the focal organizational authority in the region, is responsible for collecting, processing and analyzing epidemiological data from all healthcare providers within its region.
- The healthcare provider, which can be any healthcare entity among of the previously cited organizations. Its main role is to provide healthcare services to patients.
- The medical analysis laboratories, which are the medical laboratories authorized by the RECs to issue PCR tests to patients.

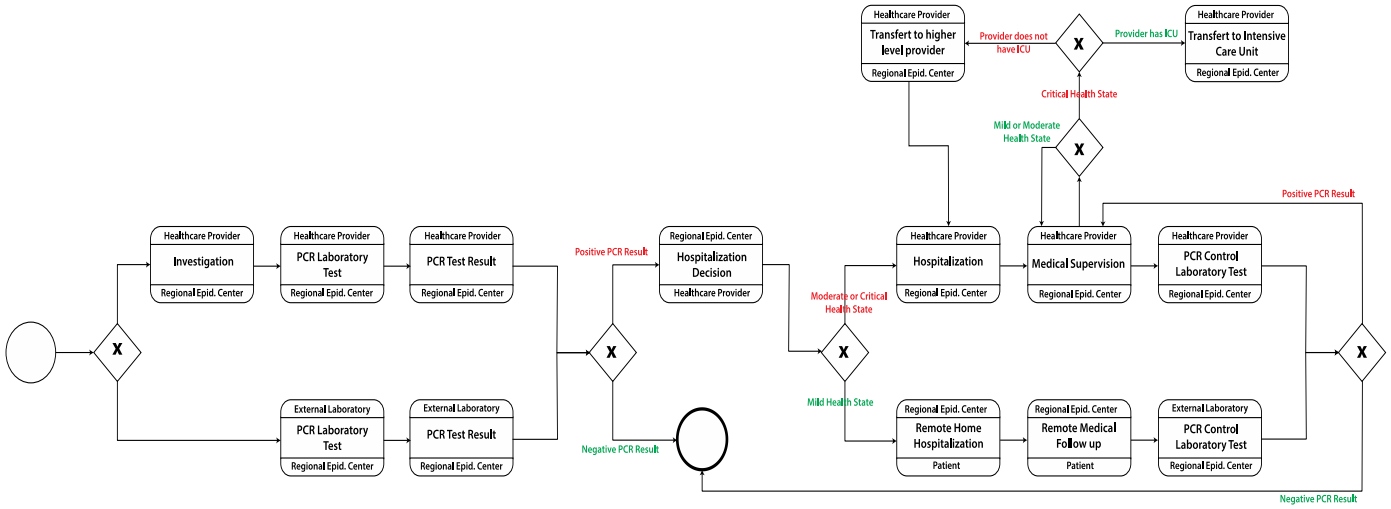


Fig. 5. A Proposed Choreography Diagram for the Adopted COVID-19 Business Processes.

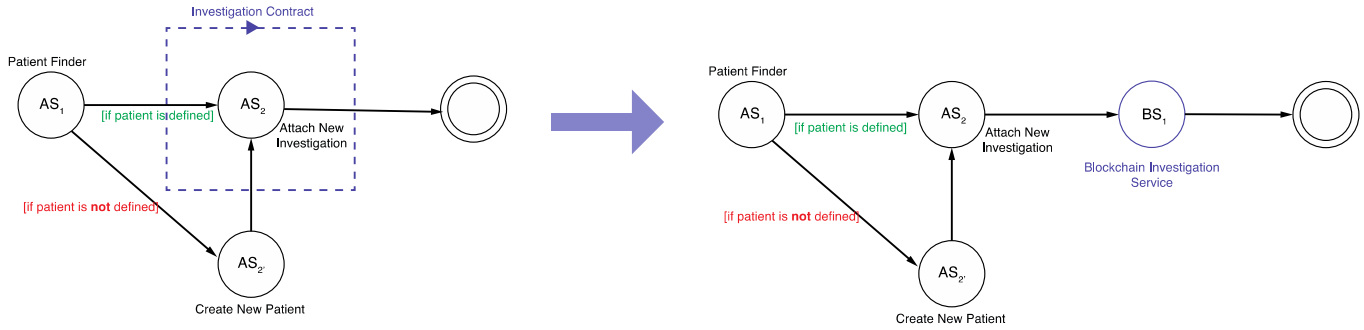


Fig. 6. A Proposed Modeling for the Investigation Process using the AFA.

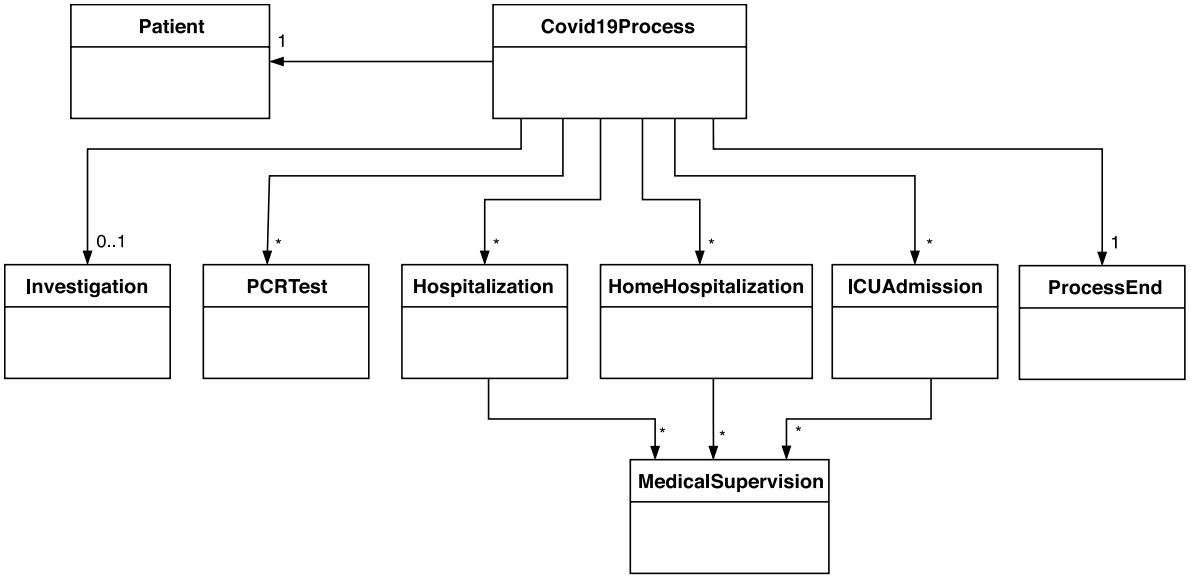
In the same modeling perspective, we illustrate in Fig. 6 a simple modeling example of interacting services that describes the investigation process. The service composition described in the automaton allows the creation of a new investigation for a patient. In this AFA modeling, we design the desired FRs using state interactions, in addition to the new scope defining the blockchain-oriented behavioral NFR. Thus, the concrete service composition integrates a service invocation that sends a transaction to the blockchain. Consequently, in addition to the usual persistence of the investigation data in the provider's internal information system, a new transaction will be executed in the blockchain, and propagated to local copies of all contributing participants.

The provided automata formalization uses descriptive names or regular expressions to index behavioral NFR [7]. In the context of blockchain, we integrate the possibility of performing the NFR-lookup operation taking as a parameter the label used to index the transaction or the smart contract to execute. In the provided example, a forward-direction arrow is used in the *InvestigationContract*, which means that the transaction service fulfilling the behavioral NFR should be executed after the scoped service. This integration implies a lookup operation performed in the behavioral registry, returning the concrete service metadata that matches the specified query.

C. The Asset Modeling

Recall that assets constitute the main resources in Hyperledger technology; they can range from tangible to intangible entities. Hyperledger Fabric provides the ability to modify assets using chaincode transactions. They are represented in Hyperledger Fabric as a collection of key-value pairs and can be represented in binary and/or JSON forms. In Composer, assets are defined in .cto files using a structure similar to the class structuration in object-oriented programming. In our case study, we used assets to model all phases of the COVID-19 healthcare processes.

Our case study modeling based on the presented choreography diagram can be defined in a fine-grained manner using appropriate assets. This allows us to design the desired information system exhaustively, and consequently, improve inter-organizational collaboration. Fig. 7 presents some asset definitions extracted from the modeling file *ma.gov.ehealthcare.assets.cto*. In the same figure we also provide the class diagram illustrating the relationships between the presented assets. In the current phase of implementation, the patient is integrated as an asset, which constitutes a subject and not an interacting participant.




```
namespace ma.gov.ehealthcare.assets
import ma.gov.ehealthcare.concepts.*
import ma.gov.ehealthcare.organizations.*
import ma.gov.ehealthcare.participants.*

asset Patient identified by govId {
  o String govId
  o String EHRId
  o PersonalInformation personalInformation
}

asset Investigation identified by investigationId {
  o String investigationId
  o DateTime investigationDate
  --> HealthcareOrganization healthcareOrganization
  --> Doctor investigator
}

asset PCRTest identified by PCRId {
  o String PCRId
  o DateTime sampleDate
  o DateTime resultDate optional
  o PCRResult PCRResult optional
  o PCRPhase PCRPhase
  --> HealthcareOrganization healthcareOrganization
}

asset Hospitalization identified by hospitalizationId {
  o String hospitalizationId
  o DateTime hospitalizationStartDate
  o DateTime hospitalizationEndDate
  o HealthState entryHealthState
  --> HealthcareProvider healthcareProvider
  --> Doctor authorizer
}

asset ICUAdmission identified by ICUAdmissionId {
  o String ICUAdmissionId
  o DateTime ICUAdmissionStartDate
  o DateTime ICUAdmissionEndDate
  o HealthState entryHealthState
  --> HealthcareProvider healthcareProvider
  --> Doctor authorizer
}

asset Covid19Process identified by covid19ProcessId {
  o String covid19ProcessId
  --> Patient patient
  --> Investigation investigation
  --> PCRTest[] laboratoryTests optional
  --> Hospitalization[] hospitalizations optional
  --> ICUAdmission[] ICUAdmissions optional
  --> HomeHospitalization[] homeHospitalizations optional
  --> HealthcareProviderTransfert[] healthcareProviderTransferts
optional
  --> ProcessEnd processEnd
}

asset HealthcareProviderTransfert identified by healthcareProvider-
TransfertId {
  o String healthcareProviderTransfertId
  o Boolean acknowledgement
  o DateTime transfertDate
  o HealthState currentHealthState
  --> HealthcareOrganization healthcareRequester
  --> HealthcareOrganization healthcareReceiver
  --> Doctor transfertRequester
}

asset DailyPCRTestStatistics identified by dailyPCRTestStatisticsId {
  o String dailyPCRTestStatisticsId
  o DateTime endDate
  o Integer dailyPCRTestNumber
  --> Region region
}

asset HomeHospitalizationidentified …
asset ImagingExam …
asset MedicalSupervision …
asset ProcessEnd …
asset DailyPositiveCaseStatistics …
asset DailyNegativeCaseStatistics …
asset DailyRecoveryCaseStatistics …
asset DailyDeathCaseStatistics …
asset DailyICUAdmissionsStatistics …
```

Fig. 7. A Proposed Example of Asset Definitions of COVID-19 Processes and their Appropriate Class Diagram Modeling.

The main assets constituting the COVID-19 healthcare processes are the following:

1) The investigation asset usually constitutes the first point of contact between the patient and the healthcare provider, allowing a doctor to investigate the patient's health state, symptoms and medical history for the recent period before his/her admission by the healthcare organization.

2) The PCR test asset gathers all PCR-related information and constitutes the key element in defining the consequent phases according to the test results. In other words, in addition to other factors, a positive result may imply a patient hospitalization, whereas a negative result implies a non-affection for diagnosis tests or a patient recovery for control tests.

3) The hospitalization asset describes the process of taking care of a patient in a healthcare organization during a period of time according to medical instructions related to his/her health state.

4) The home hospitalization asset defines the newly introduced process of taking care of a patient remotely in his/her home through continuous supervision based on telecommunication and regularly filled reports.

5) The ICU admission asset defines the process of transferring a patient internally to the intensive care unit (ICU). This transfer depends on the health status of the patient and the availability of the ICU room.

6) The medical supervision asset describes the operation of performing regular medical checks in order to follow up and update the patient's health state.

On the other hand, in order to keep track of all participating organizations in each phase of the workflow, we adopt a model that considers collaborating organizations as assets. This allows us to identify the healthcare providers responsible for any performed action, thereby improving supervision and auditability. In each region, these healthcare organizations communicate with the REC, which constitutes their higher focal authority. In their turn, RECs are continuously reporting to the NEC. The collaborating organizations defined using assets in the studied scenario are the following:

1) The healthcare organization asset is the parent of the overall organizational assets, containing common attributes such as the identifier, name, address, and region.

2) The healthcare provider asset inherits from the healthcare organization asset and gathers all key information regarding the provider's capacities in terms of hospitalization, laboratory tests, and imaging examinations.

3) The public healthcare center asset inherits from the healthcare provider and unites both university healthcare centers and regional healthcare centers.

4) The private clinical center asset inherits from the healthcare provider asset, and defines the external private healthcare entities collaborating with the public healthcare organizations during the pandemic.

5) The private imaging center asset inherits from the healthcare organization and defines the key capacity information for the main imaging examinations related to COVID19 affections.

6) The private laboratory asset inherits from the healthcare organization and defines the laboratory daily PCR exam capacity supported during the pandemic.

7) The regional epidemiological center asset inherits from the healthcare organization and defines the administrative entity responsible for collecting, processing, analyzing and reporting all epidemiological-related information through the built collaborative system.

8) The national epidemiological center asset inherits from the healthcare organization and defines the head entity responsible for collecting, processing, analyzing and reporting all epidemiological-related information gathered from RECs through the built collaborative system.

D. The Participants

Participants are the members of a business network. They constitute the main actors of the blockchain, managing assets and submitting transactions. For each participant, one or more identities can be assigned, allowing him/her, depending on the attributed rules, to perform actions and interact with the ledger. In the first phase of this project, we omit adding the patient as a participant, as in this initial phase of the project, we consider only inter-organizational collaborative processes, which excludes patient interactions. In other words, all patient data are updated by coordinators of healthcare organizations that they interact with, such as doctors, healthcare provider coordinators, etc.

We introduce different role definitions, which all inherit from the healthcare participant role. They are listed below:

1) The healthcare provider coordinator role defines the main actor representing his/her *healthcare organization*. It allows to manage internal data and also to update the ledger. We link this role to *the healthcare provider's asset* to easily identify the organization of each coordinator.

2) The regional epidemiological center coordinator plays a key role in our case study. On the one hand, this participant is responsible for managing and supervising all blockchain information related to his/her attributed region. On the other hand, it allows interaction with the NEC through the main national channel.

3) The national epidemiological center coordinator role allows users of the healthcare ministry to investigate and follow the daily summaries received through the national channel. It has higher administrative privileges allowing to manage participants, organizations, and entities on a national scale.

4) The laboratory coordinator role allows interaction with the ledger as a coordinator of a private laboratory, in order to update the patients' COVID-19 processes by integrating information related to PCR tests and their respective results.

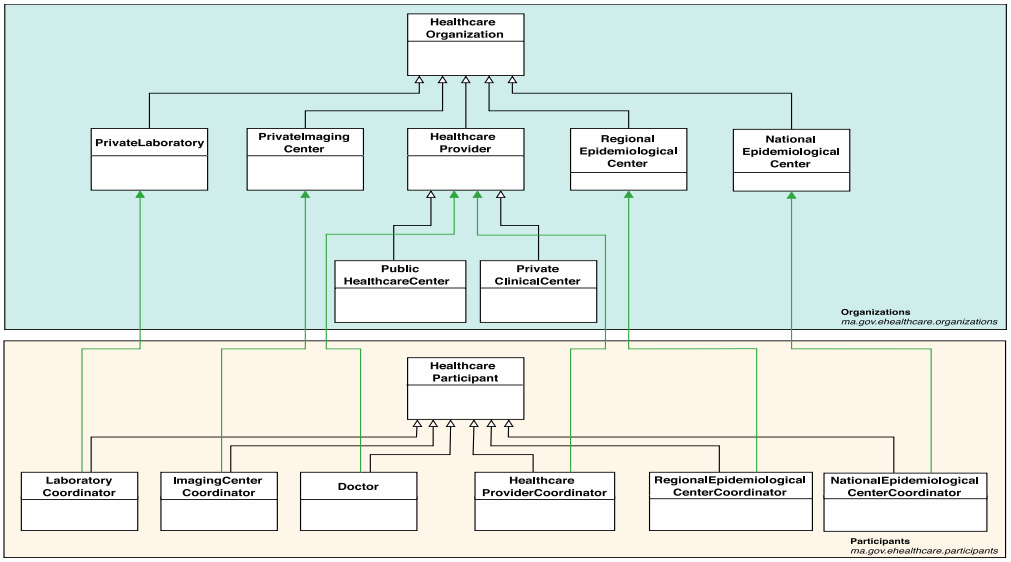
5) The imaging center coordinator defines the role assigned to the participants responsible for coordinating the private imaging centers. Their main contribution consists of

integrating all affected patients' imaging data in the blockchain, such as chest X-ray, CT scan, or MRI results.

In Fig. 8, we present some examples of collaborating organizations and their associated participants supported by the appropriate class diagram.

E. The Transactions

Transactions are the mechanisms by which participants interact with assets. They define the core business logic of the process. Once executed, the state of the associated assets changes. Consequently, the world state can be recreated by replaying all transactions with respect to their accurate order, which constitutes one of the strengths of blockchain technology.




```
namespace ma.gov.ehealthcare.organizations
import ma.gov.ehealthcare.concepts.*
import ma.gov.ehealthcare.assets.*

asset HealthcareOrganization identified by
healthcareOrganizationId{
    o String healthcareOrganizationId
    o String healthcareOrganizationName
    o Address healthcareOrganizationAddress
    --> Region region optional
}

asset HealthcareProvider extends HealthcareOrganization{
    o Integer capacity
    o Integer dailyPCRSampleCapacity optional
    o Integer dailyChestXRayCapacity optional
    o Integer dailyCTScanCapacity optional
    o Integer dailyMRICapacity optional
}

asset PublicHealthcareCenter extends HealthcareProvider{
    o HealthcareCenterType type
}

asset PrivateImagingCenter extends HealthCareOrganization{
    o Integer dailyChestXRayCapacity
    o Integer dailyCTScanCapacity
    o Integer dailyMRICapacity
}

asset PrivateLaboratory extends HealthCareOrganization{
    o Integer dailyPCRSampleCapacity
}

asset PrivateClinicalCenter extends HealthcareProvider …
asset PrivateLaboratory extends HealthcareOrganization …
asset RegionalEpidemiologicalCenter extends
HealthcareOrganization …
asset NationalEpidemiologicalCenter extends
HealthcareOrganization …
```

```
namespace ma.gov.ehealthcare.participants
import ma.gov.ehealthcare.concepts.*
import ma.gov.ehealthcare.organizations.*

abstract participant HealthcareParticipant identified by
participantId {
  o String participantId
  o PersonalInformation personalInformation
}

participant HealthcareProviderCoordinator extends
HealthcareParticipant {
  --> HealthcareProvider healthcareProvider
}

participant RegionalEpidemiologicalCenterCoordinator extends
Healthcare-Participant {
  --> RegionalEpidemiologicalCenter
regionalEpidemiologicalCenter
}

participant NationalEpidemiologicalCenterCoordinator extends
Healthcare-Participant {
  --> NationalEpidemiologicalCenter
nationalEpidemiologicalCenter
}

participant LaboratoryCoordinator extends
HealthcareParticipant {
  --> PrivateLaboratory laboratory
}

participant ImagingCenterCoordinator extends
HealthcareParticipant {
  --> PrivateLaboratory laboratory

participant Doctor extends HealthcareParticipant {
  o Boolean isSpecialist
  o Speciality speciality optional
  --> HealthcareProvider healthcareProvider
}
```

Fig. 8. A Proposed Example of Organizations and Participants Assets and their Appropriate Class Diagram Modeling.

We use transactions to populate blockchain assets using organizational and healthcare data. First, we use our proposed automata-based approach to illustrate the modeling of each business process to produce a valid abstract service composition. Then, we integrate the appropriate blockchain transactions using behavioral scopes associated to the abstract services. In Fig. 9, we present a model of the patient investigation phase, which consists of collecting patient data, such as health state, symptoms, and comorbidities, etc. According to patient collected data, the investigating doctor estimates if there is a risk that patient might be affected, and consequently redirects him to the appropriate service for a PCR test or for an imaging examination. On the other hand, the *healthcare provider coordinator* performs a lookup to verify whether the patient already has an EHR, and creates a new patient profile otherwise. The second step consists of creating a specific *COVID-19 process* blockchain asset to assign to the patient if he/she interacts for the first time with the *healthcare organization*. All these scenarios are modeled as part of the investigation process, where states identify the abstract services described as follows: The states outlined with a black color define the abstract services denoted AS. Each of these abstract services will be associated with a pool of concrete services fulfilling the desired FRs defined in the AFA. The states outlined in blue are abstract services fulfilling behavioral blockchain NFRs of requesting data from the ledger on a read-only basis. Thus, their associated concrete services will interact with the blockchain, but will not impact the world state of the ledger. Finally, the states outlined in green constitute the abstract services defining blockchain transactions, and whose concrete services have the ability to update the world state of the ledger.

We provide below a brief description of the provided Automata modeling example:

- The concrete service CS1 associated with the first abstract service AS1 (Patient Finder) allows to return, from the internal information system, the unique electronic health record (EHR) for a given patient, using a set of personal information as input parameters.
- The concrete services CS2/CS2' associated with the abstract services AS2 and AS2' respectively, are conditioned by the result of execution of the first concrete service CS1 during the runtime. Thus, if CS1 returns a valid patient, the concrete service CS2 will be executed to fulfill the query "Covid Process Finder". In this case, CS2 will only interact with the blockchain on a read-only basis to obtain the requested data. However, in the second case, the concrete service CS2' will be executed to fulfill the query "New Covid Process and Patient Assignment". In this case, CS2' will call a blockchain transaction to create a Covid Process asset with a nested patient asset filled with the patient information.
- The concrete service CS3 fulfills the requirements of the AS3 associated with the query "Investigation Assignment". This service will perform a blockchain transaction to create a new investigation asset and assign it to the previously returned Covid Process asset.

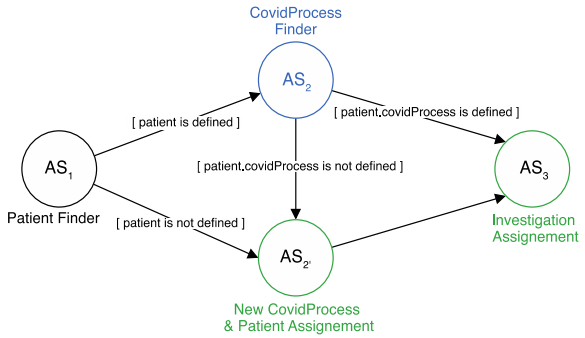


Fig. 9. The Proposed Automata Modeling of the Investigation Process Including Blockchain-Oriented Abstract Services.

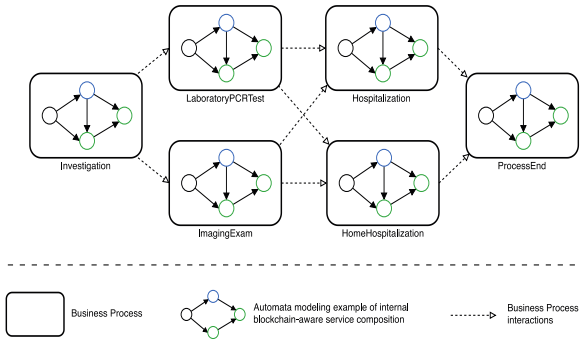


Fig. 10. Simplified Representation of the Process Choreography and Associated Internal Service-Oriented Automata.

Similarly, we extend the same modeling to design the remaining processes. In Fig. 10, we describe a simplified example representing how collaborating business processes can be seen from a wider and closer perspective, combining both FSA modeling and choreography diagrams.

Thus, in practice once a busines process is required, the system executes the appropriate service composition, which includes regular business-oriented services and blockchain-oriented services injected as behavioral NFRs in the designed AFA.

Finally, in the context of our proposed case study, we present in Table II a comparative table describing the implementation and collaboration aspects of some of the main COVID-19 operations. It specifies how these operations are implemented, and whether (1) they allow an automatic collaboration with external organizations or (2) are limited to the internal information systems and are shared manually with external organizations through messaging or data transfers. Most public healthcare providers are using local web applications, handling common operations regarding hospitalizations, medical consultations and exams. This application is exhaustive for handling medical data, however, does not cover the specificities of some important COVID-19 operations, such as investigation which comes with a long questionary storing meticulous information about patient's heath state, laboratory and imaging exams data which are handled using complementary applications connected to automates, etc. Accordingly, the healthcare ministry deployed urgently a new COVID19-oriented web application to gather investigation information and daily statistics. These applications

are not connected to each other, and do not provide web services or APIs to allow their interoperability. Thus, an automatic collaboration is only available using the web application that can be accessed by multiple participants and roles to handle investigation, patient health state, recovery/death and statistic data. For other daily operations not covered by any of implemented applications, manual registration is used then to ensure follow-ups and data transfers to higher organizations. The proposed contribution based on service-oriented collaboration and using decentralized ledgers to parallelly store and update information overcomes all these collaboration limitations, and provides accessible operations through blockchain transactions, with respect to the implemented access-control policies and regional channel configuration.

F. The Events

Events in Hyperledger Fabric, as for their common use in software engineering, provide a way to interact with external components or systems. In Hyperledger Composer, they are defined in the same way as assets or participants, and can be emitted by transactions to indicate to external systems that something of importance has happened to the ledger. Applications can then subscribe to events through the composer-client API [57]. In our context, events are mainly associated with the blockchain services updating the ledger, outlined in green in the described automata model. In the present case study, events allow the notification of the higher authority of updates regarding the overall healthcare status related to COVID-19. For instance, transactions launched by healthcare organizations trigger events to update statistics and notify the REC. Similarly, at a higher level, transactions launched by the RECs trigger events notifying the NEC of updates regarding the epidemiological status in their respective regions.

TABLE II. A RECAPITULATIVE TABLE CLASSIFYING COLLABORATION LIMITATIONS FOR SOME OF COVID-19 PROCESS ACTIONS

Action	**Organization**	**Initial Collaborative Environment**					**Current Implementation**	
		Local Application	Complementary Documents/Apps	COVID-19 Web Application	On demand Collaboration (Messaging, Data Transfers)	Automatic Collaboration	Decentralized Service-oriented Solution	Automatic Collaboration
COVID19 Investigation	Healthcare Provider	X	✓	✓	X	✓	✓	✓
Laboratory PCR Data	Healthcare Provider	Partial	✓	X	✓	X	✓	✓
	Private Laboratory	✓	X	X	✓	X	✓	✓
Laboratory PCR Results	Healthcare Provider	X	✓	✓	X	✓	✓	✓
	Private Laboratory	✓	X	On approval	X	On approval	✓	✓
Hospitalization	Healthcare Provider	✓	X	X	✓	X	✓	✓
Medical Supervision History	Healthcare Provider	X	✓	X	✓	X	✓	✓
Imaging Exam	Private Imaging Center	✓	X	✓	X	✓	✓	✓
Health Status History	Healthcare Provider	X	✓	X	✓	X	✓	✓
Home Hospitalization	REC	X	✓	X	✓	X	✓	✓
ICU Admission	Healthcare Provider	✓	X	X	✓	X	✓	✓
External Transfer	Healthcare Provider	Partial	✓	X	✓	X	✓	✓
Process End (Recovery)	Healthcare Provider	✓	X	✓	X	✓	✓	✓
Process End (Death)	Healthcare Provider	Partial	✓	✓	X	✓	✓	✓
Data Anonymization Process		X		X			✓	
Access Control – Application Level		✓		✓			✓	
Access Control – Module Level		✓		✓			✓	
Access Control – Action Level		Partial		-			✓	
Access Control – Resource/Data Level		X		-			✓	
Channel Restrictions		X		-			✓	

To make our proposed implementation more exhaustive, we implement some statistic-oriented assets allowing the COVID-19 statistics to be updated daily for each region. These assets are automatically updated when appropriate events are triggered. In our case study, we use events to notify REC coordinators about updates regarding COVID-19 daily status in their respective regions. Another concrete example consists of using events to notify a target healthcare provider about a transfer request for a patient in a critical health state, where the requesting participant receives an acknowledgement as soon as the target provider validates the transfer.

Finally, in a service-oriented approach ensuring a choreography between different organizations, events constitute an essential pillar allowing an efficient collaboration, and facilitating the implementation of real-time notification systems to increase reactivity towards critical scenarios. In Fig. 11, we provide a definition of the investigation transaction associated with an event emission to subscribed components.

G. Access Control and Security Implementation

Hyperledger proposes a high-end permissioned system that allows to manage and control participants' permissions for each action. This constitutes the backbone and main purpose behind using an identity-based permissioned blockchain. Hyperledger provides an access control language based on declarative permissions over each element of the domain model. Access control rules are defined in specific files and are described using two different methods. The first definition type, which is the simplest, allows to control the access to a namespace, an asset, or an asset's property for a participant type or participant instance. The second type is a more advanced conditional rules system, based on JavaScript Boolean expressions. These expressions are evaluated in runtime to allow or deny access to a resource for a participant. Participants can have their access to transactions restricted based on their role's permissions defined in the access control files. In our perception, collaborating organizations and their attributed participants are dynamically managed. Identities and business network cards can be managed under administrative privileges through the Hyperledger API. The roles of coordinators for both regional and national epidemiological centers are given the ability to manage participants, identities and network cards, for their regional and national organizations, respectively.

In the present case study, we notice that none of the essential assets constituting the main COVID-19 process contain or point to the patient's assets. The aim behind omitting this relationship is to enhance confidentiality and ensure data anonymization, which constitutes an essential feature implemented in our contribution, as highlighted in Table II. Accordingly, participants with restricted permissions can only access and handle the data of assets associated to their scope. For example, they can handle COVID-19 hospitalization information, PCR results information or statistics information without accessing the corresponding patients' data (e.g., personal information). This modeling-based separation of assets, combined with the fine-grained layers of access-control management provided by the Hyperledger Fabric blockchain, constitute the backbone of security attributes implemented in this contribution, in order to provide a meticulous management of roles and permissions for overall participants, according to their administrative attributions.

On the other hand, and in order to meet the ambitions and efforts of the Moroccan government to deploy the advanced regional decentralization, we aimed in the proposed case study to use an exclusive channel for each individual region to promote data privacy. From our perspective, these channels allow to concentrate the efforts of regional epidemiological centers coordinators to securely manage the permissions of collaborating organizations falling under their administrative perimeters. Similarly, a national channel is implemented, allowing communication between all RECs with the ministry entities (e.g., National Epidemiological Center, Ministry Secretary). The aim of this channel is to ensure continuous data forwarding and reporting based on the daily statistics generated for each region.

```
/**
 * Create Investigation Transaction
 * @param
   {ma.gov.ehealthcare.assets.CreateInvestigation}
investigationData
 * @transaction
 */

function    createInvestigation(investigationData) {

  // 1. Get the asset registry
  return
getAssetRegistry('ma.gov.ehealthcare.assets.Investigation')
    .then(function(investigationRegistry){

        // 2. Get resource factory
        var factory = getFactory();
        var NS = 'ma.gov.ehealthcare.assets';

        // 3. Create the Resource instance
        var investigationId = generateTxId();

        var investigation = factory.newResource(
          NS, 'Investigation', investigationId);

        // 4. Set the data and relationships
        investigation.investigationDate = new Date(
          investigationData.investigationDate);
       var doctorRelationship = factory.newRelationship(
          'ma.gov.ehealthcare.participants','Doctor',
          investigationData.investigator.participantId);

        investigation.investigator = doctorRelationship;

        var healthcareOrganizationRelationship =
          factory.newRelationship(
          'ma.gov.ehealthcare.organizations',
          'HealthcareOrganization',
          investigationData.healthcareOrganization.healthcar
          eOrganizationId);

        // 5. Emit the event InvestigationCreated
        var event = factory.newEvent(NS,
'InvestigationCreated');
        event.investigationId = investigationId;
        emit(event);

        return
investigationRegistry.addAll([investigation]);
    });
}
```

Fig. 11. A Definition Example of the Investigation Transaction and the Associated Emitted Event.

VI. Conclusion and Future Work

In this study, we aimed to provide a complete solution for the design and implementation of inter-organizational collaborative processes based on a two-phase modeling using BPMN's choreography diagram and FSA. We provided the ability to integrate blockchain transactions as services to promote decentralization, immutability, integrity, and trustless interactions between collaborating organizations. To illustrate our approach, we presented a proof-of-concept implementation adapted to the healthcare domain, and handling COVID19-related collaborative processes. Although we opted for Hyperledger Fabric as a permissioned blockchain platform to meet the collaboration requirements of the studied scenario, the introduced blockchain as a service concept can be adapted to any blockchain platform as long as it supports the automation of transaction calls. Our proposed proof-of-concept modeling is based on real business processes aiming to promote collaboration between healthcare organizations in Morocco.

Using an architecture based on a specific blockchain topology will always rely on its limitations in term of performance, scalability, security, and other key aspects characterizing the chosen blockchain platforms. On the other hand, the misuse of permissioned blockchains can cause colossal damage, especially in the context of inter-organizational collaboration, as the reliability of the ledger relies on the integrity of its members. Thus, identity and role management are key features implemented to ensure healthy collaborations. In this contribution, the proposed service-oriented approach does not rely on any specific blockchain topology and can integrate both permissioned and permissionless blockchain platforms, since the execution of a transaction or a smart contract is wrapped into a concrete service.

In future work, we will focus on two principal aspects: (1) Covering intra-organizational collaboration by encouraging progressive migration of internal information systems toward the blockchain, requiring the deployment of new blockchain transactions as concrete services and an exhaustive definition of interacting assets using the proposed modeling process. (2) Integrating end-users (e.g., patients) as participants in the collaborative ledger by experimenting the integration of a hybrid topology, pairing permissionless and permissioned blockchains. To achieve this objective, numerous challenges arise and are associated, for example, to the correlation of performance with the scalability, shared identity management, and the implementation of unified security requirements while merging the heterogeneous platforms.

References

[1] L. Chung and J. C. S. do Prado Leite, "On Non-Functional Requirements in Software Engineering," in Conceptual Modeling: Foundations and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 363–379.

[2] W. Cai, Z. Wang, J. B. Ernst, Z. Hong, C. Feng, and V. C. M. Leung, "Decentralized Applications: The Blockchain-Empowered Software System," in IEEE Access, vol. 6, pp. 53019-53033, 2018, DOI: 10.1109/ACCESS.2018.2870644.

[3] M. Crosby, P. Pattanayak, S. Verma, and V. Kalyanaraman, "Blockchain technology: Beyond bitcoin", Applied Innovation, vol. 2, pp. 6–10, 2016.

[4] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in 2016 IEEE Symposium on Security and Privacy (SP), 2016.

[5] B. Niehaves and R. Plattfaut, "Collaborative business process management: status quo andquo vadis," in Business Process Management Journal, vol. 17, pp. 384–402, 2011, DOI: https://doi.org/10.1108/14637151111136342.

[6] Q. Chen and M. Hsu, "Inter-enterprise collaborative business process management," in Proceedings of 17th International Conference on Data Engineering, 2002.

[7] I. E. Kassmi, R. Belkeziz, and Z. Jarir, "Deep Attention on Measurable and Behavioral-driven Complete Service Composition Design Process," in International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 3, 2021, DOI: 10.14569/IJACSA.2021.0120377.

[8] C. Cachin, "Architecture of the Hyperledger blockchain fabric," in Workshop on Distributed Cryptocurrencies and Consensus Ledgers (DCCL 2016), Chicago, Jul. 2016.

[9] M.Swan, "Blockchain: Blueprint for a New Economy," Newton, MA, USA: O'Reilly Media, 2015.

[10] M. Pilkington, "Blockchain Technology: Principles and Applications", in Research Handbook on Digital Transformations, 2016.

[11] A. Muddasar and B. Sikha, "Introduction to NFTs: The Future of Digital Collectibles" International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12 no. 10, 2021. DOI: http://dx.doi.org/10.14569/IJACSA.2021.0121007.

[12] A. Reyna, C. Martín, J. Chen, E. Soler, M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," Future Generation Computer Systems, Vol. 88, pp. 173-190, 2018, DOI: https://doi.org/10.1016/j.future.2018.05.046.

[13] A. Jawad, A. Toqeer, M. Shahrulniza and A. Zahrani, "Towards Secure IoT Communication with Smart Contracts in a Blockchain Infrastructure" International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 10, 2018. DOI: http://dx.doi.org/10.14569/IJACSA.2018.091070.

[14] S. Al-Rakhami M. and Al-Mashari M., "Blockchain and Internet of Things for Business Process Management: Theory, Challenges, and Key Success Factors" International Journal of Advanced Computer Science and Applications (IJACSA), vol, 11, no. 10, 2020. DOI: http://dx.doi.org/10.14569/IJACSA.2020.0111069.

[15] Y. Chen and C. Bellavitis, "Blockchain disruption and decentralized finance: The rise of decentralized business models," Journal of Business Venturing Insights, vol. 13, 2020, DOI: https://doi.org/10.1016/j.jbvi.2019.e00151.

[16] S. Solat, P. Calvez, and F. Naït-Abdesselam, "Permissioned vs. Permissionless blockchain: How and why there is only one right choice", in Journal of Software, pp. 95–106, 2021.

[17] S. Pahlajani, A. Kshirsagar, and V. Pachghare, "Survey on private blockchain consensus algorithms", in 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019.

[18] E. Androulaki et al., "Hyperledger fabric: A distributed operating system for permissioned blockchains," in Proceedings of the Thirteenth EuroSys Conference, 2018.

[19] K. Olson, M. Bowman, J. Mitchell, S. Amundson, D. Middleton, and C. Montgomery, "Sawtooth: An introduction", in Hyperledger: Blockchain Technologies for Businesses, 2018.

[20] M. Valenta, P. Sandner, "Comparison of Ethereum Hyperledger Fabric and Corda," in FSBC Working Paper June 2017, Frankfurt School Blockchain Center, 2017.

[21] S. Bano et al., "SoK: Consensus in the age of blockchains," in Proceedings of the 1st ACM Conference on Advances in Financial Technologies, 2019.

[22] W. Wang et al., "A survey on consensus mechanisms and mining strategy management in blockchain networks," IEEE Access, vol. 7, pp. 22328-22370, 2019, doi: 10.1109/ACCESS.2019.2896108.

[23] J. A. Garay, A. Kiayias, and G. Panagiotakos, “Proofs of Work for Blockchain Protocols,” IACR Cryptology ePrint Archive, Report 2017/775, Aug. 2017.

[24] F. Saleh, “Blockchain without Waste: Proof-of-Stake,” The Review of Financial Studies, vol. 34, pp 1156–1190, 2021, DOI: https://doi.org/10.1093/rfs/hhaa075.

[25] S. De Angelis, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri and V. Sassone, “PBFT vs proof-of-authority: applying the CAP theorem to permissioned blockchain,” Italian Conference on Cyber Security, 2018.

[26] M. Castro and B. Liskov, “Practical byzantine fault tolerance and proactive recovery,” ACM Transactions on Computer Systems, vol. 20, no. 4, pp. 398–461, 2002.

[27] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system”, Self-published Paper, May 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[28] V. Buterin, “Ethereum: A next-generation smart contract and decentralized application platform”, Ethereum Foundation, Technical Report, 2014. [Online]. Available: https://github.com/ethereum/wiki/wiki/White-Paper.

[29] J. Mendling et al., “Blockchains for business process management - challenges and opportunities,” ACM Transactions on Management Information Systems, vol. 9, no. 1, pp. 1–16, 2018.

[30] J. A. Garcia-Garcia, N. Sánchez-Gómez, D. Lizcano, M. J. Escalona and T. Wojdyński, "Using Blockchain to Improve Collaborative Business Process Management: Systematic Literature Review," in IEEE Access, vol. 8, pp. 142312-142336, 2020, doi: 10.1109/ACCESS.2020.3013911.

[31] L. D. Xu and W. Viriyasitavat, “Application of Blockchain in Collaborative Internet-of-Things Services,” IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1295–1305, 2019.

[32] I. Weber, X. Xu, R. Riveret, G. Governatori, A. Ponomarev, J. Mendling, “Untrusted business process monitoring and execution using blockchain”. In International Conference on Business Process Management, pp. 329-347, 2016, Springer.

[33] L. García-Bañuelos, A. Ponomarev, M. Dumas, I. Weber, “Optimized execution of business processes on blockchain”. In International conference on business process management, pp. 130-146, 2017, Springer.

[34] F. Corradini, A. Marcelletti, A. Morichetta, A. Polini, B. Re, and F. Tiezzi, “Engineering trustable choreography-based systems using blockchain,” in Proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020.

[35] C. Dannen, “Solidity Programming,” in Introducing Ethereum and Solidity, Berkeley, CA: Apress, 2017, pp. 69–88.

[36] J. Ladleif, M. Weske, and I. Weber, “Modeling and enforcing blockchain-based choreographies,” in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 69–85.

[37] C. Di Ciccio et al., “Blockchain support for collaborative business processes,” Informatik Spektrum, vol. 42, no. 3, pp. 182–190, 2019.

[38] A. B. Tran, Q. Lu and I. Weber, "Lorikeet: A model-driven engineering tool for blockchain-based business process execution and asset management", CEUR Workshop Proceedings, 2018.

[39] O. López-Pintado, L. García-Bañuelos, M. Dumas, I. Weber, and A. Ponomarev, “Caterpillar: A business process execution engine on the Ethereum blockchain,” Software: Practice and Experience, vol. 49, no. 7, 2019.

[40] V. Pourheidari, S. Rouhani, and R. Deters, “A Case Study of Execution of Untrusted Business Process on Permissioned Blockchain,” in 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2018.

[41] M. Autili, F. Gallo, P. Inverardi, C. Pompilio, and M. Tivoli, “Introducing Trust in Service-oriented Distributed Systems through Blockchain,” in 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2019.

[42] A. Kernahan, U. Bernskov and R. Beck (2021). “Blockchain out of the Box – Where is the Blockchain in Blockchain-as-a-Service?”, 54th Hawaii International Conference on System Sciences, 2021, DOI: 10.24251/HICSS.2021.520.

[43] L. Daming, D. Lianbing, C. Zhiming and S. Alireza, “Blockchain as a service models in the Internet of Things management: Systematic review,” Transactions on Emerging Telecommunications Technologies, 2020, DOI: https://doi.org/10.1002/ett.4139.

[44] W. Zheng, Z. Zheng, X. Chen, K. Dai, P. Li and R. Chen, "NutBaaS: A Blockchain-as-a-Service Platform," in IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2941905.

[45] Q. Lu, X. Xu, Y. Liu, I. Weber, L. Zhu and W. Zhang, “uBaaS: A unified blockchain as a service platform,” Future Generation Computer Systems, vol. 101, pp. 564-575, 2019, DOI: https://doi.org/10.1016/j.future.2019.05.051.

[46] I. El Kassmi and Z. Jarir, “Measurable and Behavioral Non-Functional Requirements in Web Service Composition,” in Handbook of Research on Contemporary Perspectives on Web-Based Systems, pp. 340-361, 2018, IGI Global, DOI: http://doi:10.4018/978-1-5225-5384-7.ch015.

[47] I. El Kassmi and Z. Jarir, "Towards security and privacy in dynamic web service composition," 2015 Third World Conference on Complex Systems, 2015, doi: 10.1109/ICoCS.2015.7483260.

[48] Hyperledger Fabric: Main Documentation [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/whatis.html (Visited on 02/02/2021).

[49] Hyperledger Fabric: Channel Capabilities [Online]. Available: https://github.com/hyperledger/fabric/blob/release-2.2/docs/source/capabilities_concept.md (Visited on 02/21/2021).

[50] W. Fu, X. Wei and S. Tong, “An Improved Blockchain Consensus Algorithm Based on Raft,” Arabic Journal for Science and Engineering, 2021, DOI: https://doi.org/10.1007/s13369-021-05427-8.

[51] E. Kinory, S. S. Smith, and K. S. Church, “Exploring the playground: Blockchain prototype use cases with Hyperledger Composer,” Journal of Emerging Technologies in Accounting, vol. 17, no. 1, pp. 77–88, 2020.

[52] Hyperledger Fabric: Smart Contracts and Chaincode [Online]. Available: https://github.com/hyperledger/fabric/blob/release-2.2.

[53] /docs/source/smartcontract/smartcontract.md (Visited on 02/23/2021).

[54] J. Xu et al., "Healthchain: A Blockchain-Based Privacy Preserving Scheme for Large-Scale Health Data," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8770-8781, Oct. 2019, DOI: 10.1109/JIOT.2019.2923525.

[55] S. H. Alsamhi and B. Lee, “Blockchain-empowered multi-robot collaboration to fight COVID-19 and future pandemics,” IEEE Access, vol. 9, pp. 44173–44197, 2021.

[56] D. Marbouh et al., “Blockchain for COVID-19: Review, opportunities, and a trusted tracking system,” Arabian Journal for Science and Engineering, vol. 45, no. 12, pp. 1–17, 2020.

[57] A. A. Abd-alrazaq et al., “Blockchain technologies to mitigate COVID-19 challenges: A scoping review,” Computer Methods and Programs in Biomedicine Update, vol. 1, p. 100001, 2021.

[58] Hyperledger Fabric: Peer channel-based event services [Online]. Available: https://github.com/hyperledger/fabric/blob/release-2.2/docs/source/peer_event_services.rst (Visited on 03/01/2021).

A Language Tutoring Tool based on AI and Paraphrase Detection

Anas Basalamah
Computer Engineering Department
Umm Al-Qura University

***Abstract*—A language tutoring tool (LTT) helps learning a language through casual human-like conversations. Natural language understanding (NLU) and natural language generation (NLG) are two key components of an LTT. In this paper, we propose a paraphrase detection algorithm that is used as the building block of the NLU. Our proposed tree-LSTM with a self-attention method for paraphrase detection shows accuracy of 87% with a lower parameter of 6.5m, which is much robust and lighter than the other existing paraphrase detection algorithms. Furthermore, we discuss an LTT prototype using the proposed algorithm with having some featured components like- message analysis, grammar detection, dialogue management, and response generation component. Each component is discussed in detail in the methodology section of this paper.**

***Keywords*—LTT; NLG; NLU; paraphrase detection; LSTM**

I. Introduction

The conversation is an effective technique for learning a second language [1]. When people try to learn a new language, they learn it through verbal and written communication, including reading, writing, listening, and speaking skills. The former includes listening and speaking, and the latter includes reading and writing. A partner is mandatory for whatever conversation a learner wants to make. It is easy these days to make friends from different countries, cultures, and tribes with the rise of the internet and social media. People can easily share their cultures and values and can teach their languages to others. However, learning a second language requires regular practice for a long time. It isn't easy to find someone available for live conversation every day for a long time because of different time-zone, work schedules, and many other reasons. Although movies, prerecorded videos, or online courses that provide video lessons can be useful equipment to some extent for understanding a language but not adequate for making conversations.

In recent times, smartphones and computers with internet connection are available to everyone from developed to emerging countries. What if someone develops an LTT that can act as a language expert and help the learner with grammar errors, spelling or syntax errors, dictionary and sentence corrections. People can speak and chat with it at their convenient time to learn a new language. For example, if a traveler visits a country without knowing its languages, it becomes difficult to communicate with the local people. This LTT can help the traveler in such a case. Many researchers have already researched on various conversation applications for both pedagogical and industrial purpose [2], [3], [4] but the idea of language specific LTTs (language other than English) is relatively new in this domain.

The concept of an LTT comes from the $chatbots$ providing 24-hours customer support or technical support on online platforms [2]. $Chatbots$ are widely used to serve as a customer service agent in many web-platforms. However, an AI-based LTT is different in functionality comparing to a $chatbot$. A $chatbot$ is only capable of conversation related to customer service management whether an AI-based LTT performs conversation related to language tutoring. The LTT consistently learns through conversations with the user. It is a game changer as a language tutor and impacts significantly in learning a second language effectively.

Generally, human agents provide service to the customers in online platforms, but recently they are getting replaced by $chatbots$. Because a $chatbot$ saves time, money, energy, and physical costs for the company. Besides, it generates responses with fewer grammatical errors and typos. Since the text-based conversation part is also a feature of the conversation agent, it is important to adapt the mechanism of a $chatbot$ in the development of a conversation agent. However, it is never easy to develop such a $chatbot$ system to continue a meaningful conversation with customers and solve their problems. The critical issue here is to detect different sentences with the same meaning, i.e., paraphrase detection. Naturally, people use different expressions and structures of sentences to express the same thing when they make conversation in different circumstances. For example-

Sentence-1: "I have to learn this language very quickly."

Sentence-2: "I need to pick up this language as soon as possible."
Clearly, 'Sentence-1' and 'Sentence-2' have two distinct structures but express the same meaning. Paraphrase detection and its natural language understanding (NLU) component play an essential role in detecting these sentences. Therefore, paraphrase detection is considered a vital element of a conversation agent or a dialogue system. The next important step is natural language generation (NLG), as the system has to respond to customers after understanding their problems. This paper follows these NLU and NLG steps for making a conversation agent that acts as a language learning tool to the user.

We develop an AI-based human-like interactive system that can understand the context of natural language sentences and help learn a new language. When the system is trained on a massive dataset of semantic relatedness, it retains the relative behavior of words and vocabulary in the sentence.

The developed application is efficient for people learning or exploring a second language. A novel tree-LSTM with self-attention is proposed that is implemented as the system core of the developed application. The proposed tree-LSTM with self-attention shows moderate accuracy than the existing models but offers the best efficacy as a lightweight and robust model. Performance comparison of the proposed and existing methods are demonstrated in Section III.

The remaining part of the paper is structured as follows. The methodology of our proposed LTT with its demonstration, components, architecture, and workflow is discussed in Section II. The performance of the proposed method is analyzed and compared in section III. Previous research works on language learning agent and paraphrase detection are studied in Section IV, and finally, the paper is concluded in Section V.

II. Methodology

This section describes the components, architecture, and workflow of the LTT. In learning, tourism, visiting, hosteling, and exposing to new places, LTTs are very efficient as the way of interaction is human-like. Additionally, the mapping of sentences to intent makes it more flexible for users to interact by using any phrases. Contextual conversation makes the user or learner feel more active and realistic. A user-LTT sample communication is depicted in Fig. 1. Fig. 1 illustrates how the grammar detection component of an LTT corrects grammatical errors and helps with dictionary.

The primary goal is to create a task-oriented AI-based contextually aware LTT which can interact with humans naturally. This research proposes and trains the system to learn a second language through chat or voice LTT.

Instead of traditional language learning platforms or apps, the LTT will communicate with the learner. It is inspired by the natural phenomenon that humans can learn a new language better when he visits a new place and communicates with natives. In this situation, the primary concern of the visitor is to convey his intention through a message. This case study motivates to create a contextually aware platform that can talk, understand and reply to the user more personally and intelligently. Unlike Google assistant or Alexa, the system does not need to be a knowledge-based assistant; it will be solely trained for language practicing. User will share his feeling: e.g., "how was the day, what he took in the lunch, what he observed today" in second language, and the bot would extract the context, understand the user and reply appropriately. The proposed system can recognize the same sentence in multiple ways as it is trained to paraphrase detection. The language learning contextual system would teach the language to the learner by analyzing the input sentence and informing the user about sentence accuracy.

Fig. 2 represents the machine learning-based architecture of the LTT. The message analysis component (as NLU) extracts entities and classifies intents. This module recognizes the intention of the user and the semantic of the sentence. The grammar detection component checks grammar and punctuation errors, correct sentences, and acts as a dictionary. The dialogue management module tracks the text, implements relevant policy, and triggers the appropriate action based on the database. Finally, the Response generation component provides the corresponding response of the query based on the context. This system communicates with user training on the proposed architecture based on tree-LSTM and attention.

RvNNs are efficient for natural language tasks but require structural input, which is hard for data preparation and implementation. This paper proposes a natural language task-specific model that can compose a tree structure of a plain text. The architecture of the system consists of Tree Long-Short term memory and Attention mechanism. Tree-LSTM is a refined version of RvNN and LSTM that manages the information flow from child to parent more structurally than LSTM. In tree-LSTM, the cell state computes parent representation by utilizing cell dependencies at a vertical distance.

Tree-LSTM extracts information from unstructured sentences by creating a structure of input dynamically. The information about the structure of the input is not provided to the Tree-LSTM model. Therefore, an extra process is required to compose task-specific tree structures of the input sentence. In the equation, $q.h$ express the score of representation validity $r = [h; c]$.

The trainable composition query vector is required to measure the representation validity for building the tree structure from the unstructured sentence. At the first layer, the candidacy for the parent representation is computed using equations 1 to 3. Query vector calculates the validity score for each candidate. Then, the validity of each candidate is prioritized based on their score. During training, the softmax estimator samples the parent node among candidates weighted by the query vectors, and for the testing, the model itself selects the candidate with the highest validity.

$$\begin{bmatrix} i \\ f_l \\ f_r \\ o \\ g \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w_{comp} \begin{bmatrix} h_l \\ h_r \end{bmatrix} + b_{comp} \right) \tag{1}$$

$$c_p = f_l \odot c_l + f_r \odot c_r + i \odot g \tag{2}$$

$$h_p = \mathbf{o} \odot tanh(c_p) \tag{3}$$

$$v_i = \frac{exp(q.\tilde{h}_i^{t+1})}{\sum_{j=1}^{M_{t+1}} exp(q.\tilde{h}_j^{t+1})} \tag{4}$$

Additional components are also added to encode the input sentences into vectors dynamically using the bottom-up tree structure technique. We focused on a type-specific attention mechanism that aims to encourage the model to focus on salient latent information of the composition or constituency tree relevant to the classification decision. We denoted the output of the intermediate nodes and leaf nodes in the layers of the tree-LSTM as $\underset{h1}{\rightarrow}, \underset{h2}{\rightarrow}, ..., \underset{h_{2n-1}}{\rightarrow}$ where 'n' is the sentence length. The information of the architecture is as follows:

$$\tilde{a_i} = exp(W_a \times h_i) \tag{5}$$

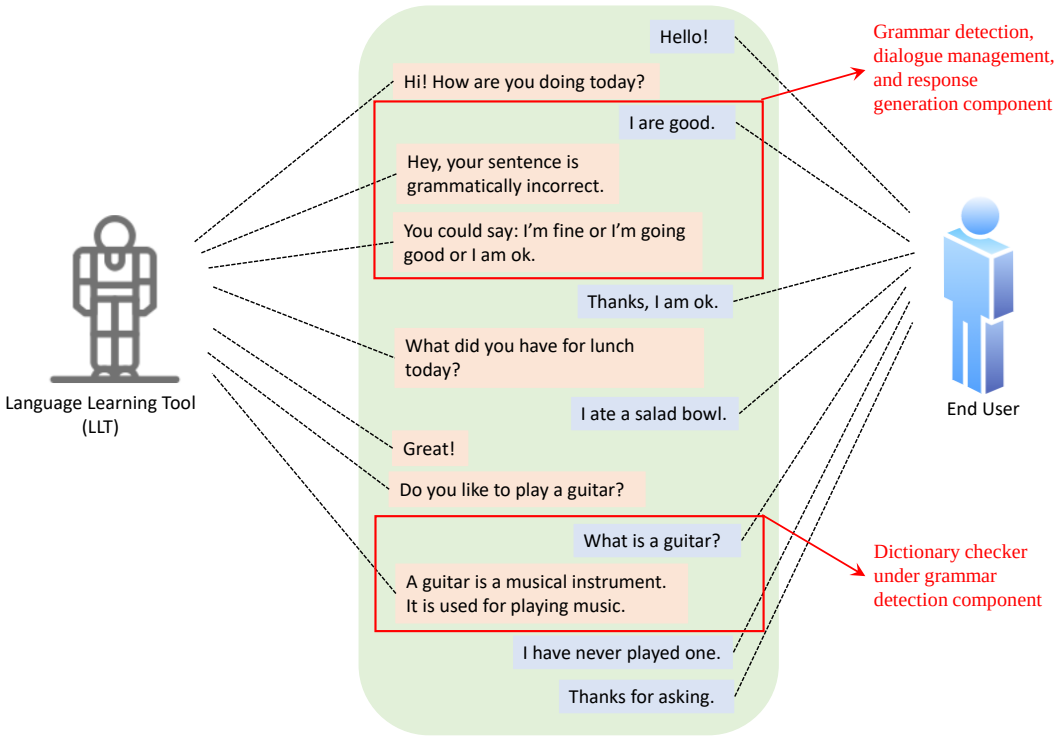


Fig. 1. A Sample Chat Window of End user with the LTT.

$$a_i = \frac{\tilde{a}}{\sum_{i=1}^{2n-1}(\tilde{a})} \tag{6}$$

$$\vec{H_c} = \sum_{i=1}^{2n-1}(a_i \vec{h_i}) \tag{7}$$

In our architecture, the context vector H_c is learned using attention-weighted contextual representation by passing Node vectors to the fully connected layer with Da hidden units in a hidden layer. For each such hidden state vector e_i, a scalar unit a_i is learnt with the weight matrix $W_a \epsilon R^{1Da}$. These scalar units are then normalized to 1, and the whole formulation of the context word vector is defined in equation 7. The process is visually summarized in the diagram of the architecture in Fig. 3.

III. Experimental Results

The section describes a complete experimental setup, hyperparameters, and dataset to analyze and evaluate the proposed model for paraphrasing and contextual similarity in conversation agents. The natural language inference (NLI) task is conducted to investigate the semantic relationship between two sentences. The proposed model is trained on the standard SNLI dataset that contains 570K pairs of English sentences.

The SNLI dataset* and Glove embedding* are available at resources. The model is trained on a large dataset containing sentences and learns from vector representation considering grammar and language constraints. The GPU Geforce RTX 2080 ti with 11GB memory and Cuda version 10.1 was used to train the model. Experimental results are evaluated in the table. The accuracy matrix in the table is employed to assess the model's efficiency and compare the results. The proposed model is reached based on pre-trained vectors' dimensions and several parameters used in training.

The presented table shows that Bi-LSTM with inner Attention is one of the stable models that use fewer parameters, 2.8m trained on 600D word vectors. The model achieved 84.2% testing accuracy, which is reasonably comparable with other models. The DiSAN uses 300D word vectors and achieves 85.6% accuracy with 2.4m parameters. This model is time efficient and consumes fewer parameters, but the proposed approach would only work for context-aware representations with temporal information encoded. Gumble tree-LSTM consists of tree-LSTM based on 600D word vectors and attains 86% inadequacy, but the deficiency is that it utilizes 10M parameters. BERT base model is proficient at reaching 90% testing accuracy. However, it stands among the list of large models as it utilizes 335m parameters.

TABLE I. Different Model Results

Model	Emb Dim	Params	Testing Accuracy
Bi-LSTM with inner Attention	600D	2.8m	84.20%
DiSAN	300D	2.4m	85.6%
Gumbel Tree-LSTM	600D	10m	86%
Tree-LSTM with attention	300D	6.5m	87%
BERT BASE		335m	90.7%

The above discussed models can be utilized for the same task, but they are not entirely appropriate in terms of task,

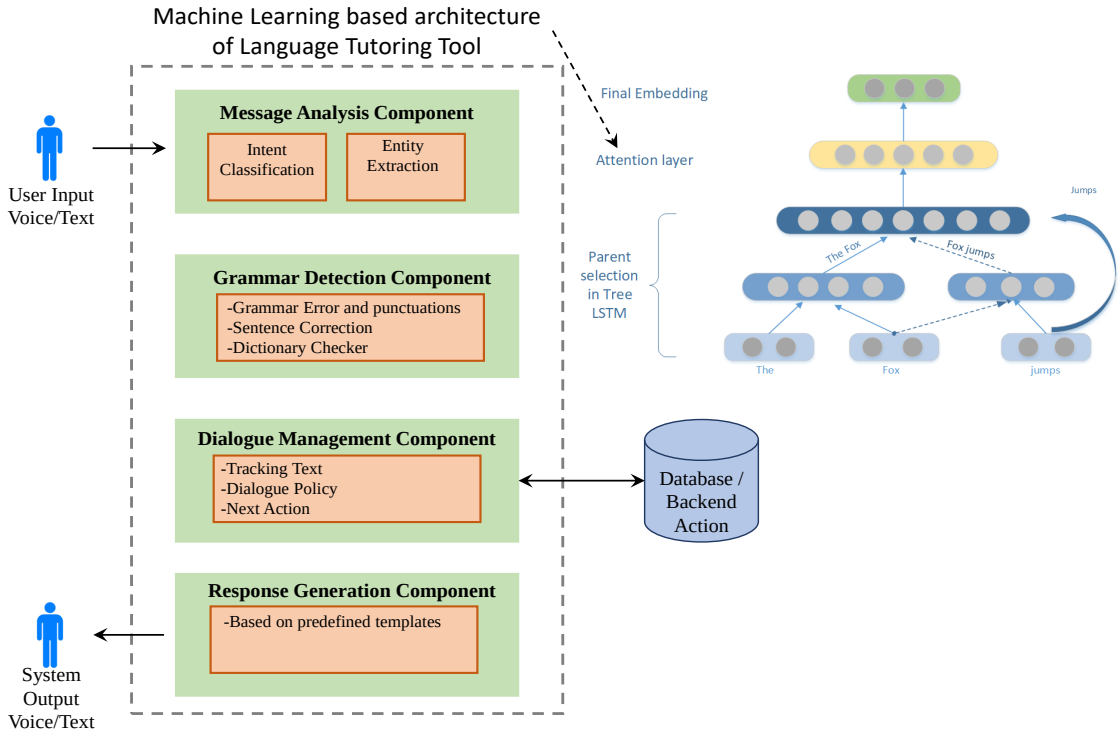


Fig. 2. Primary Components of the LTT based on the Proposed Architecture.

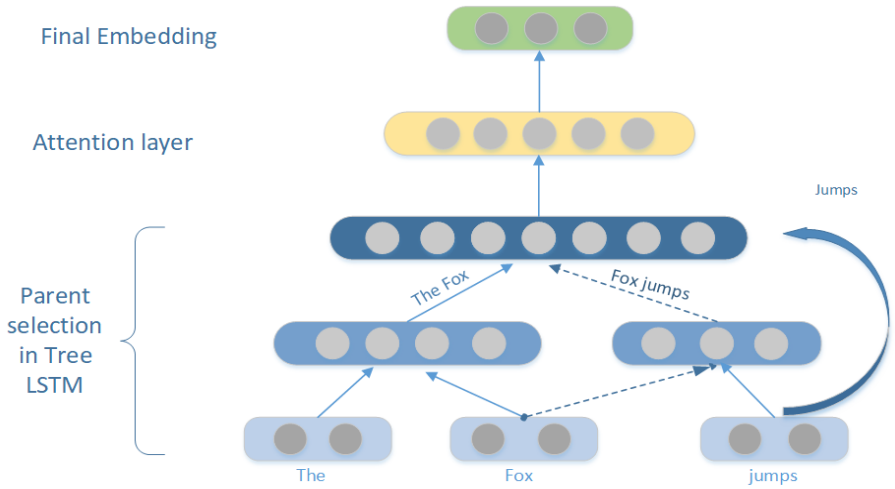


Fig. 3. Proposed Architecture for the LTT.

resources, and parameters. The proposed model is more robust, lightweight, and remembers the hidden states that help it in the question answering system. Therefore, our proposed model is efficient to be used in an LTT.

IV. Related Work

AI-based LTTs have the potential for human-like conversation. They work well for companies that have a lot of data. Despite the fact they need colossal time to train, they do save many in the long run. Based on previous information, they gather the information that learns new patterns from the input and improves the system after iterative training. Assistants understand patterns and behavior, and have a broader range of decision-making than any person, and over time they get to be very sophisticated. Many researchers investigated foreign language learning using a *chatbot*, where the *chatbot* mainly helps to learn english as a foriegn language [5], [6], [7], [8], [9], [10]. Our proposed model is generic and can be used for learning any language if the model is trained with respective languages' data set.

Our proposed platform comes under the task-oriented dialogue system in which end users will learn second language

through conversation. The proposed platform is initially trained on English language which will be extended further using datasets of other languages. However, the NLU and NLG are two must-have components that we needed to develop in the LTT. In the past decades, these modules are implemented individually and optimized using statistical models[11], [12]. Due to rapid and growing advancements in technology and algorithms, deep learning and reinforcement learning have reached the stage of producing human-like conversation agents. The trend to talk with a machine is not as new as more than 50 years ago; an agent is released that tries to maintain a conversation as mentioned in [13]. In recent years, messaging platforms offer developers to custom their chats and develop AI-based LTTs. The integration of AI modules to task oriented $chatbot$ can be complicated.

After the vector representation to evaluate the semantics of natural language text, sentence representation came into the ground to perform NLP tasks. The LSTM and its variants (RNN) sequentially process sentences to extract sentence meaning where the output depends on all previous hidden states. Recursive NN is the generalized version of RNN based on the structured input for sentence encoding as it extracts semantics through a hierarchical structure. The number of rvNNs were proposed in Socher et al. [14] which consider phrases instead of the complete sentence.

Matrix-Vector RvNN was reported as the best model that represents the word as vector and matrix at the same time. After that, Tai et al. [15] proposed two variants of standard LSTM (Child sum and N-ary tree-LSTM) that perform better for hierarchical and structural data. Additionally, hidden and cell states do not depend on the entire previous sequences in these variants as only the children's hidden and cell conditions contribute to the parent's states.

The Tree RNNs were extended by Zhou et al. [16] into attention-based tree RNN for semantic relatedness. In this model, semantic similarity is extracted by encoding attention in the tree structure of one sentence with vector representation of other sentences. Attention is used in machine translation by aligning the source or target sentence. It can focus on the most concerning sentences. The pioneers of the machine translation [4], [17] developed a model based on RNN variants that use attention for source words to generate target words.

A task-specific conversational agent is a kind of $chatbot$ that tries to imitate a human-like response bypassing the Turing test. Unlike pattern matching and simple algorithm-based $chatbot$, AI-based conversation agents contain complex algorithms and structures. Whether, we present a system in which an AI-based voice bot as a tutor provides services to learn a second language. A person or a student or a learner can speak a sentence multiple-way regardless of the vocabulary's dimension, as the system would efficiently detect the paraphrased sentences. At the beginning stage, when a person is learning the second language, it demands freedom to speak sentences and require hint to improve himself. This system considers the language constraints, grammar knowledge, and contextually aware algorithms trained on paraphrase detection.

V. Conclusion

This paper proposed an AI-based system that can perform casual conversation with a user in the second language for teaching and training purposes. The system based on the tree-LSTM with self-attention efficiently extracts contextual information from unstructured sentences and long conversations. The system is trained on a standard SNLI dataset that has been used widely for contextual understanding and paraphrasing detection systems. The experimental results expressed that the proposed model has achieved efficient results compared with state-of-the-art models.

References

[1] A.-M. Barraja-Rohan, "Using conversation analysis in the second language classroom to teach interactional competence," *Language Teaching Research*, vol. 15, no. 4, pp. 479–507, 2011.

[2] M. Adam, M. Wessel, and A. Benlian, "Ai-based chatbots in customer service and their effects on user compliance," *Electronic Markets*, vol. 31, no. 2, pp. 427–445, 2021.

[3] N. Haristiani, "Artificial intelligence (ai) chatbot as language learning medium: An inquiry," in *Journal of Physics: Conference Series*, vol. 1387, no. 1. IOP Publishing, 2019, p. 012020.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[5] J. Jia, "The study of the application of a web-based chatbot system on the teaching of foreign languages," in *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 2004, pp. 1201–1207.

[6] L. Fryer and R. Carpenter, "Bots as language learning tools," *Language Learning & Technology*, vol. 10, no. 3, pp. 8–14, 2006.

[7] J. Jia, "Csiec: A computer assisted english learning chatbot based on textual knowledge and reasoning," *Knowledge-Based Systems*, vol. 22, no. 4, pp. 249–255, 2009.

[8] Y. Goda, M. Yamada, H. Matsukawa, K. Hata, and S. Yasunami, "Conversation with a chatbot before an online efl group discussion and the effects on critical thinking," *The journal of information and systems in education*, vol. 13, no. 1, pp. 1–7, 2014.

[9] L. K. Fryer, M. Ainley, A. Thompson, A. Gibson, and Z. Sherlock, "Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners," *Computers in Human Behavior*, vol. 75, pp. 461–468, 2017.

[10] D. Coniam, "Evaluating the language resources of chatbots for their potential in english as a second language," *ReCALL*, vol. 20, no. 1, pp. 98–116, 2008.

[11] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[12] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[13] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[14] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 151–161.

[15] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[16] Y. Zhou, C. Liu, and Y. Pan, "Modelling sentence pairs with tree-structured attentive encoder," *arXiv preprint arXiv:1610.02806*, 2016.

[17] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

Secured 6-Digit OTP Generation using B-Exponential Chaotic Map

Rasika Naik, Udayprakash Singh
Electronics and Communication Engineering,
Sir Padampat Singhania University, Udaipur,
Rajasthan, India 313001

***Abstract*—Today, the traditional username and password systems are becoming less popular on the internet due to their vulnerabilities. These systems are prone to replay attacks and eavesdropping. During the Coronavirus pandemic, most of the important transactions take place online. Hence we require a more secure method like one-time password generation to avoid any online frauds. one-time password generation has multiple techniques. With one-time password generation it has become possible to overcome the drawbacks posed by the traditional username and password systems. The one-time password is a two-way authentication technique and hence secure one-time password generation is very important. The current method of one-time password generation is time-consuming and consumes a lot of memory on backend servers. The 4-digit one-time password system limits its uses to 9999 users and with advance deep learning approaches and faster computing it is possible to break through the existing one-time password generation method. Hence we need a system that is not vulnerable to predictive learning algorithms. We propose a 6-digit one-time password generation technique based on a B-exponential chaotic map. The proposed 24-bit (6-digit) long one-time password system offers 120 times higher security as compared traditional 4-digit systems, with a faster backend computing system that selects 24-bits out of 10^8 bits in 89 seconds at 1.09 Kilo-bits per milliseconds. The proposed method can be used for online transactions, online banking, and even automated teller machines.**

Keywords*—*One-time password generation; B-exponential chaotic map; 6-digit one-time password; online transactions; security

I. Introduction

Our manuscript introduces a 6-digit OTP generation system using the B-exponential chaotic map. The methods that are currently used to generate OTPs such as time-based OTP and hash-based OTP are prone to brute force attacks, forging attacks, etc. This leads to the research question, "Whether there was another method that could be used to generate a highly secure OTP system that was less prone to such threats?". Chaotic maps that are widely used for their highly random behavior were chosen to check whether they can be used to produce highly random OTPs. Also, instead of 4-digit OTPs, our objective is to make a system that generates 6-digit OTPs with a limited validity time makes it extremely difficult for a general computer to crack the OTP with brute force.

To prevent unauthorized access, access restriction methodologies are used [1]. The unauthorized attacker should be prohibited from making any changes to the system, at the same time the authorized user should not face any difficulties in accessing or updating the system [2]. Traditional identification control used badges and passwords to control the access authentication and provided a security safeguard [3]. Money transfers, mobile merchanting, account checks, and payment of different types of expenses (school fees, medical bills, and residential maintenance) are some of the examples of rapidly expanding mobile banking operations. Because an attacker can perceive some part of the key, conventional passwords are vulnerable to replay attacks and type scheme assaults [4]. Although ID/password methods are vulnerable to eavesdropping and replay attacks they are still better than only password systems because of the added unknown factor of the user name [5]. Most of the users tend to forget their passwords and hence they write them down or store them on a PC. This poses a greater threat to the traditional security system. To overcome the weaknesses of the traditional method, the One-Time Password (OTP) solution has been proposed [6]. OTP is an additive system that requires a new key to be entered by the user every time along with the user name and password. OTP was initially called a One Time Authorization Code (OTAC) [7]. It is a dynamic password that remains valid for a certain amount of time or till the successful login within a session. In earlier days, the OTPs were sent to a keyring fobe device or pager. The OTP generation algorithm is typically a pseudo-random algorithm that is difficult to be guessed by the attacker. The additive cryptographic hash functions make it very difficult to derive or guess the OTP. Some of the recent literature [8] has shown time-dependent OTPs, thus making it hard for the attacker to guess them.

An OTP is suitable for signing in to sessions or financial deals on any digital device like a computer, mobile. OTPs avoid several flaws associated with traditional (static) password-based authentication; some implementations also include two-factor authentication by ensuring that the one-time password requires access to both something a person has (such as a device with the OTP calculator built-in, or a smartcard or specific cell phone) and something (such as a PIN).

This manuscript proposes a mechanism for generating an OTP using a bit file generated by a pseudo-random sequence generator. This pseudo-random sequence generator uses the B-exponential chaotic map. Each session generates a 6-digit random OTP that will provide more security than the 4-digit OTP systems.

Fig. 1 shows the conceptual diagram for 6-digit (24-bit) OTP generation using B-exponential chaotic map. As per this concept, after inserting an ATM card into the machine, the user enters the correct user name and password. After validating the username and password, the bank server generates a 6-digit

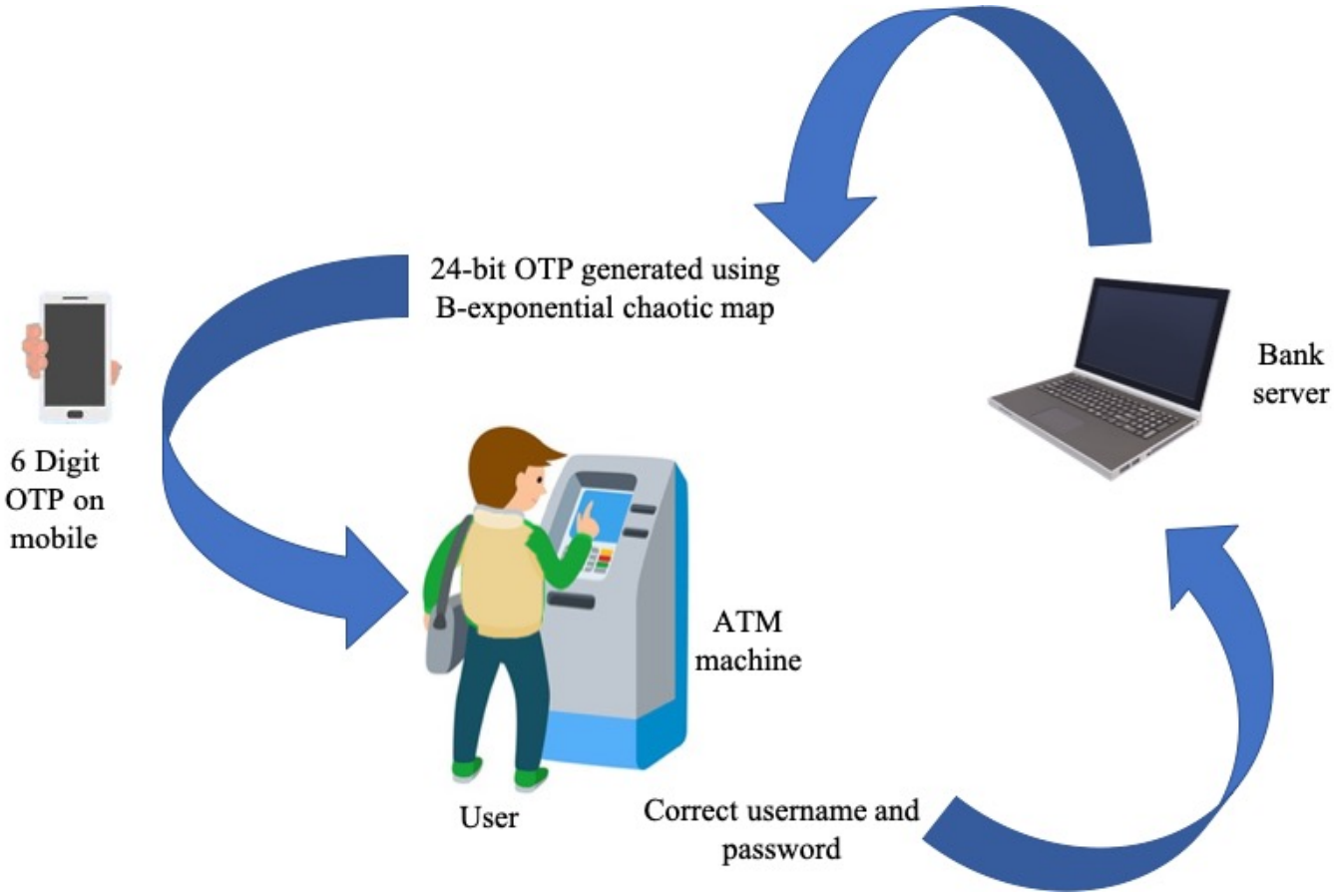


Fig. 1. Concept Diagram of 6-digit (24 bit) OTP Generation using B-exponential Chaotic Map. A user Enters the Correct User Name and Password after Inserting an ATM Card into the Machine. Upon Validation of Correct User Name and Password, Bank Server Generates 24-bit OTP using B-exponential Chaotic Map. This 6-digit OTP is sent on a Registered Mobile Number where the User will Enter the same in the ATM Machine.

(24-bit) OTP using a B-exponential chaotic map. This 6-digit OTP is sent to the user's registered mobile number, which the user enters into the ATM machine. This ensures a transfer to be secured on four levels. The first level is the user must have an ATM card. The second level is the user should know his or her personal username and password. Third, the mobile on which the OTP is sent should be in the network and in the same cell as the ATM machine. Finally, the B-exponential chaotic map is used to generate the time dependant OTP that makes the system more secured.

In this proposed methodology, we have shown how a 6-digit OTP can be generated using a novel B-exponential chaotic map method. Although this method was reported earlier in 2006, no one has ever used it generates a 6-digit OTP and validated using the NIST SP800-22.

We have proposed this system as with the increasing demand for online transactions, there is a need to develop a system that offered higher accuracy and security. The existing systems are more susceptible to brute force attacks. Also, as compared to 4-digit OTP systems 6-digit OTPs provide higher security. There are many ways that OTPs can be created but by using chaotic maps we provide a fast and simple method for OTP generation. Chaotic maps are simple to implement and also produce thousands of bits in a few milliseconds providing more security. They also have a special feature where they create a unique key that allows us to decode the OTP.

The manuscript has been organized into five distinct sections. The first section is the 'Introduction' section. It introduces the need to develop secure OTP systems and also highlights how chaotic maps can be used in them. The 'Literature review' section summarizes the most recent literature available on OTP generation systems. The 'Methodology' section explains in detail the steps we followed to implement our proposed OTP generation system. The 'Results and discussions' section mentions all the results that were obtained while implementing the system and also discusses these results and the future scope of the system. The final 'Conclusion' section briefly summarizes the entire manuscript.

II. Literature Review

Many efforts have been taken by different research groups in order to develop robust OTPs.

Most of the OTP generation techniques suggest time-based OTP [9], [8] and others have used Hash-based Message Authentication Code (HMAC) [10]. Recently, a captcha based OTP [11], real-time eye-tracking based OTP [12] where also proposed. Some of the researchers have also tried a combination of hashed and time-based OTP [9], [13]. RSA SecureID time-bas generates a safe OTP after specific seconds based on

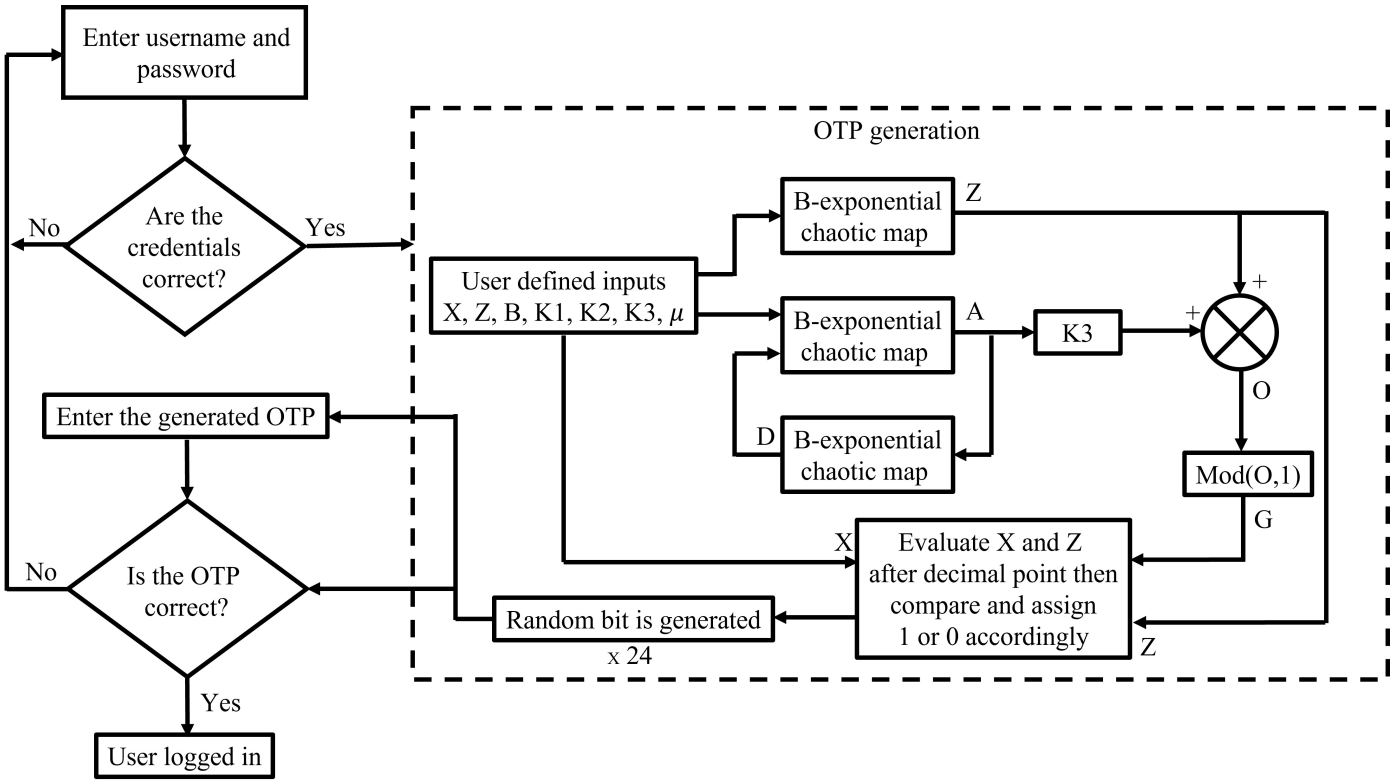


Fig. 2. Flow Diagram of OTP Generation using B-exponential Chaotic Map. The User Enters the User Name and Passwords and its Credentials are Checked. If the Credentials are False then the user is asked to Enter User Name and Password again. If the Credentials are Correct then the OTP Generation Algorithm is called on the Server-side. Once the OTP is Generated, User is asked to Enter that OTP and its Validation is Done. If the User Enters a Valid OTP then the OTP System allows him to Log in, else User Name if prompted for User Name and Password again. During the OTP Generation, User-defined Inputs are Passed through Three B-exponential Chaotic Maps. A Closed-loop Output of the B-exponential Chaotic Map Loop is Weighted and then Added to the Non-loop Output. This Summation is Modulo Operated and then Generated Bit is Evaluated against User Input and B-exponential Chaotic Map Output. The Final Evaluation is One Random Bit and this Process is Repeated 24 Times to Get the Final OTP.

arithmetic functions using the internal clock and stored seed. Each token carries its own initial value or conditions. However, these codes can be hacked because there is no reciprocal authentication. Time-based OTPs suffer from man-in-the-middle attacks. There were some attempts made to generate soft-token systems having unrivaled tamper resistance. Hardware tokens or keyring fobe devices or USB-based tokens with embedded chips were also implemented.

Lamport's technique [14] is the most common algorithm to generate hash chain-based OTPs (HOTP). These techniques are either OTP mechanism based on time to produce the OTP, like the algorithms suggested by El *et al.* [15] and Nugroho *et al.* [16], or hardware-based HOTP algorithm, like Lamport's [14] or S/Key [17].

Despite the widespread use of HOTPs in protocols like Secure Socket Layer (SSL), IPsec, and others. Algorithms used in HOTPs are typically vulnerable to attacks like collision, forging, and birthday attacks [18]. In comparison to time dependant OTPs, HOTP systems have additional flaws like more hashing steps and complicated computations, thus increasing resource utilization. Nontraditional bilinear map-based OTP developed by Lee *et al.* [19] was found to be vulnerable to insider attacks.

Chaotic maps can also be used for implementing other applications. There are many chaotic maps that are available. Akgul *et al.* [20] have proposed a random number generator using chaotic order systems.They test the algorithm with NIST 800-22, Federal Information Processing Standards (FIPS) 140-1, and ENT. Flores *et al.* [21] implemented a chaotic cryptosystem using a multi-precision algorithm. They had used four chaotic maps Rossler, TinkerBell, logistic, and Henon. Saber *et al.* [22] have developed a PRNG using a Lemniscate Chaotic Map (LCM). Deng *et al.* [23] have used chaotic maps to encrypt digital images using a scrambling algorithm.

Chaotic maps are not the only mechanisms for OTP generation. There are many recently reported works that generate OTP in a novel manner. Kumar *et al.* [24] have built an OTP generation system that uses the Vigenère cipher algorithm. They use this algorithm as it is not complex and still provides very high randomness which is required for OTP generation. Goel *et al.* [25] have proposed a system that uses cryptography and cloud computing to create a secured connection for Internet of Things (IoT) systems. They implemented the algorithm on MATLAB software and compared it with other methods. Kadum *et al.* [26] developed a novel OTP generation algorithm by generating an unsystematic key. They have created a random number and then ciphered plaintext. The random key created can make different ciphering texts.

The OTP generation is not only useful for secure banking and transactions but can also be used in encryption as a key.

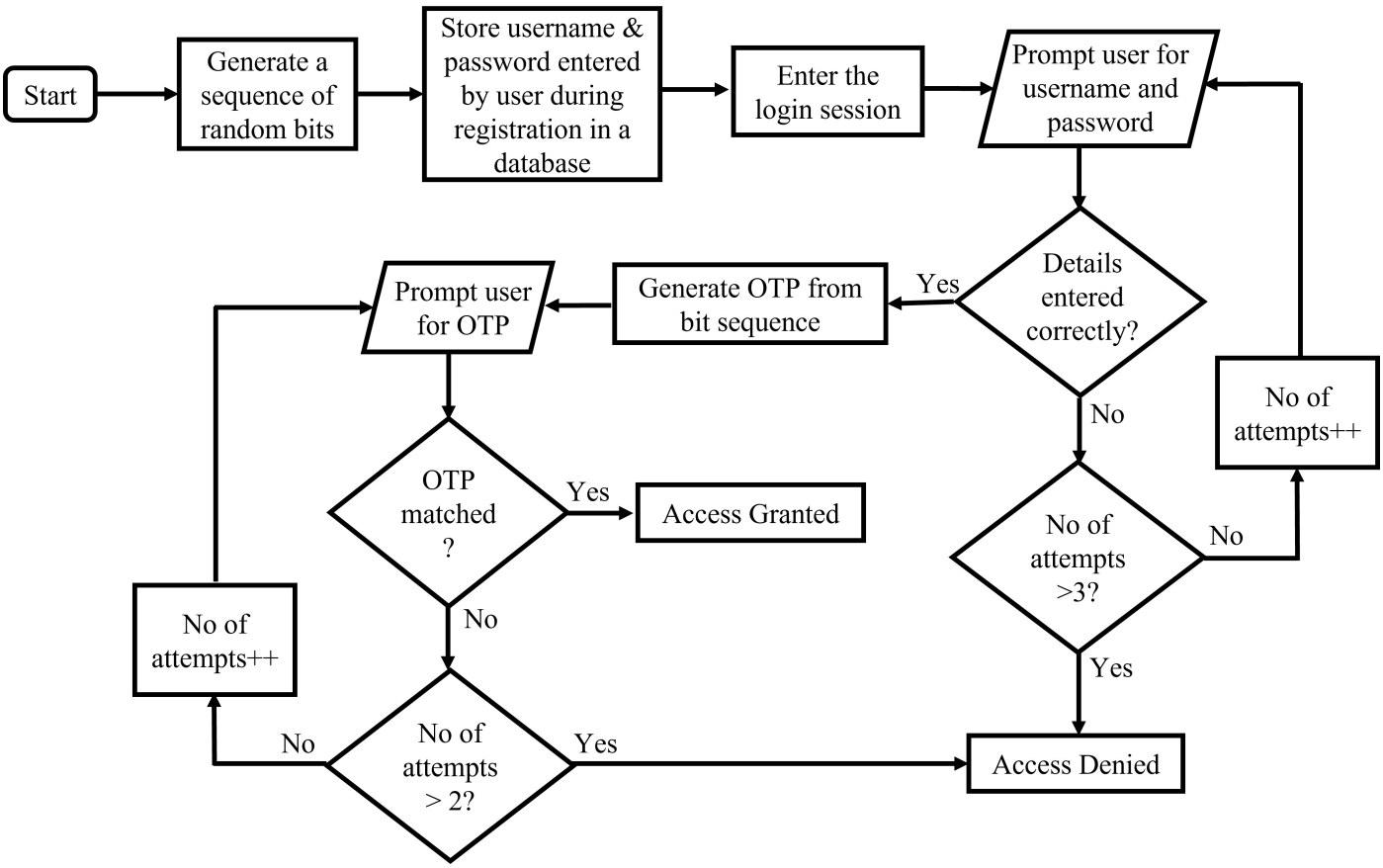


Fig. 3. Flowchart of the Proposed Methodology. A Random Bit Sequence is Generated Anticipatorily to Save Time during the Real-time Execution of the Proposed Method. A Database is Created to Store all the User Names and Passwords. All the Valid User Names and Passwords were Entered into the Database during Registration. A Login Session is Created for the User. The User is Prompted for the User Name and Password. If the Details are Entered Correctly then a 24-bit OTP is Generated else three more Attempts are given to the User before the Access is Denied. Once the OTP Generation is Completed, the User is Prompted for the OTP. If the OTP is Matched, the Access is Granted else two more Attempts are given to the User before the Access is Denied.

Thus a secure OTP generation algorithm will help in other applications as well. Recently, Shakir *et al.* [27] have developed an image encryption system. Using the Haar transform, One-Time Pad, and Playfair algorithms they have created an image encryption algorithm and then applied the Inverse Fourier Transform to get a ciphered image. The decryption is done by reversing the encryption method.

A. Challenges and Limitations of Existing Systems

A hash-based OTP suffers from an attacker who can position himself in front of hash and can access clear private information. A time-based OTP system relies on seed sequence and right counter. Such systems can be vulnerable if an attacker knows the right time and seed sequence. Although, it is very difficult to predict the time in advance as well as seed-sequence determination is still a challenge. physical token have concerns like being stolen, destroyed, running out of batteries, and clock drifting that takes hours to correct. There have been attacks reported using malware and sheet-based phishing.

From the survey done it was seen that the existing methods suffer from various attacks that can put private information at risk. Man-in-the-middle attacks and seed determination attacks are some of the major flaws of these OTP generation systems. Even tokens generated for OTP creation can be hacked into. To work on these drawbacks we have proposed a B-exponential chaotic map-based 6-digit OTP generation technique.

III. Methodology

Fig. 2 shows the flow diagram for OTP generation using a B-exponential chaotic map. According to the concept, when a user enters the username and password, the credentials are checked; if the credentials are incorrect, the user is prompted to enter the username and password again. If the credentials are correct, the server-side OTP generation algorithm is invoked. Once the OTP is generated, the user is prompted to enter it, and the OTP is validated. The OTP system allows the user to log in if he enters a valid OTP. If the OTP entered is incorrect then the user is prompted to enter the username and password again. User-defined inputs are passed through three B-exponential chaotic maps during the OTP generation process. The B-exponential chaotic map loop's closed-loop output is weighted and then added to the non-loop output. This summation is modulo-operated, and the generated bit is compared to the user input and the output of the B-exponential chaotic map. The final evaluation is one random bit, and the process is repeated 24 times to produce the final OTP.

Fig. 4. Different Output Cases. Case (a): In the First Attempt, the User Enters the Wrong Password after which the User is Prompted to Enter the Login Details again. In the Second Attempt, the User Enters the Correct Credentials and is then Prompted to Enter the Generated OTP. Once the Correct OTP is Entered by the User in the First Attempt, they are Granted Access. Case (b): The User Enters the Wrong Credentials in the First Attempt. The User is then Prompted to Enter the Login Details for the Second Time. The Details Entered by the User are Invalid in this Attempt as Well and the User is then Offered a Third Chance. Since the Details Entered in the Third Attempt are also Incorrect, the User is Denied Access. Case (c): Initially the User Enters the Correct Credentials. The User is then Prompted to Enter the Generated OTP. The OTP Entered by the User is Wrong in the First and Second Attempts. Hence, the User is Diverted Back to the Login Page. This Time the User Enters the Correct Login Credentials and OTP and is given Access. Case (d): The User Enters the Incorrect Credentials in the First Two Attempts and is given a Third Chance to Enter the Right Details. Once the User Enters the Right Login Details in the Third Attempt they are Prompted to Enter the OTP. As the OTP Entered by, the OTP Matches the Randomly Generated OTP, the User is Granted Access.

A. OTP Generation using B-exponential Chaotic Map

During the OTP generation, user-defined inputs (X, Z, B, K1, K2, K3, μ) are passed through three B-exponential chaotic maps. The B-exponential chaotic map closed-loop output (A) is weighted (multiplied by K3) and then added to the non-loop output (Z). This summation (O) is modulo-operated by 1, and the generated bit (G) is compared to the user input (X) and the output of the B-exponential chaotic map (Z). The final evaluation is one random bit, and the process is repeated 24 times to produce the final OTP.

B. OTP Generation Process Flow

Fig. 3 shows the flowchart of the proposed methodology. As per this concept, to save time during the proposed method's real-time execution, a random bit sequence is generated ahead of time. A database is created to store all of the usernames and passwords. During registration, all valid usernames and passwords were entered into this database. For the user, a login session is created. The user is prompted to enter his or her username and password. If the details are entered correctly, a 24-bit OTP is generated; otherwise, the user is given three more attempts before access is denied. When the OTP generation is finished, the user is prompted to enter the OTP. If the OTP is matched, access is granted; otherwise, the user is given two more attempts before access is denied.

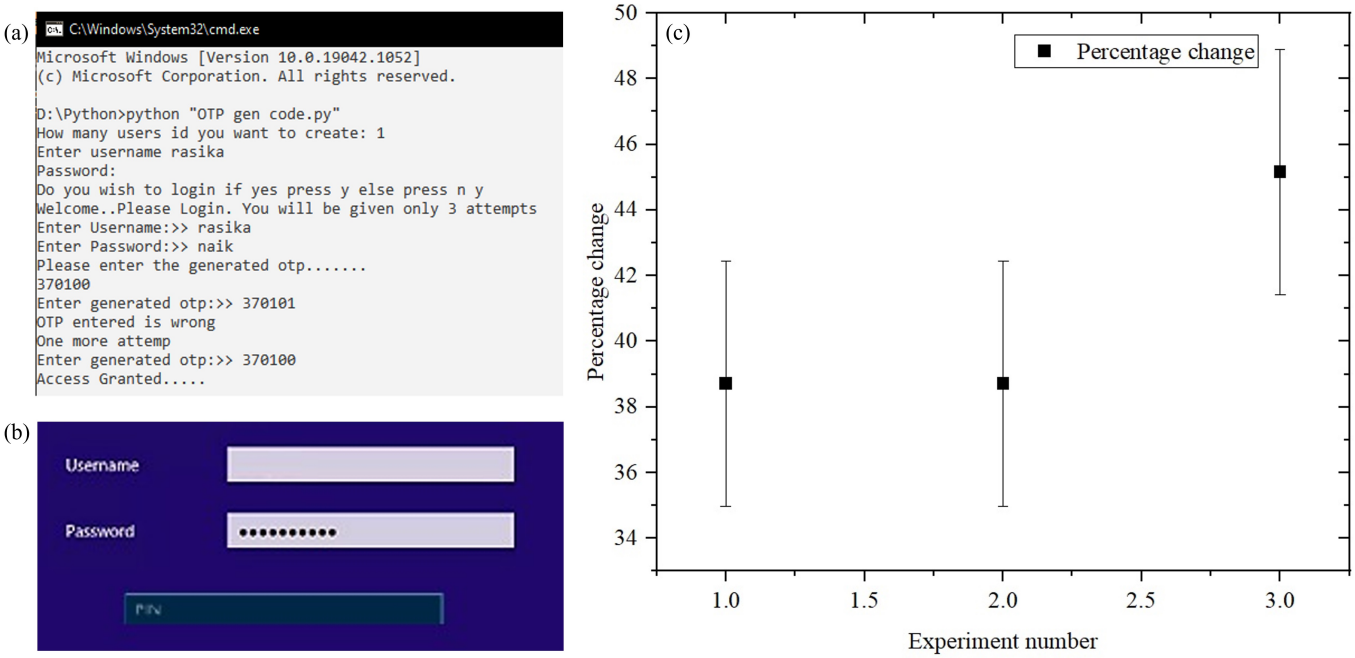


Fig. 5. (a) Console Output shows User has Entered the Right Password, one Wrong OTP and still gets Access Granted since the User has Entered Correct OTP on the Second Attempt. (b) Final Front-end UI showing User Name, Password and Pin. The Password is Masked and the Pin is Deactivated. (c) Different Experiments Carried with User Input Versus Percentage Change in Bits. In the First and Second Attempts, almost 40 % bits were Changed and in the Third Attempt, almost 46 % Bits Changed showing the Time Dependency on the OTP Generation for the same Set of Inputs.

TABLE I. VALIDATION OF BIT-STREAM GENERATED USING NIST SP800-22 TESTING PARAMETERS

Statistical case	P-Value	Proportion of passing	Result
ApproximateEntropy	0.437274	0.98	Success
BlockFrequency	0.153763	0.99	Success
CumulativeSums (Forward)	0.699313	0.97	Success
CumulativeSums (Backword)	0.595549	0.98	Success
(Fast Fourier Transform) FFT	0.304126	0.99	Success
Frequency	0.032923	0.98	Success
LinearComplexity	0.987896	0.98	Success
LongestRun	0.275709	0.98	Success
NonOverlappingTemplate	0.5125662466	0.98	Success
OverlappingTemplate	0.191687	0.98	Success
RandomExcursions	0.439636625	0.99	Success
RandomExcursionsVariant	0.645271	0.99	Success
Rank	0.678686	0.99	Success
Runs	0.401199	0.98	Success
Serial	0.1782745	0.975	Success
Universal	0.383827	0.97	Success

IV. RESULT AND DISCUSSION

Fig. 4 shows the various output scenarios obtained. Initially, the users are given the option to enter the number of login IDs they want to create. They are then asked to enter the login ID and password. This information is stored in a database. Fig. 4(a) shows the first case where the user enters the incorrect password on the first attempt. The user is now left with two more attempts to enter the right credentials. In the second attempt, the user enters the correct credentials. After this, he/she is prompted to enter the OTP generated by the proposed B-exponential chaotic map. The duration of the OTP generation process is 00:01:29 and the entire process takes 00:1:53. Once the OTP is generated by the system and received by the user, he/she is then prompted to enter the OTP. Since the OTP entered by the user matches the generated OTP, the user is granted access. Fig. 4(b) shows the second case where the user enters the incorrect credentials on the first try and is then left with two more attempts to enter the right credentials. On the second attempt, the user enters incorrect details again and is left with only one more try. Here on the third attempt, the user enters the wrong details again and is denied access. Fig. 4(c) shows the third case where the user initially enters the correct credentials. After that, the user is prompted to enter the OTP generated by the proposed B-exponential chaotic map. In the first attempt, the user's OTP is incorrect. This leaves the user with one more attempt to enter the OTP. As the OTP entered by the user does not match the generated OTP, he/she is redirected to the login page. The user is then prompted to enter the login credentials once again. Upon entering the correct credentials in the first attempt the user is prompted to enter the generated OTP. Once the OTP entered by the user matches the generated

TABLE II. COMPARISON BETWEEN THE EXISTING STUDIES REPORTED IN LITERATURE AND THE PROPOSED METHOD. THE ACCURACY IS OBTAINED FROM THE NIST TEST SUITE. OUR PROPOSED METHOD DEMONSTRATED BETTER CORRELATION COEFFICIENT AND ENTROPY THAN THE ONES REPORTED

Method	Accuracy (%)	Correlation coefficient	Entropy
De *et al.* [28]	99	-	-
Flores *et al.* [21]	98.39	-	-
Saber *et al.* [22]	-	0.0014	7.9980
Akgul *et al.* [20]	94	-	-
Tang *et al.* [29]	-	0.0857	7.990
Deng *et al.* [23]	-	0.0032	7.9931
Proposed method	98.45	0.00076	7.9999

OTP, he/she is granted access. This entire process takes a total time of 00:02:45 for completion. Here, the duration of the OTP generation process where the login details and OTP entered by the user are both entered correctly in the first attempt is 00:01:45. Fig. 4(d) shows the fourth case where the user enters the wrong credentials in the first attempt and is left with two more attempts to enter the correct details. As the details entered in the second try are incorrect the user is asked one last time to enter the right details. When the user enters the correct login information on the third try, they are prompted to enter the OTP. the entire process takes 00:1:53. Since the OTP entered by the user matches the OTP generated by the proposed B-exponential chaotic map the user is granted access.

Fig. 5 (a) shows the console output demonstrating the entries made by the user. The user entered the correct password, then entered one incorrect OTP but was still granted access because the correct OTP was entered on the second attempt. Fig. 5(b) shows the final front-end UI, displaying the username, password, and pin. The password has been masked, and the pin has been disabled. Fig. 5(c) shows the graph of user input versus percentage change in bits acquired from performing various experiments. In the first and second attempts, nearly 40 % of the bits were changed, and in the third attempt, nearly 46 % of the bits were changed, demonstrating the time dependency on OTP generation for the same set of inputs.

The proposed system was able to select 24-bits out of 10^8 bits in 89 seconds at 1.09 Kbits/ms. We have also checked the 4-digit password generation using the same B-exponential chaotic map and found that the probability of hacking 4-digital systems is 0.00012 using brute force attack. Whereas, the probability of hacking a 6-digit OTP generation system created using the B-exponential chaotic map was 0.000000991. This shows that the 6-digit OTP system provides 120 times higher security than a 4-digit OTP system. The maximum time-out period observed for the 6-digit OTP generation system was 15 minutes i.e. the OTP has to be reset within 15 minutes before it can become susceptible to brute force attack. The B-exponential chaotic map was also able to obtain a correlation coefficient of 0.00076 and an entropy of 7.9999. This proves that the chaotic map algorithm we have proposed is highly random and can produce secure output.

A. NIST Test Suite Result and Comparison

Table I shows the accuracy obtained for the bit-streams generated. The NIST SP800-22 testing parameters were used for validation purposes. The results obtained were, the system achieved a 98 % accuracy with a p-value of 0.43 and 196 successful attempts out of 200 for the approximate entropy statistical case. For the block frequency test, the system achieved a 99 % accuracy with 198 successful tests out of 200 with a p-value of 0.15. For the forward cumulative sums test, the system showed 97 % accuracy with 194 successful tests out of 200 with a p-value of 0.69. For the backward cumulative sums test, the system achieved a 98 % accuracy with 196 successful tests out of 200 with a p-value of 0.59. In the Fast Fourier Transform (FFT) test, the system successfully passed 196 attempts out of 200 with a 98 % accuracy and a p-value of 0.3. The system achieved 196 successful attempts out of 200 with a 98 % accuracy and p-value of 0.03 for the frequency test. For the linear complexity test, the system achieved a 98 % accuracy with 196 successful tests out of 200 with a p-value of 0.98. In the longest run test, the system successfully passed 196 attempts out of 200 with a 98 % accuracy and a p-value of 0.27. The system achieved 196 successful attempts out of 200 with a 98 % accuracy and p-value of 0.51 for the non-overlapping template test and also achieved 98 % accuracy for the overlapping template test with a p-value of 0.19. In the random excursions test, the system successfully passed 198 attempts out of 200 with a 99 % accuracy with a p-value of 0.43 and in the random excursions variant test it achieved a 99 % accuracy with a p-value of 0.64. The system also achieved a 99 % accuracy in rank parameter test with 198 successful attempts out of 200 and a p-value of 0.67. In the runs test the system was able to achieve a 98 % accuracy with 196 successful attempts out of 200 and a p-value of 0.4. The serial test showed an accuracy of 97.5 % with 195 successful attempts out of 200 with a p-value of 0.17 and the universal test achieved a p-value of 0.38 with a 97 % accuracy with 194 successful attempts out of 200. All the tests were passed with a successful result.

Table II provides a comparison between some of the works reported in the literature that use chaotic maps for different applications and our proposed method. The accuracy of 98.45 % which is obtained as a result of the NIST test suite indicates that our proposed system is one of the best. The correlation coefficient and entropy indicate the randomness of the proposed algorithm. It can be seen that with a 0.00076 correlation coefficient and 7.9999 entropy our system has outperformed the exiting works reported.

B. Data Science and OTP Generation

Chaotic maps are used in applications in which creating confusion in the initial data to encrypt it is required. They are dynamic systems. But due to their chaotic behavior, chaotic maps on their own are generators of huge amounts of random data. These maps lead to the generation of millions of unique

bits before repeating themselves due to their large periodicity. The huge amounts of data and random trajectory that the maps produce make it difficult for hackers to decipher them. Another application of chaotic maps relating to data science is that these maps are capable of encrypting the data available. A vast amount of data is available on the internet today and it is crucial to encrypt it to provide security and avoid hacking attacks. Chaotic maps are also valuable in encrypting images and audio files.

Due to the frequent usage of online websites and online transactions it has become important to secure data. OTP generation using chaotic maps plays a big role in this application. We have seen that OTP generated using chaotic maps is highly unpredictable and secure. This will allow users on the internet to continue secure browsing, protect their data and be safe from hackers.

V. Conclusion

Due to the vulnerabilities in traditional username and password systems, there was a need to develop a more secure system, especially due to the increased use of online transactions during the COVID-19 pandemic. We have developed a 6-digit OTP generation method over traditional 4-digit OTP using a novel B-exponential chaotic map. As the current methods of generating OTPs are time-consuming and uses a large amount of memory on backend servers, we developed a fast and less memory-consuming system. The 4-digit OTP system is limited to 9999 users, but with our 6-digit we could expand this database to 100 times more users which will be suitable for the upcoming 5G technology. The proposed 24-bit (6-digit) long OTP system was able to achieve 120 times more security than traditional 4-digit systems with a faster backend computing system that selected 24-bits out of 10^{8} bits in 89 seconds at 1.09 Kbits/ms. The proposed method is applicable to any online transactions or banking applications.

Declaration

Funding Information

No funding was involved in the present work.

Conflicts of Interest

Authors R. Naik and U. Singh declare that there has been no conflict of interest.

Code Availability

The codes will be made available upon reasonable request to the authors.

Authors' Contribution

Conceptualization was done by R. Naik (RN) and U. Singh (US). All the literature reading and data gathering were performed by RN. All the experiments and coding was performed by RN. The formal analysis was performed by RN and US. Manuscript writing- original draft preparation was done by RN. Review and editing was done by US. Visualization work was carried out by RN and US.

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.

Consent to Participate

This article does not contain any studies with animals or humans performed by any of the authors. Informed consent was not required as there were no human participants. All the necessary permissions were obtained from Institute Ethical committee and concerned authorities.

Consent for Publication

Authors have taken all the necessary consents for publication from participants wherever required.

References

[1] T. Breaux and A. Antón, "Analyzing regulatory rules for privacy and security requirements," *IEEE transactions on software engineering*, vol. 34, no. 1, pp. 5–20, 2008.

[2] G. Ho, D. Leung, P. Mishra, A. Hosseini, D. Song, and D. Wagner, "Smart locks: Lessons for securing commodity internet of things devices," in *Proceedings of the 11th ACM on Asia conference on computer and communications security*, 2016, pp. 461–472.

[3] A. Jøsang and S. Pope, "User centric identity management," in *AusCERT Asia Pacific information technology security conference*. Citeseer, 2005, p. 77.

[4] A. G. Chefranov, "One-time password authentication with infinite hash chains," in *Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics*. Springer, 2008, pp. 283–286.

[5] M. H. Barkadehi, M. Nilashi, O. Ibrahim, A. Z. Fardi, and S. Samad, "Authentication systems: A literature review and classification," *Telematics and Informatics*, vol. 35, no. 5, pp. 1491–1511, 2018.

[6] S.-D. Park, J.-C. Na, Y.-H. Kim, and D.-K. Kim, "Efficient otp (one time password) generation using aes-based mac," *Journal of Korea Multimedia Society*, vol. 11, no. 6, pp. 845–851, 2008.

[7] W. N. W. Muhamad, N. A. M. Razali, K. K. Ishak, N. A. Hasbullah, N. M. Zainudin, S. Ramli, M. Wook, Z. Ishak, and N. J. A. MSaad, "Enhance multi-factor authentication model for intelligence community access to critical surveillance data," in *International Visual Informatics Conference*. Springer, 2019, pp. 560–569.

[8] I. S. Shaik and M. Manoj, "Time based dynamic password (tbdp) system using variable insertion technique," *International Journal of Computer Applications*, vol. 113, no. 8, 2015.

[9] S. S. Gosavi and G. K. Shyam, "A novel approach of otp generation using time-based otp and randomization techniques," in *Data Science and Security*. Springer, 2021, pp. 159–167.

[10] S. ShanmugaPriya, A. Valarmathi, and D. Yuvaraj, "The personal authentication service and security enhancement for optimal strong password," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 14, p. e5009, 2019.

[11] R. K. Devi, M. Muthukannan, S. H. Babu, A. Sivadasan, and S. Abinivesh, "Novel authentication mechanisms for hash code, captcha and otp in cyber security domain," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2021, pp. 62–68.

[12] H. K. SM, G. Pradyumna, B. Aishwarya, and C. Gayathri, "Development of personal identification number authorization algorithm using real-time eye tracking & dynamic keypad generation," in *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE, 2021, pp. 1–6.

[13] J. Keller and S. Wendzel, "Reversible and plausibly deniable covert channels in one-time passwords based on hash chains," *Applied Sciences*, vol. 11, no. 2, p. 731, 2021.

[14] L. Lamport, "Password authentication with insecure communication," *Communications of the ACM*, vol. 24, no. 11, pp. 770–772, 1981.

[15] S. A. El-Booz, G. Attiya, and N. El-Fishawy, "A secure cloud storage system combining time-based one-time password and automatic blocker protocol," *EURASIP Journal on Information Security*, vol. 2016, no. 1, pp. 1–13, 2016.

[16] E. P. Nugroho, R. R. J. Putra, and I. M. Ramadhan, "Sms authentication code generated by advance encryption standard (aes) 256 bits modification algorithm and one time password (otp) to activate new applicant account," in *2016 2nd International Conference on Science in Information Technology (ICSITech)*. IEEE, 2016, pp. 175–180.

[17] N. M. Haller, "The s/key (tm) one-time password system," in *Symposium on Network and Distributed System Security*, 1994, pp. 151–157.

[18] V. Shivraj, M. Rajan, M. Singh, and P. Balamuralidhar, "One time password authentication scheme based on elliptic curves for internet of things (iot)," in *2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW)*. IEEE, 2015, pp. 1–6.

[19] Y. Lee and H. Kim, "Insider attack-resistant otp (one-time password) based on bilinear maps," *International Journal of Computer and Communication Engineering*, vol. 2, no. 3, p. 304, 2013.

[20] A. AKGÜL, C. ARSLAN, and B. ARICIOĞLU, "Design of an interface for random number generators based on integer and fractional order chaotic systems," *Chaos Theory and Applications*, vol. 1, no. 1, pp. 1–18, 2019.

[21] A. Flores-Vergara, E. García-Guerrero, E. Inzunza-González, O. López-Bonilla, E. Rodríguez-Orozco, J. Cárdenas-Valdez, and E. Tlelo-Cuautle, "Implementing a chaotic cryptosystem in a 64-bit embedded system by using multiple-precision arithmetic," *Nonlinear Dynamics*, vol. 96, no. 1, pp. 497–516, 2019.

[22] M. Saber and M. M. Eid, "Low power pseudo-random number generator based on lemniscate chaotic map." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 11, no. 1, 2021.

[23] Z. Deng and S. Zhong, "A digital image encryption algorithm based on chaotic mapping," *Journal of Algorithms & Computational Technology*, vol. 13, p. 1748302619853470, 2019.

[24] M. Kumar and S. Tripathi, "A new method for otp generation," in *Healthcare and Knowledge Management for Society 5.0*. CRC Press, pp. 213–228.

[25] A. Goel, D. K. Sharma, and K. D. Gupta, "Leobat: Lightweight encryption and otp based authentication technique for securing iot networks," *Expert Systems*, p. e12788, 2021.

[26] S. A. Kadum and S. M. K. Jawad, "New programmatic otp algorithm," in *Journal of Physics: Conference Series*, vol. 1818, no. 1. IOP Publishing, 2021, p. 012097.

[27] H. R. Shakir and S. A. Yassir, "Image encryption-compression method based on playfair, otp and dwt for secure image transmission," in *International Conference on Advances in Cyber Security*. Springer, 2021, pp. 95–113.

[28] L. G. de la Fraga, E. Torres-Pérez, E. Tlelo-Cuautle, and C. Mancillas-López, "Hardware implementation of pseudo-random number generators based on chaotic maps," *Nonlinear Dynamics*, vol. 90, no. 3, pp. 1661–1670, 2017.

[29] Z. Tang, Y. Yang, S. Xu, C. Yu, and X. Zhang, "Image encryption with double spiral scans and chaotic maps," *Security and Communication Networks*, vol. 2019, 2019.

Mobile Application Aimed at Older Adults to Increase Cognitive Capacity

Ricardo Leon-Ayala, Gerald Gómez-Cortez, Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—The research work focuses on people with dementia of the Alzheimer's type, since, among the types of dementia, this is the most common worldwide. In Peru, more than 200 thousand adults over 60 years of age suffer from this disease and many others who still do not know it or are in its initial stage. Therefore, it was decided to create a prototype of a mobile application with memory games, riddles, reminders and different types of physical activities to perform during the day. The scrum methodology was implemented to promote good practices for team and collaborative work, in terms of us phases from inception to launch of the product which is the mobile application. In addition, balsamiq was used as a prototype design tool. And so the objective of creating the prototype for its development was achieved. The goal of creating the prototype for the application was achieved. Positive results were obtained in terms of user and customer satisfaction. This will allow the benefit of adults for the improvement of cognitive ability, being able to perform their daily activities in the best way and socializing with family and friends.**

***Keywords*—Alzheimer's; balsamiq; mobile; prototype; scrum**

I. Introduction

Dementia prevents us from performing the activities of daily living properly, as it is a syndrome that affects memory, understanding and behavior. Worldwide, dementia has become a major public health problem. Risk factors for dementia vary from country to country. It is estimated that the number of new cases of dementia each year worldwide is close to 7.7 million, with one new case every four seconds. Each year, there are 3.6 million new cases (46%) in Asia, 2.3 million (31%) in Europe, 1.2 million (16%) in the Americas and 500,000 (7%) in Africa [1].

Epidemiologists around the world have conducted research on the prevalence of dementia, with predicted estimates coming from a variety of sources, including epidemiological surveys, hospital records and electronic medical record databases. A meta-analysis included 157 epidemiological studies conducted worldwide between 1980 and 2009. The report showed that the prevalence of dementia among people over 60 years of age was 5.8% to 8.0%, showing an exponential increase with age, doubling every five years. A different global meta-analysis study reported similar results, with a prevalence rate of 4.8% for people over 60 and a period prevalence rate of 6.9%.

Although dementia affects the world, it represents a unique danger for Latin American and Caribbean countries compared to the lower, stable and declining prevalence rates in Europe and the United States, the prevalence of dementia in people over 65 years of age in this region of the continent is high, and increasing, between 7.1% and 11.5%. In addition, there are several related risk factors, including significant genetic heterogeneity and social determinants of health [2]. In Peru, according to the Ministry of Health (Minsa) indicated that in 2019 more than 200 thousand people over 60 years of age suffer from dementia. This report was presented by the psychiatrist Manuel Escalante.

Due to the COVID-19 pandemic, the government of Peru, through the former president Martín Alberto Vizcarra Cornejo, has been taking preventive measures since the first wave of this disease, such as the use of masks, face shields, social distancing, etc. One of these provisional measures was social isolation, a strict but necessary posture. In this situation, there was a statistically significant increase in the population in the levels of agitation, depression, appetite, eating disorders, nocturnal behavioral disturbances and aberrant motor activities [1]. It is not yet known what cognitive damage these measures may have caused in the population, according to experts the results of these new routines may be chaotic.

Among the types of dementia, Alzheimer's disease (AD) is the most common, accounting for 60-70% of dementia cases. In the initial stage of AD, cognitive impairment progresses progressively, showing orientation difficulties, language alterations and cognitive dysfunctions. During the course of the disease many psychological symptoms occur in behavior, affecting activities of daily living [3].

According to the author [4], there is an inverse relationship between sedentary time and cognitive ability, although studies on this topic are inconsistent. However, not all sedentary challenges are equal. Passive sedentary behaviors, such as watching television, appear to be harmful; whereas more cognitively stimulating activities, such as reading, using a computer or solving puzzles, are shown to be associated with improved cognitive ability and recognition.

By means of the analysis carried out research work is to develop a mobile application model using basic principles of mental health experts. Enabling the improvement of FE and delaying future Alzheimer's problems in older adults.

The article is made up of sections. In Section II for literature review, in Section III, the methodology was applied to, In Section IV a case study was conducted, in Section V result and discussion, and finally in Section VI the conclusion and future work.

II. Literature Review

The research topic of this article is cognitive ability in older adults, likewise, of the various applications created on the basis of cognitive enhancement, based on this I design an application to enhance the cognitive capacity of older adults through interactive games.

The author [5], reviewed the studies of the last decade in the field of cognitive training using communication and information technologies, also recorded those cognitive improvement strategies and thus evaluated the effectiveness of these programs, that is why I come to the following conclusion, entertainment applications, video games among others could be used in intervention studies for cognitive improvement of healthy or cognitively impaired individuals.

On the other hand, the author [6], It is recognized that there is limited understanding of how older adults use smartphones and how their use changes compared to younger users, in addition, mentions that based on data collected from the telephones of 84 healthy adults during the last three months, the most common characteristic is that they open fewer applications, take longer to perform daily tasks and send fewer text messages, then a cognitive analysis was performed on each respondent and it turns out that 79% had cognitive impairment, This is why their study suggests that researchers and developers should take cognitive impairment into account when developing any type of project.

On the other hand, according to the author [7], tells us that taking into account the illnesses of people due to their age, an application consisting of four modules was developed: news, reasoning games, reasoning questions and a calendar. The app was put to the test and the results show that there is an increased interest in memory games and entertaining brain games.

According to the author [8], modern cell phones have generated new interactive scenarios that require complex interfaces. The major operating system developers of these devices provide APIs for developers to implement their own apps, including different solutions for graphical interfaces, sensor control and voice interaction. While these resources are useful, there is no clear strategy to combine the multimodal interface with the possibilities offered by the device to identify and adapt to user needs, which is particularly important in areas such as environmentally assisted living.

In addition, the author [9], argues that digital apps for seniors should be evaluated to support senior independence and home care. Apps provide increasing opportunities for older adults and their family caregivers to educate, participate and share health information through digital platforms. Few applications have recorded evidence of availability for the elderly and their caregivers. By conducting a survey, the author concluded that technology use among older adults and caregivers was high. Usability and engagement of the mobile application was average. Additional training is needed for older adults and their caregivers, including that on specific behaviors for digital maintenance.

In conclusion, the good contributions of the authors mentioned for the development of this topic were analyzed, also, certain shortcomings were observed in the interactivity with the application, so a new contribution will be added by applying what was analyzed and incorporating new knowledge.

III. Methodology

In order to carry out the research, the agile Scrum methodology was used, is considered agile because it uses incremental and iterative process approaches. On the other hand, has proven to be more useful than the traditional waterfall model because it improves the productivity of the processes and helps to reduce the time consumed for their realization. In a traditional waterfall model, planning is performed prior to testing, in addition, the process is managed in phases and once it has been completed, it is not possible to return to the previous phase. On the other hand, Scrum at any stage changes can be made to improve the outcome. In Fig. 1, the Scrum flow is shown [10].

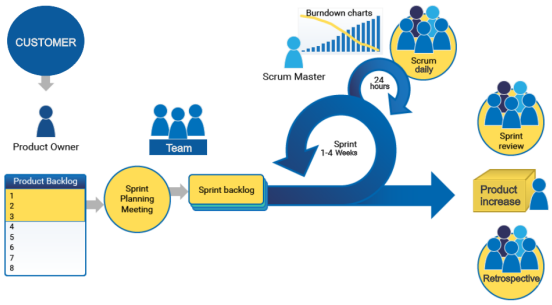


Fig. 1. Scrum Flow.

A. Scrum

The use of scrum involves integrating roles between work teams, an advantage of using scrum is that it allows teams to adapt quickly to manage and plan their work, since, each step of scrum allows you to plan, design, develop and test the code, all these activities are divided between roles [10], Fig. 2 shows the roles proposed by Scrum [11].

Scrum has three roles: Product Owner, Scrum Master and the Team:

1) The Product Owner: For this role, the manager is responsible for the definition and prioritization of the client's needs, in addition to communicating all the information of the client's requirements to the entire team and in this way maximize the value of the product.

2) The Scrum Master: Who assumes the role is the "Servant Leader" of the team, also known as moderator, is responsible for solving problems of the team, also motivates the same to develop its activities.

3) The Team: For this role, the person in charge is the entire team in charge of the development of the project, usually made up of 3 or 9 members. It is responsible for making the deliverables established in the Product Backlog, on the other hand, a characteristic of the team is its proactivity, ability to multitask and have the necessary knowledge to help in the activities of its colleagues [12].

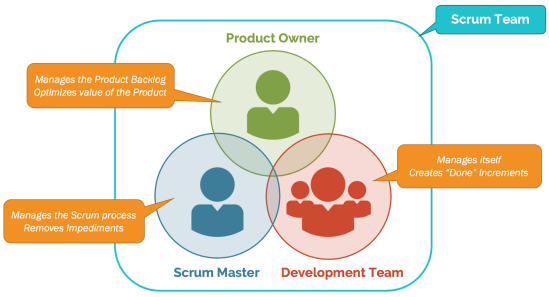


Fig. 2. Roles in Scrum.

B. Scrum Practices

Scrum has practices also called Scrum formalities: Daily Meeting, Sprint Review, Sprint Planning and Sprint Retrospective Meetings:

1) Daily Meeting: The meeting is held every day, the team discusses the problems and the process of the project.

2) Sprint Review: A meeting is held in which the team presents the results at the end of each Sprint to the owner.

3) Sprint Planning Meetings: A meeting is held to define the activities to be carried out during a Sprint.

4) Sprint Retrospective: The entire Scrum team meets with the objective of evaluating the team's performance during the Sprint and the practices to be carried out to improve team productivity [12].

C. Structure of Scrum Development

1) Requirements Stage: It is established to obtain the project planning requirements.

2) Definition of User Stories: It is the description of the functionality of the system, for which the requirements are analyzed in collaboration between the client and the team, the user stories will be improved throughout the life of the project.

3) User Stories Prioritization: The priority of each of the user stories is established and the order in which they will be developed is determined, taking into account which one is the most fundamental for the project and thus following a hierarchy [13].

4) Analogous Estimation: When the user stories are obtained, they go through an examination where the use of a tool or estimation method is required, this consists of a study of each of the user stories concerning the time that the scrum team believes it can develop it, this phase involves both time and the resources and expenses that may be required.

5) Definition of Sprint Speed: The speed and time of development of each Sprint is performed through an analysis that varies according to the experience of the Scrum team [14].

6) Creation of the Product Backlog: User stories, priorities and required tasks are included in a Product Backlog. t includes the features or short term requirements and long term functionalities that have been defined jointly by the development team and Product Owner, moreover, the points or level of effort for each task are established. During the Sprint process, assigned tasks are movable depending on the Product Owner's requirements [15].

7) Sprint Presentation: The Sprint is presented according to the requirements and the determined times, in this phase, the scrum team performs tests on what is presented and it is either accepted or rejected, this is repeated according to the number of sprints a project contains [16].

8) Feedback: At the end of the project or each sprint, the scrum team holds a feedback meeting where they analyze everything developed in order to know their successes and failures in all aspects, para poder mejorar en un próximo proyecto o Sprint.

D. Development Tools

1) Kotlin Programming Language: Kotlin is a programming language with 100% interoperability with Java that combines functionality and object oriented features so that developers can use it. Write a new file in an existing Java project or write a new file Use Kotlin applications from scratch. In addition, the official IDE is used for android development [17].

2) Android Studio: The Android operating system is a linux-based open source software, this means that it can be used by anyone for free and free of charge. It was developed primarily for portable devices, including tablets and smartphones. Its architecture allows compatibility with the Java language. Ever since the first version of Android 1.0 was released in 2008, have been developing it up to version 11. Android studio is being powered by IntelliJ IDEA, what is integrated development Environment (IDE) [18]. IntelliJ IDEA's contribution makes it even simpler to create Android applications.

3) FireBase: Firebase is a web application platform. Helping developers create quality applications. The data are stored as objects using the JSON format. Firebase does not depend on the most common queries, how to add, insert, delete or update. This tool is basically a backend for your system, used to store data used to store the data and use it according to your needs [19].

4) Balsamiq: It is software used to create diagrams of a web page, which is widely used by application designers [20]. Balsamiq encourages us to focus on the structure and content of a website, avoiding position and detail errors, at the same time it allows us to create prototypes quickly, recreating the experience of a computer board [21].

IV. Case Study

A. Start-up Stage

1) Requirements Identification: Table II lists the functional requirements of the system., based on the survey and results shown in Table I.

2) User Stories: Table II shows the list of User Stories created based on the functional requirements shown in Table II.

TABLE I. REQUIREMENTS LIST

N°	Requirements
R-1	The application must have a section that allows me to register or log in with an alternative account, Gmail, Facebook, etc.
R-2	The application should perform a test where I evaluate the phase in which I am in encounter, early stage, middle stage or final stage.
R-3	The application allows me to select the preferred activities depending on the phase I'm in.
R-4	The application must show my progress during its use, in order to be able to evaluate my improvement during the use of the application.
R-5	The application should show me recommendations, advice or information based on my results or the phase I am in.
R-6	The application must have activities that include logic and mathematics.
R-7	The application should have a game customization stage to improve my user experience.
R-8	The application must show my results at the end of each section or activity.

TABLE II. USER STORIES

N°	User Stories
H1	As a user, I want the application to allow me to login or register to have access to all functionalities..
H2	As a user, I want the application to have a test or entry game to know what stage I am in.
H3	As a user, I would like to have different sections where I can evaluate my knowledge according to the stage I am in to perform the activities your preferences.
H4	I as a user, I want the application to show my progress as I use the application to see my progress over time.
H5	As a user, I want the application to suggest activities, information and recommendations based on my results to improve my current state.
H6	I, as a user, want the application to include mental reasoning games to improve my cognitive memory.
H7	I as a user, I want the application to allow me to customize the games to enhance my user experience.
H8	As a user, I want the application to show me the results at the end of each section to observe my performance in each activity.

B. Planning Phase

1) Analog Estimation: At this stage, the complexity estimate is made with respect to the development of each story, for this purpose, the rating range is from 1 to 13, 1 is the lowest level of complexity and 13 is the highest. he estimation is based on H1, since this is where the "Login" is developed, then by way of example, H5 is located in column 5, which indicates that H5 is 5 times more complex to develop than H1, the same logic applies to the other stories. Thanks to the analog estimation, an idea of the functional development of each story for a future project was presented. To arrive at the following result, the work team held a discussion on the estimation of each story, were also classified according to personal criteria. As shown in Table III.

TABLE III. ANALOG ESTIMATION

	1	2	3	5	8	13
H1	1-Login					
H2						13-Login
H3						13-Login
H4		2-Login				
H5				5-Login		
H6					8-Login	
H7				5-Login		
H8			3-Login			

2) User Stories Prioritization: In order to arrive at the following result, the work team performed an analysis of each story to determine those that are most relevant to the operation of the mobile application, that is, is sorted according to priority, as shown in Table IV in the first column all the user stories are ordered, then the second column contains the information obtained in Table III, and each estimate is added according to the corresponding history and finally, in the last column, the prioritization is carried out, for this it is considered that user stories are important to develop first for the correct functioning of the application.

TABLE IV. USER STORIES PRIORITIZATION

User Stories	
H-U	Priority
H1	1
H2	2
H3	3
H4	4
H5	5
H6	6
H7	7
H8	8

3) Definition of Sprint Speed and History Points: In order to carry out the following chart it was decided together with the development team to divide the Sprint into three deliverables, also details which user stories belong to each Sprint, for example, H1 and H2 belong to Sprint 1. As shown in Fig. 3, Sprint 1 has 14 points, this result was obtained thanks to the information in Table IV, the effort points obtained by the stories were added up there, H1 only obtained 1 point and H2 13 points, Adding both values together gives the result. The sprint speed refers to the estimated time for the completion of a sprint, speed is determined by the development team, the equipment is new, is why at the beginning the speed is high, however, Sprint 3 the speed slows down as the team is already more consolidated.

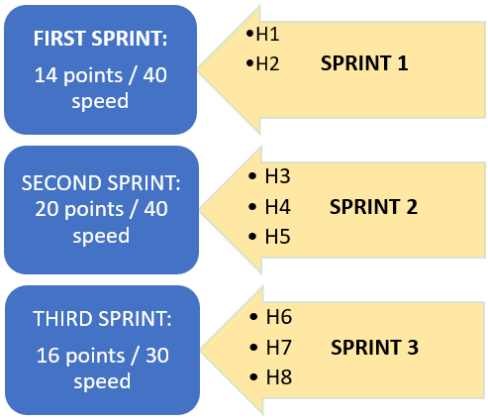


Fig. 3. History and Speed Points.

The Sprint Backlog table shows the three Sprint with their respective functionality to be developed, also, the estimated time for its development is defined. Estimated project development time is 3 months and 2 weeks. also, deliverables are completed in a shorter time. All this is shown in Table V:

4) Creation of the Product Backlog: For the creation of the complete backlog, ordered according to their history and priority, likewise, in the "Sprint" column, each story is sorted

TABLE V. SPRINT BACKLOG

Interface	Duration
Cognitive memory enhancement application	3 months and 2 weeks
Sprint 1: Login Interface	1 week
Sprint1: Registration Interface	1 week
Sprint 1: Home Interface	2 week
Sprint 1: Input evaluation	2 weeks
Sprint 2: Creation of activities	2 weeks
sprint 2: Creation of interactive games	2 weeks
Sprint 3: Activity customization functionalities	2 weeks
Sprint 3: Results reports	2 weeks

according to which Sprint it belongs to, is why it has been numbered from 1 to 4. Finally, the Estimate is listed from 1 to 13, and this information is obtained from the analog estimating Table. The Product Backlog is shown in Table VI.

TABLE VI. PRODUCT BACKLOG

User Stories	Priority	Sprint	Estimation
H1: As a user, I want the application to allow me to login or register to have access to all functionalities.	1	1	1
H2: As a user, I want the application to have a test or entry game to know what stage I am in.	2	1	5
H3: As a user, I would like to have different sections where I can evaluate my knowledge according to the stage I am in to perform the activities I prefer.	3	2	13
H4: As a user, I want the application to show my progress depending on how long I have been using the application to see how much progress I have made.	4	2	2
H5: As a user, I want the application to suggest activities, information and recommendations based on my results to improve my current state.	5	2	5
H6: As a user, I want the application to include mental reasoning games to improve my cognitive memory.	6	3	8
H7: As a user, I want the application to allow me to customize the games to improve my user experience.	7	3	5
H8: As a user, I want the application to show me the results at the end of each section to observe my performance in each activity.	8	3	3

V. RESULT AND DISCUSSION

A. Presentation of Prototypes by user Stories

This stage presents the design of the prototypes created based on the requirements and user stories, in order to graphically represent the functionalities of the application.

First Sprint: As shown in Fig. 3, the first sprint has a total of 14 story points, the estimated time is one month and two weeks.

H1: I as a user, I want the application to allow me to login or register to have access to all functionalities. As shown in Fig. 4, the option to log in via Facebook and Gmail has been implemented, also, in Fig. 5, in case you do not have any of the accounts you have the option to register to start using the application.

H2: I as a user, I want the application to have a test or entry game to know what stage I am in. After logging in, the application automatically directs you to a 10-activity entry test to assess your current status (see Fig. 6), the app presents you with several interactive games, among them the following stand out: 4 pictures one word, crossword puzzles, mathematical problems, sound recognition (sounds of animals, things, instruments, etc.). in addition, each test has a time limit for answering. These are shown in Fig. 4 and Fig. 5.

Fig. 4. Login Screen.

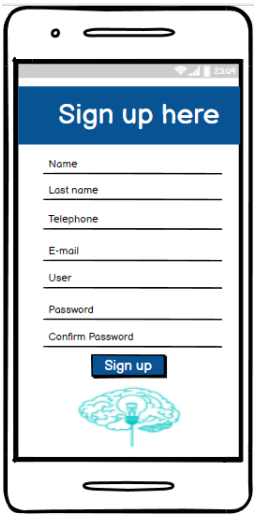


Fig. 5. Registration Screen.

At the end of the test, the results are shown, in the initial part it shows you where you are based on your results, Fig. 8 shows that at the end of the 10 activities, at the top shows you the result, in this case the status is ”Initial”, which indicates that the tests were performed in the estimated optimal time.

Second sprint:

H3: As a user, I would like to have different sections where I can evaluate my knowledge according to the phase I am in to perform the activities I prefer. As shown in Fig. 9, 6 interactive games have been developed for this story, which are intended to stimulate the mental agility of our users.

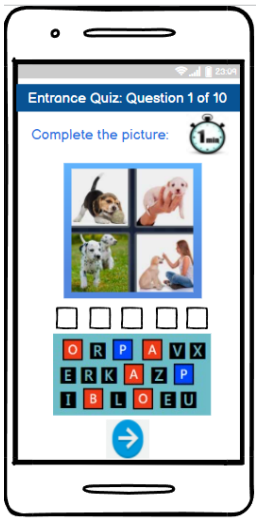


Fig. 6. Question 1 of the Test.

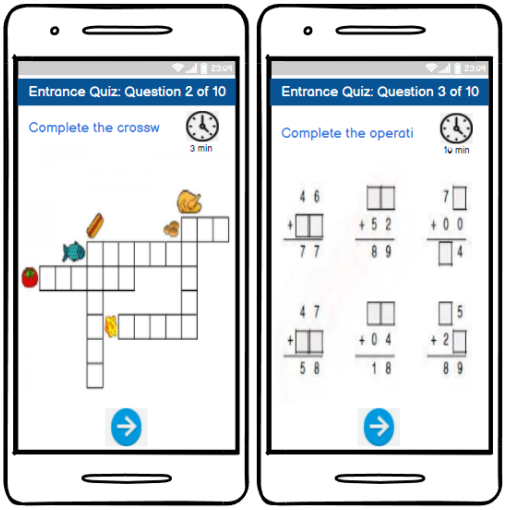


Fig. 7. Test Questions 2 and 3.

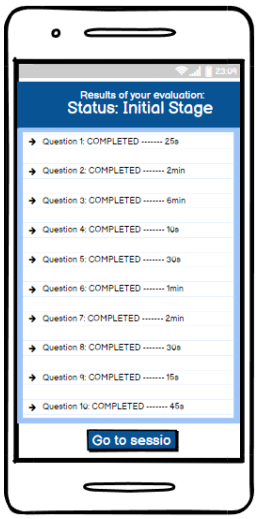


Fig. 8. Entry Test Results.

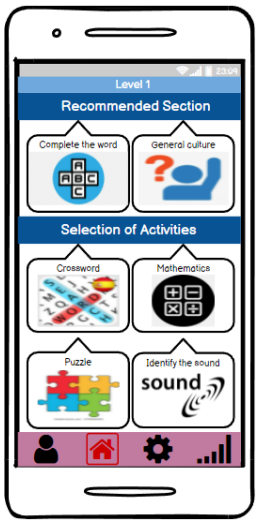


Fig. 9. Games Section.

H4: As a user, I want the application to show my progress depending on how long I have been using the application to see how much I have progressed, this functionality was implemented through general and game levels. When you select a game the app directs you to an interface where you find all the levels, this stage was designed for the user to progress over time, to pass to the next level you must necessarily finish the previous level. As you progress through the games your overall level will increase. this section is shown in Fig. 10.

H5: I, as a user, want the application to suggest activities, information and recommendations based on my results to improve my current state. As shown in Fig. 9, this can be seen in the initial part, where it is shown that "Complete the word" and "General culture" are the recommended games, this recommendation depends on your initial test result and your current condition.

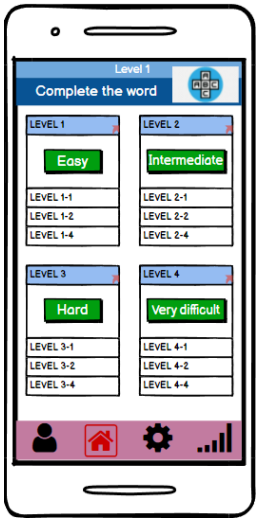


Fig. 10. Game Levels Section.

Third Sprint:

H6: I as a user, I want the application to include mental

reasoning games to improve my cognitive memory. As shown in Fig. 7, n this section, mathematical reasoning problems are developed, since the team considers that it is essential to maintain an active mind.

H7: I as a user, I want the application to allow me to customize the games to improve my user experience. As shown in Fig. 11, this section allows the user to customize all games, you can change the background of each image, this function was implemented in order to make the user more familiar with the app by including images of their choice, also, has the option to include custom music for each game, since music is an active stimulant of the mind and influences the mood generally in a positive way.

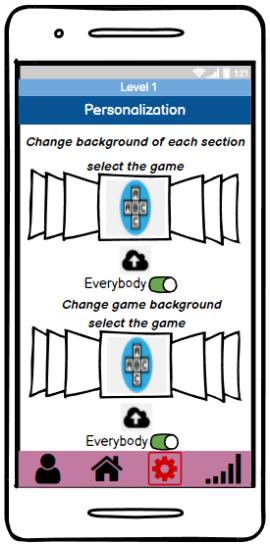


Fig. 11. Customize Games.

H8: I, as a user, want the application to show me the results at the end of each section to observe my performance in each activity. As shown in Fig. 12, this section shows the user's progress in each game, depending on the time of use, this information varies depending on the time of use. The information presented is updated daily and shows the results of the last month, order to allow users to graphically observe their progress and thus motivate them to continue making progress.

All the prototypes shown in the research were validated by testing the functionality.

B. Comparison between Methodologies

The following are a series of methodologies, which have different standards, covering all aspects of requirements, development, design and quality. Table VII shows a comparison of the requirements considered for each method studied. In this table, agile methodologies are compared, as well as with traditional methodology; This allows us to know the importance of the use of scrum in research.

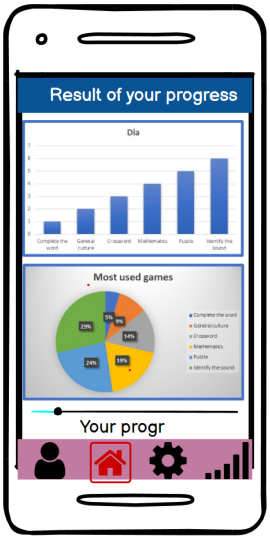


Fig. 12. Result of the use of the Games.

1) Assessment between Methodologies: Table VIII shows the evaluation of the methodologies studied and the scores obtained. The criteria are scored 1, 3, and 5, with 1 = low, 3 = medium, and 5 = high. The criteria have allowed a score to be made according to the methodology under study, where the scrum methodology obtained the highest score. In addition, Table IX shows the 11 questions asked in the survey; this allowed for an exhaustive analysis.

C. Analysis of the Survey Carried out for user Stories

For the creation of the requirements, a survey of 35 people between 35 and 60 years of age was conducted, using the GoogleForms tool, the survey was conducted directly from their cell phones. The following questions were created by the development team in order to know the user's preferences. Table I shows the questions asked to the 35 respondents.

In Fig. 13, the results and user presence can be seen, thanks to this information, the functional requirements of the application were created, the following are the most relevant results for the system.

Of the 35 people surveyed, 31 prefer that the application allows you to log in through an alternative account: Gmail or Facebook, this is why it is necessary to implement this functionality in the system.

Of the 35 people surveyed, 32 agree with the creation of an initial test, which should be developed in order to determine the user's level: "Beginner", "Intermediate" or "Moderate".

D. Analysis of Each Sprint using Burndown Chart

At this stage, each Sprint was analyzed with the help of a graph, the X-axis shows the time range, which has a duration of 1 month and 2 weeks (42 days). The Y-axis shows the score for Sprint 1, the analogous estimate seen in Table VI indicates that it has a total of 14 history points. The Blue line is the expected result and the Orange line is the actual result.

TABLE VII. COMPARISON OF METHODOLOGIES

	METHODOLOGIES		
CRITERIA	**SCRUM**	**RUP**	**XP**
Description	A model that is strengthened by the active participation of all project members.	Characterized by the iterative and incremental model.	Extreme programming model, allowing for the incorporation of new functionalities.
Type of Review	A daily review is needed, analyzing the following: 1. Work done the day before. 2. Work to be performed. 3. Tasks with impairments or that can be performed.	In this methodology, the phases are developed one after the other, thus perfecting the objectives. If a phase is not completed, the next phase is not continued.	It must be integrated at least once a day, and tests must be performed on the entire process.
Objectives	Suitable for projects in complex environments: Innovation and competitiveness. Get fast results. Changing requirements.	Object-oriented that establishes the basis, templates and examples for all aspects and phases of software development.	Defined by giving priority to work with direct results. Customer satisfaction. Group work. Acting on variables: Cost, time, quality and scope.
Stages	Planning. Mounting. Development. Release.	Home. Prepared by. Construction. Transition.	Define roles Estimate the effort Choosing what to build Program Repeat.
Characteristics of the model	Increased active customer collaboration.	Focuses on using use cases with an incremental model.	Emphasis on Programming.

TABLE VIII. EVALUATION BY PROCESSING CRITERIA

	METHODOLOGIES		
CRITERIA	**SCRUM**	**RUP**	**XP**
KNOWLEDGE	5	5	3
TIME	5	1	5
ADAPTABLE	5	5	5
APPLICABLE	5	3	5
TOTALS	20	14	18

This leads to the conclusion that the team at the time was not consistent in its work, since the actual line is not lines, has inconsistent variations, also, the Sprint was successfully completed, this is shown in Fig. 14.

For Sprint 2, has a total of 20 story points, with a duration of 1 month (28 days). As shown in Fig. 15, the development of the project was carried out with a higher consistency, the first two weeks progress was slow and from the second week onwards the activity intensified until the Sprint was finally completed.

Finally, in Sprint 3, has a total of 16 history points, in addition, has a duration of 1 month (28 days). At this stage of the project, the team is already consolidated, since, as can be seen the expected time and the actual time does not vary much compared to the first Sprint, which indicates a greater commitment of the team to the project, the activities were developed gradually until the end. As can be seen in Fig. 16.

TABLE IX. QUESTIONS ASKED OF RESPONDENTS

No	Survey Questions
1	Have you played games that are based on strategy to win?
2	Would you like the application to allow you to enter through your Gmail account?
3	Would you like the application to take an entrance test to know if you are in the ”initial”, ”medium” or ”moderate” stage?
4	Would you like the application to include basic math operations?
5	Have you solved solving puzzles and riddles? health?
6	Would you like an application that allows you to play your favorite songs? for the diagnosis?
7	Would you like to have several games within the application?
8	Would you like to see your progress after a certain time of use of the application
9	Would you like the application to show you the results of your progress in a graph?
10	Would you like to include levels of difficulty for each game?
11	Do you think an application can improve your cognitive memory?

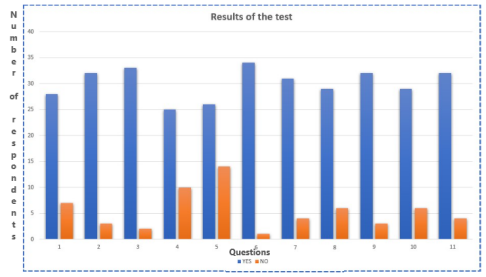


Fig. 13. Results of the Survey.

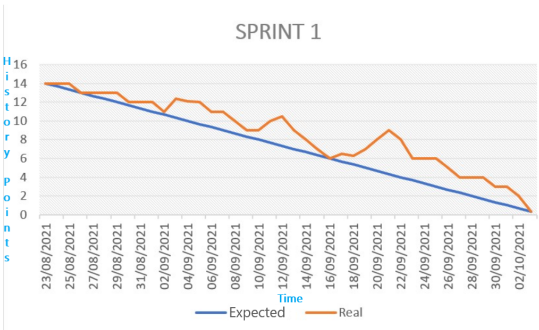


Fig. 14. Result of the First Sprint.

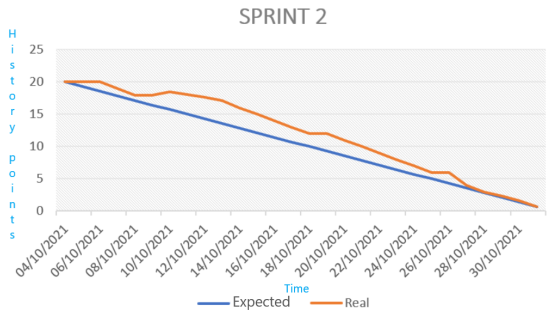


Fig. 15. Result of the Second Sprint.

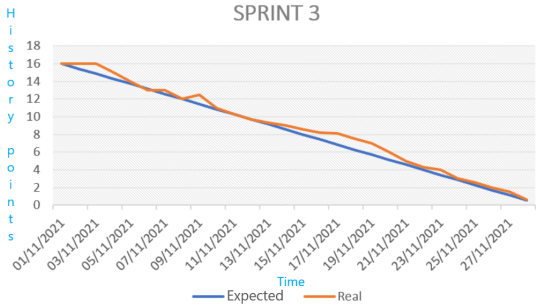


Fig. 16. Third Sprint Result.

E. Analysis of Electronic Glasses to Improve Vision

In the research, the design of glasses is presented whose objective is to improve the vision of the elderly, applying various technologies [22], including the use of augmented reality, real-time motion sensors and the use of GPS, among others, as seen in Fig. 17 and Fig. 18. The development of this research [23], will be of great help to those older adults who consume applications from their Smartphone, including our application, since it would be a pity if due to visual problems it is not possible to use our application, that is why the development of electronic glasses is a complement to take into account when thinking about the scope and comfort of the user.

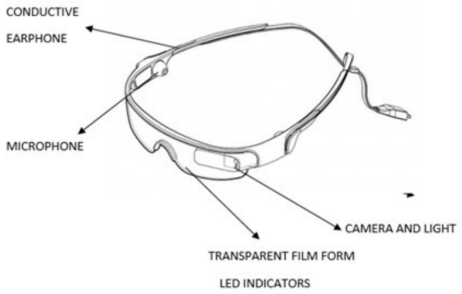


Fig. 17. Design of the Electronic Glasses [23].

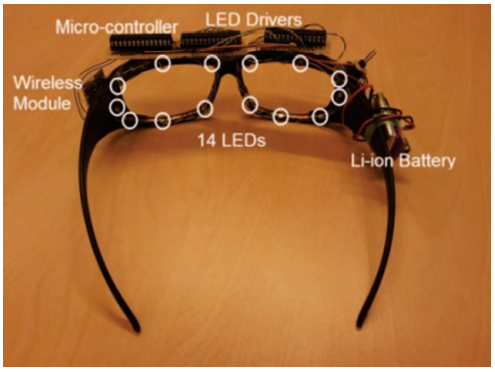


Fig. 18. LED Indicators Implanted in the Frame of the Glasses [23].

VI. Conclusion and Future Work

The research work concludes that a mobile application prototype has been developed to improve cognitive ability in older adults. The research contribution will allow the improvement of the health of these citizens. Using the Scrum methodology and applying the Balsamiq design tool, the functionalities required by the respondents were successfully represented.

In terms of methodology, the Scrum framework was of vital importance, since it allows a high-value sequential development, thus improving the quality of the application. In addition, the variety of tables and graphs provided by Scrum are the key to presenting information clearly and accurately.

In the future, it is expected that this research will be complemented with the development of a mobile application implementing the use of Artificial Intelligence to improve the initial diagnosis, also, it is suggested to apply all the steps presented above and in this way make the application a reality, as it will be very useful for people in the early stages of Alzheimer's disease, thus preventing the disease from progressing. It is expected that this mobile application will soon contribute to the health sector and be of great help to people.

Acknowledgment

Acknowledge the University of Sciences and Humanities, and its research institute, for their support in research. We acknowledge the support of the University of Sciences and Humanities, and its research institute, for their suggestions and recommendations.

References

[1] R. Brito-Aguilar, "Dementia around the World and the Latin America and Mexican Scenarios," vol. 71, no. 1, 2019.
[2] M. A. Parra, S. Baez, L. Sedeño, C. G. Campo, H. Santamaría-García, I. Aprahamian, and P. H. Bertolucci, "Dementia in latin america: Paving the way toward a regional action plan," *Alzheimer's and Dementia*, vol. 17, pp. 295–313, 2 2021.
[3] L. K. Huang, S. P. Chao, and C. J. Hu, "Clinical trials of new drugs for Alzheimer disease," vol. 27, no. 1, 2020.
[4] A. E. Altinöz, F. Kösger, and A. Eşsızoğlu, "Relationship between selective attention, cognitive flexibility, response inhibition and theory of mind functions in OCD," *Anadolu Psikiyatri Dergisi*, vol. 20, no. 1, 2019.
[5] M. A. Pappas and A. S. Drigas, "Computerized training for neuroplasticity and cognitive improvement," *International Journal of Engineering Pedagogy*, vol. 9, no. 4, pp. 50–62, 2019.
[6] M. L. Gordon, L. Garys, C. Guestrin, J. P. Bigham, A. Trister, and K. Patel, "App usage predicts cognitive ability in older adults," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12, 2019.
[7] B. A. Leonardo, A. Prieto Taborda, S. Grajales Agudelo, and J. D. Pérez Rodríguez, "Aplicación móvil para mejorar la capacidad cognitiva en adultos mayores utilizando juegos mentales," *Teknos revista científica*, vol. 16, no. 2, p. 11, 2016.
[8] D. Griol and Z. Callejas, "Mobile Conversational Agents for Context-Aware Care Applications," *Cognitive Computation 2015 8:2*, vol. 8, no. 2, pp. 336–356, aug 2015. [Online]. Available: https://link.springer.com/article/10.1007/s12559-015-9352-x
[9] C. C. Quinn, S. Staub, E. Barr, and A. Gruber-Baldini, "Mobile Support for Older Adults and Their Caregivers: Dyad Usability Study." [Online]. Available: http://aging.jmir.org/2019/1/e12276/
[10] W. Mahmood, N. Usmani, M. Ali, and S. Farooqui, "Benefits to organizations after migrating to Scrum," *Proceedings of the 29th International*

Business Information Management Association Conference - Education Excellence and Innovation Management through Vision 2020: From Regional Development Sustainability to Global Economic Growth, no. May, pp. 3815–3828, 2017.

[11] 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

[12] M. Kumar and R. Dwivedi, "Applicability of Scrum Methods in Software Development Process," *SSRN Electronic Journal*, 2020.

[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] B. Singh, "Comparative Study and Analysis of Scrum and Lean Methodology," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, no. 3, pp. 3441–3448, 2018.

[15] M. Morandini, T. A. Coleti, E. Oliveira, and P. L. P. Corrêa, "Considerations about the efficiency and sufficiency of the utilization of the Scrum methodology: A survey for analyzing results for development teams," *Computer Science Review*, vol. 39, p. 100314, feb 2021.

[16] 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

[17] B. G. Mateus and M. Martinez, "On the adoption, usage and evolution of Kotlin features in Android development," 2020.

[18] S. Periyanayagi, A. Manikandan, M. Muthukrishnan, and M. Ramakrishnan, "BDoor App-Blood Donation Application using Android Studio," vol. 1917, no. 1, 2021.

[19] C. Khawas and P. Shah, "Application of Firebase in Android App Development-A Study," *International Journal of Computer Applications*, vol. 179, no. 46, 2018.

[20] F. N. Khasanah, S. Rofiah, and D. Setiyadi, "Metode User Centered Design Dalam Merancang Tampilan Antarmuka Ecommerce Penjualan Pupuk Berbasis Website Menggunakan Aplikasi Balsamiq Mockups," *JAST : Jurnal Aplikasi Sains dan Teknologi*, vol. 3, no. 2, 2019.

[21] 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

[22] M. Gamboa-Ramos, R. Gómez-Noa, O. Iparraguirre-Villanueva, M. Cabanillas-Carbonell, and J. L. H. Salazar, "Mobile application with augmented reality to improve learning in science and technology," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 10, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0121055

[23] L. O. Cunyarachi, A. S. Santisteban, and L. Andrade-Arenas, "Augmented Reality Electronic Glasses Prototype to Improve Vision in Older Adults," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 704–709, 2020. [Online]. Available: https://doi.org/10.14569/IJACSA.2020.0111185

Implementation of an Expert System for Automated Symptom Consultation in Perú

Gilson Vasquez Torres, Luis Lunarejo Aponte, Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—The human being has a fragile life, and is attacked by different diseases throughout his life, neglecting or ignoring some of them because it is considered minimal, can be fatal, but many do not want to attend a health center, so they seek your symptoms on the Internet and finding pages with false information, that is the problem that we will address in this investigation. The objective of the research is to implement an expert system, creating a web page that provides real information when a user enters their symptoms. This was achieved based on the logic of rules developed in Prolog, so when a user fills out the created questionnaire, the expert system will follow. the rules to conclude with the desired diagnosis; all these steps were carried out using the buchanan methodology. The result was an improvement in the accessibility of truthful information through the Internet, facilitating the management of appointments of users if they have a serious illness, or the treatment in case of a minor illness. The beneficiaries of the research were the population that required the use of the automated query application.**

***Keywords*—Automated query; buchanan; expert system; prolog; symptoms**

I. Introduction

Since the end of 2019, there is no person who has not been affected by the pandemic experienced, every area of life itself was changed in many ways, the general population is in a state of constant alarm and concern as to whether they are sick from COVID-19 or not [1].

Globally, there were waves of excessive contagions, which brought high peaks of death, in addition, very few people considered that the COVID-19 is not the only disease in recent years, even in the midst of the pandemic, people were struggling with other diseases, from the simplest [2] , such as the common flu, to the strongest and deadliest, such as cancer.

It is for all this that millions of individuals in Peru and the world resort to the use of the Internet to learn about their illness, by entering their symptoms in the search engine of their choice, in order to get immediate help, and in most cases, proceed to self-medication to avoid the cost of medical appointments or time spent in the hospital [3].

Many times when resorting to unreliable sites, people are exposed to receive information that can be harmful to them, since many of these diagnoses are written by bloggers or people who did not have a higher education and who do not have the appropriate knowledge to derive a clinical solution [4].

The main objective is the creation and implementation of a system through a web page with accurate information, this would be achieved by affiliating doctors as editors, in addition to having a relationship with a nearby medical center for prompt attention [5] [6]. The aim is to classify the patient's level of need for care, whether he/she can be treated using telemedicine or referred to a face-to-face appointment .

To achieve the best selection for the benefit of the user, at the time of putting their symptoms, is through an expert system, which is a branch of artificial intelligence [7], that allows decision making based on a set of rules, so when a patient indicates their ailment, the system will select the diagnosis with more matches [8].

The resulting benefits are the clearest, both for the person himself, because at last, he will have a place where he can have a solution to his health problem and that it is also reliable and, on the other hand, it is very beneficial for the network of hospitals and clinics affiliated to the website [9], because they will have a greater number of patients .

On the other hand, another of the great benefits that we obtain by applying directly on a web page, is the direct interaction with the user. Peru is a country with a large sector of poverty, not everyone can buy expensive devices for downloading mobile applications, but websites are more accessible to all Peruvian citizens [10]. At a general level, there is a need to modernize the prompt care of the population, especially in a health crisis, such as a pandemic.

The paper is structured as follows: in Section II, the literature review; in Section III, the methodology, where the steps to be carried out are indicated; in Section IV the results and discussions; finally in Section V the conclusion and future work.

II. Literature Review

The research principle allows the visualization in a global way, how expanded and well executed the digital format related to the health area is, speaking specifically about the search for symptoms and treatments for non-lethal situations, in the current context of the pandemic. The number of searches on websites has multiplied, to know for sure if the symptoms suffered by a person could be a serious illness, but when doing this search for symptoms on the Internet we are faced with a main problem, of finding a sea of pages with erroneous or unverified content.

With respect the subject of study, research shows the great growth of internet searches on health issues, being this way that [11] Jozsef in its research on Health Information on the

Internet, tells us that research has revealed that more and more Internet users are accessing health-related websites to search for health information and are relying on it to answer their health questions. Therefore, in the following study [12] it is confirmed that Internet searches can help people find not only the symptoms of the disease, but also various treatment options, this information is largely hosted in social media communities such as blogs, social networks, e-mails, among others [13].

Based on this, it is known that there is a need for the population to understand the medical situation they are going through, Macrohon in a study on real-time COVID-19 data visualization and information [3] reveals that studies have shown that people of all age groups search the Internet for information about their health, illnesses, care and treatment, in addition, a great variety of texts, images, graphics, files and audio and video applications that are uploaded on different websites, supported by the poor knowledge of people on health issues, affects people's search for information, since this information is of low quality and from not very credible sources.

Knowing that most of the health information found on the internet is not very reliable [12] Waltters in a study on virtual clinic care in these pandemic times, I affirm that the websites and health-related information on the internet show inconsistent quality with incorrect data and unreliable sources, the medical community has therefore begun to question the reliability of using websites to search for information.

For this reason [11] indicates that professionals who are in charge of developing websites to provide health services with respect to people's health, focusing on the needs of the users, this help is necessary to eliminate any health ignorance.

On the other hand, in the following study on the adaptation of an expert system for medical diagnosis [14] it was found that at present most websites are not properly prepared to provide health information. Many of the sites studied violated the basic guidelines for remote access, whether mobile or web, and non-governmental sites have proven to be in a better position to adapt to the new needs of patients. This is, a limitation of cyber resources for the population, so the state or government web pages related to health must adapt to the new changes quickly in order to meet the needs of the population.

To put the situation of the websites in order, author Muhammand in his research on web-based clinic management system compared by scores the level of information of the websites [15] These showed us that the level of help provided is directly referential to the type of search, which means that there are searches that will be helpful and useful, such as searching for beneficial help on websites, but a lower percentage of useful information and help on personal help sites [16].

Reviewing other research can see that in addition to being able to score the various health sites, they were also rated by percentage of truthfulness, as in the following study on user participation in health websites, [17] shows a worrying result, 87.2% of the pages that have been examined are not verified and do not have official code certificates, which could result in false information about the current health crisis in the world. Entities should improve their satellite education dissemination system as soon as possible, which would make more people literate in the area of health and personal care.

Finally, must be clear that the population is looking for a lot of information on health food care, [18] this recent study revealed that websites in Europe do not have good information quality despite having unique criteria patterns. It should be noted that only half of the websites had correct information.

In addition, the following research on the digital inclusion of health information websites [18] tells us that none of the websites found in this study had nutritional advice that met the quality criteria 100%. In conclusion, several works have been analyzed, including scientific articles and theses. Objective to implement an expert system for automatic symptom queries.

Applying an expert system in the healthcare area is an important step in obtaining reliable medical diagnoses. At this time, people are putting their physical and mental health at risk by consulting symptoms of any kind on the Internet. What intend with this research is to establish a reliable source for anyone who wants quick help. Any person would only have to go online, enter their symptoms and the system would automatically determine what these people suffer from, helping a lot to make decisions, to go to a medical center or buy basic medicines that do not require a prescription.

III. Methodology

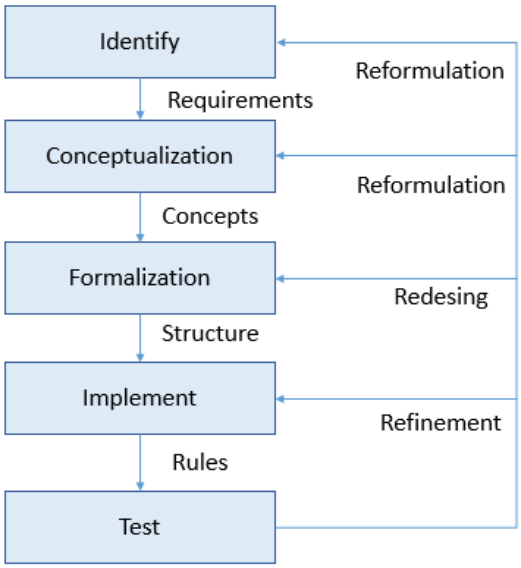


Fig. 1. Methodology Buchanan.

A. Stages of the Methodology

For the development of the expert system was used the BUCHANAN methodology shown in Fig.1. Which is a methodology based on the cascade life cycle that was used in the beginnings of software engineering, with this it is deduced that the process of construction or development of the expert

system will be carried out with a process of almost constant revision since this methodology follows the software (expert system) in its different phases [19] , which make a hierarchical life cycle; these are defined in 6 modules.

1) Identification: In this step the participants, roles, actors to build the system and their functions in the construction are identified. At this stage, the available resources and sources of knowledge are also identified, as well as the objectives or goals of the expert system.

2) Conceptualization: The knowledge of the field expert is described, knowledge obtained for the project, in order to delimit the system, this to identify and characterize the problem, the scope of the expert system is defined, this means that the specific problems to be solved by the expert system must be specified.

3) Formalization: The structure of the expert system is obtained, and with the problem well defined, was begin to identify what is required to be done in the different functions or tasks to be solved by the expert system, relevant and important concepts are identified, as well as the result of formalizing the conceptual information diagram the Bayesian network, which will be used to identify the rules of the expert system, at this stage of the methodology, the specifications are defined to first build the prototype of the knowledge base.

4) Implementation: This step begins with the development of the expert system, based on the data previously obtained, such as the identification of the study variables, the tasks of the expert system, the use cases, the Bayesian Network, the rules, among others, then the programming language will be chosen,the programming environment, which is the choice of the set of programs for the realization of the project and finally the general organization of the development, all this allows an adequate creation of the knowledge base of the expert system and of the prototype to be tested in the following step [20].

5) Test: Having the prototype, observe the behavior it has during its execution, and the key points that analyze are, the operation of the knowledge base, that is, how it reacts to the information it processes within the system and the second point is the structure of inferences, which is the response it has in the interaction.

B. Methodology Development

1) Identification:

- Problem:
  Currently, there is no website adequate to the needs of people who search for their symptoms on the Internet, and even less, an expert system that can provide a reliable basic diagnosis for the user to make health care decisions.
- Solution:
  Propose the creation of a rule-based expert system with the utility to provide basic diagnostics for a virtual triage process, thus streamlining existing medical consultations.
- Familiarization with the Domain:
  In order to carry out the familiarization and mastery, the face-to-face triage process was studied with all the corresponding procedures. Once the problems and domains have been identified, continue with the development of the expert system tasks.
- Expert System Tasks
  - Enable to register patients to the system, this module is shown in the Fig. 2.
  - Allows triage from home, a module shown on the Fig. 3.
  - To allow through a module the learning of the expert system, in order to continue expanding the knowledge base.
  - Allow the system to diagnose possible disease based on the patient's symptoms, diagnosis module.
  - Enable the system to display the results of the diagnostic, as shown in the Fig. 4.

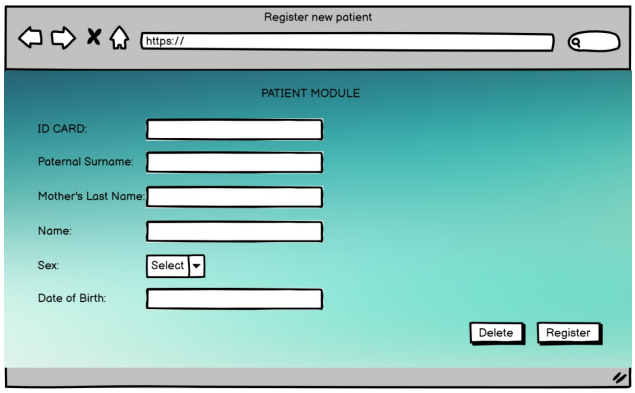


Fig. 2. Patient Registration Module.

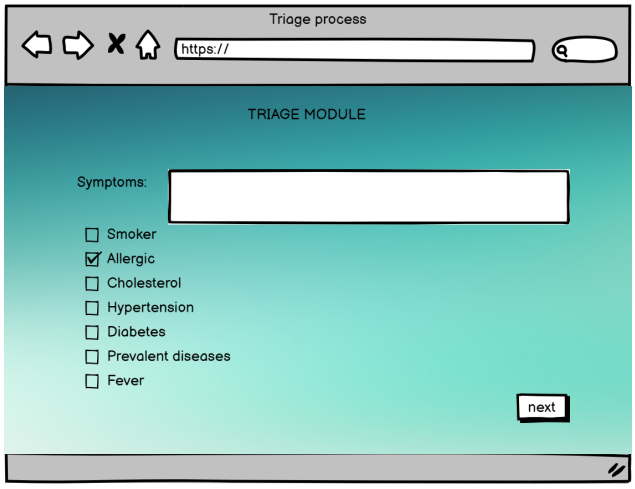


Fig. 3. Triage Module.

2) Conceptualization: In this stage of the methodology, proceeded to obtain the knowledge of the expert system, qualitative information was needed for the research, this to be able to apply it in the rule-based model to be used, from the different models that can be applied for decision-based structures, the Bayesian network was chosen, to be able to use this model and implement it in the expert system for the creation of the knowledge base, the target variable was

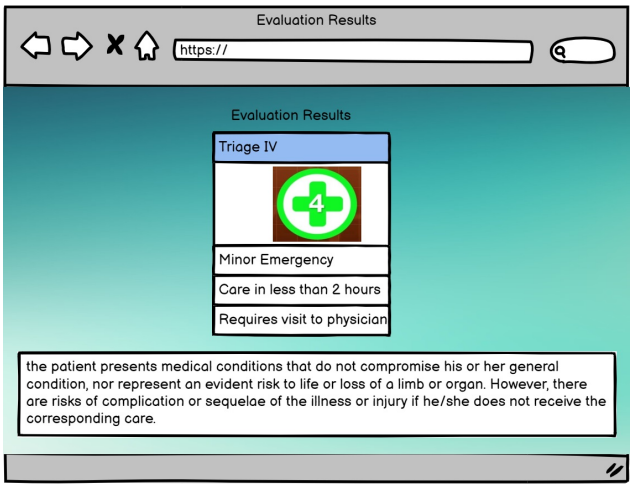


Fig. 4. Result Module.

identified as shown in Table I, as well as the observation variables that are specified in Table II, once these 2 variables were identified, the knowledge base will be created.

- Target Variable: As shown in Table I.

TABLE I. IDENTIFICATION OF THE TARGET VARIABLE

Description	Variable	Range
Symptom consultation	Triage Priority(PT)	Triage I, Triage II,Triage III, Triage IV, Triage V

- Observation Variables: As shown in Table II.

TABLE II. IDENTIFICATION OF OBSERVATION VARIABLES

Nro	Description	Variable	Range
1	Age	ED	Under Age, Adult
2	Can walk	CA	Yes, No
3	Over Weight	SP	Yes, No
4	Allergy	AL	Yes, No
5	Smoke Cigarettes	FU	Yes, No
6	Cholesterol	CO	Yes, No
7	Can Speak	HA	Yes, No
8	Fever	FI	Yes, No
9	Muscle pains	DM	Yes, No
10	Hypertension	HI	Yes, No
11	You can obey simple orders	OB	Yes, No
12	Diabetes	DI	Yes, No
13	Prevalent Diseases	EP	Yes, No
14	High pressure	PA	Yes, No
15	Can Breathe	RE	Yes, No
16	Dizziness	MA	Yes, No
17	Nausea	NA	Yes, No
18	Loss of Memory	PM	Yes, No
19	Blurred Vision	VB	Yes, No
20	Vital Signs	VB	Yes, No

3) Formalization: For this stage of the methodology, having already identified the target variable and the observation variable in the previous step, was proceed to build the chosen graphical model, which is the "Bayesian network" as shown in Fig. 5. this model is developed according to the symptoms identified in the previous section. At this stage, by using the Bayesian network graphical model, knowledge base presentation created, for this we identify the rules of inference that will be used for the construction of the knowledge base through logic programming, it should be noted that this stage of the methodology is important because it establishes the rules of inference that are used for the solution of the problem posed, the criteria for the development of the rules of inference is observed in Table III and to culminate in this stage is the one that establishes everything necessary for the implementation stage of the expert system in prolog.

Rules of inference

• Ruler 1 YES (Walk and YES SP1 and No AL and No FU and No CO) THEN TN-V
• Ruler 2 YES (Walk and YES SP1 and YES AL and No FU and No CO) THEN TN-V
• Ruler 3 YES (Walk and YES SP1 and No AL and YES FU and No CO) THEN TN-V
• Ruler 4 YES (Walk and YES SP1 and No AL and no FU and YES CO) THEN TN-V
• Ruler 5 YES (Walk and YES SP1 and YES AL and YES FU and No CO) THEN TN-V
• Ruler 6 YES (Walk and YES SP1 and YES AL and No FU and YES CO) THEN TN-V
• Ruler 7 YES (Walk and YES SP1 and No AL and YES FU and YES CO) THEN TN-V
• Ruler 8 YES (Walk and YES SP1 and YES AL and YES FU and YES CO) THEN TN-V
• Ruler 9 YES (Walk and No SP1 and No AL and No FU and No CO) THEN TN-V
• Ruler 10 YES (Walk and No SP1 and YES AL and No FU and No CO) THEN TN-V
• Ruler 11 YES (Walk and No SP1 and No AL and YES FU and No CO) THEN TN-V
• Ruler 12 YES (Walk and No SP1 and No AL and no FU and SI CO) THEN TN-V
• Ruler 13 YES (Walk and No SP1 and YES AL and YES FU and No CO) THEN TN-V
• Ruler 14 YES (Walk and No SP1 and YES AL and No FU and YES CO) THEN TN-V
• Ruler 15 YES (Walk and No SP1 and No AL and YES FU and YES CO) THEN TN-V
• Ruler 16 YES (Walk and No SP1 and YES AL and YES FU and YES CO) THEN TN-V
• Ruler 17 No (Walk and YES HA and YES SP1 and No FI and No DM and No HI) THEN TN-IV
• Ruler 18 No (Walk and YES HA and YES SP1 and YES FI and No DM and No HI) THEN TN-IV
• Ruler 19 No (Walk and YES HA and YES SP1 and No FI and YES DM and No HI) THEN TN-IV
• Ruler 20 No (Walk and YES HA and YES SP1 and No FI and No DM and YES HI) THEN TN-IV

IV. RESULT AND DISCUSSION

1) Implementation: In this step of the methodology chose to develop the Prolog expert system that gives us the tools for logic programming, this system was connected to a web environment interface that will be implemented in netbeans and will be supported by a SQL database. As step I the specifications of the use case of the system were defined.

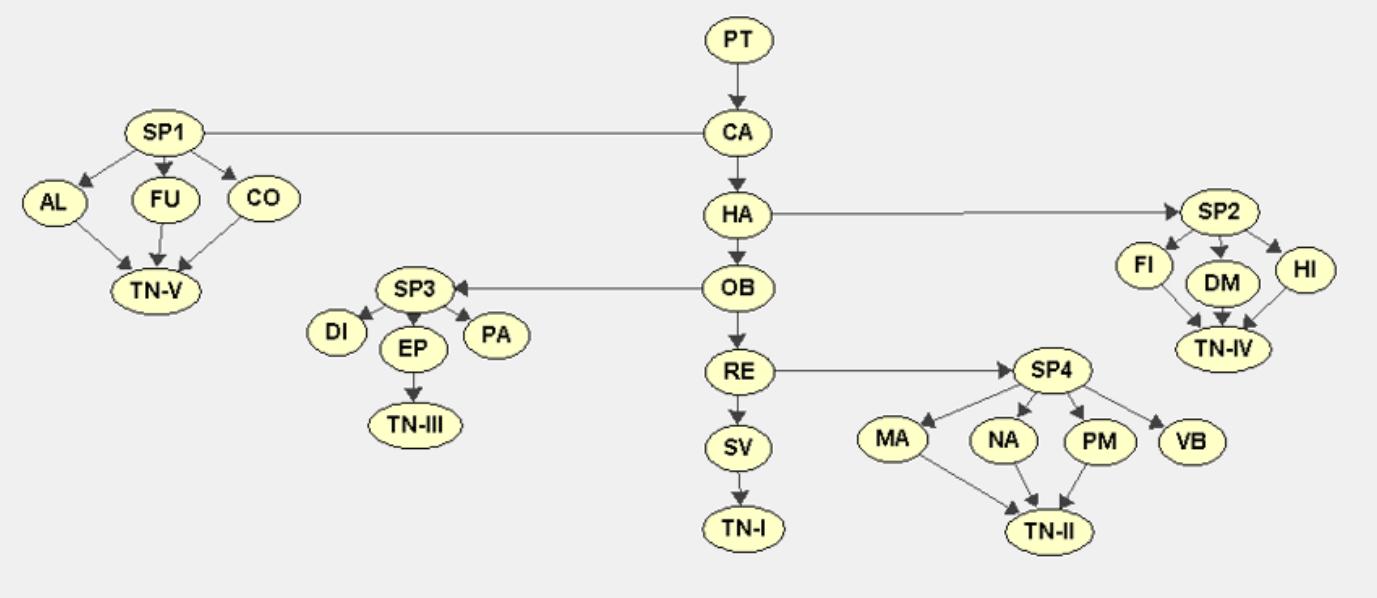


Fig. 5. Knowledge base: Graphical Model Bayesian Network.

TABLE III. RULES OF INFERENCE

Rule	Age	Weight	Commom diseases	Cholesterol	Hypertension	Memory loss	Diagnosis
Rule1	MA	Yes	Yes	Yes	No	Yes	TN-I
Rule2	ME	Yes	Yes	Yes	No	Yes	TN-I
Rule3	ME	Yes	Yes	Yes	No	No	TN-I
Rule4	MA	Yes	Yes	Yes	No	No	TN-I
Rule5	ME	Yes	Yes	Yes	Yes	No	TN-I
Rule6	MA	Yes	Yes	Yes	Yes	No	TN-II
Rule7	ME	Yes	Yes	Yes	Yes	Yes	TN-II
Rule8	MA	Yes	Yes	Yes	Yes	Yes	TN-II
Rule9	MA	Yes	Yes	Yes	Yes	No	TN-II
Rule10	ME	No	Yes	No	Yes	No	TN-III
Rule11	MA	No	No	No	No	Yes	TN-III
Rule12	ME	No	No	Yes	No	No	TN-III
Rule13	MA	No	No	Yes	Yes	Yes	TN-III
Rule14	ME	No	No	Yes	Yes	No	TN-IV
Rule15	MA	No	No	No	Yes	Yes	TN-IV
Rule16	ME	No	No	No	Yes	Yes	TN-IV
Rule17	MA	No	No	No	Yes	Yes	TN-IV
Rule18	ME	No	No	No	No	Yes	TN-V
Rule19	MA	No	No	No	No	Yes	TN-V
Rule20	ME	No	No	No	No	No	TN-V

TABLE IV. USE CASE MODULE REGISTER PATIENT

Use Case	Register Patient	
Code	CU01	
Target	Allow to register new patients in the database, in order to obtain a better diagnosis from the expert system.	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	The patient's data is recorded.	
Actors	User	
Main Flow	1.	User logs in to the system.
	2.	The user accesses the patient module.
	3.	The user presses the add button of the patient module.
	4.	The system displays the form to register the patient.
	5.	The user enters the patient's DNI, surname and first name, sex, age, date of birth and then clicks on the add button.
	6.	The system stores the patient.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

- Patient registration use case: This use case is going to allow registering patients in the database, a procedure that was performed to obtain a better diagnosis from the expert system, this can be seen in Table IV.
- Search patient use case: This use case allows the user to be able to search for patients registered in the system, as shown in Table V.
- Add triage variable use case: Use case that allows the user to add new variables for the triage process, as shown in Table VI.
- Register symptom use case: Use case that allows the user to register new symptoms that are the knowledge base for the expert system, as shown in Table VII.
- Search symptom use case: Use case that allows the user to search and list the symptoms registered in the system, with their respective characteristics, as shown in Table VIII.
- Use case obtain diagnosis: Use case that allows the user to carry out the process of diagnosing the patient, this is done by selecting the symptoms presented by the patient, as shown in Table IX.
- Consult diagnosis use case: Use case that allows the user to consult the patient's diagnosis together with a possible treatment, as shown in Table X.

This database was created in order to store the data obtained from the patient's admission as well as the results obtained from each patient's consultation to the expert system, thus expanding the knowledge base, it contains the following tables with which the database diagram shown in Fig. 6 was generated.

- Patient Table.

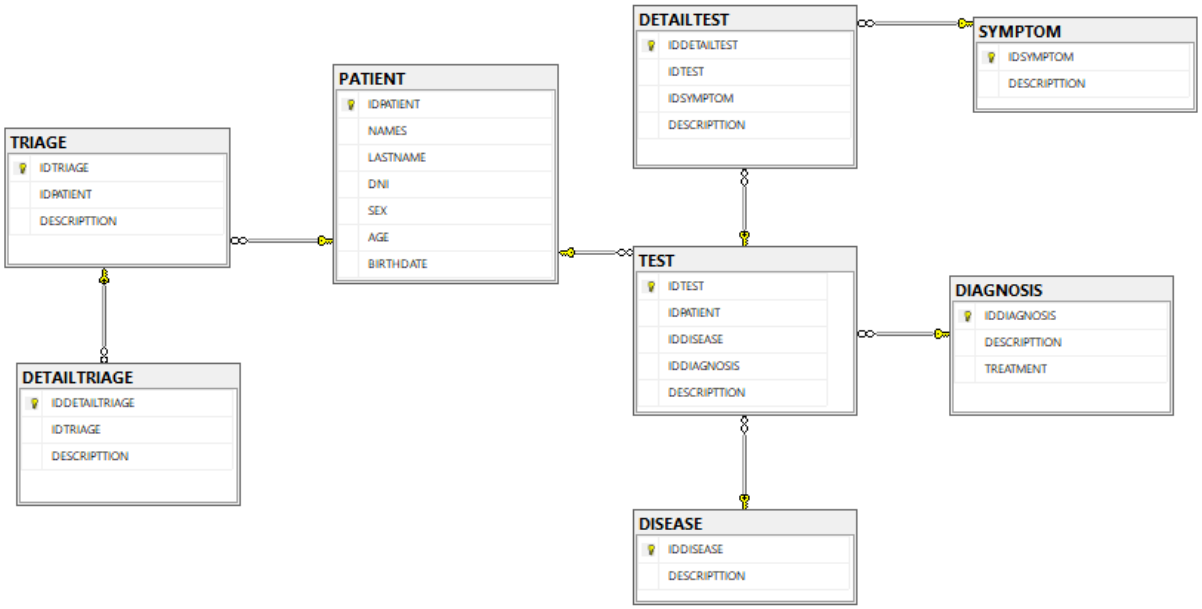


Fig. 6. Expert System Database.

```
%facts
patient(nofever).
patient(walk). %CA
patient(smoke). %FU
patient(speaks). %HA ->
patient(obeysorders). %OB
patient(prevalentdiseases). %EP
patient(vitalsigns). %SV

symptom(overweight).  %SP
symptom(allergy). %AL
symptom(cholesterol). %CO
symptom(fever). %FI
symptom(musclepain). %DM
symptom(hypertension). %HI
symptom(diabetes). %DI
symptom(highpressure). %PA
symptom(dizziness). %MA
symptom(difficultyinbreathing). %RE
symptom(nausea). %NA
symptom(memoryloss). %PM
symptom(blurredvision). %VB


reply(immediateAttention).
reply(urgentAttention).
reply(urgentcare).
reply(normalCare).
reply(nonUrgentCare).
```

Fig. 7. Statement of Facts.

TABLE V. USE CASE SEARCH FOR PATIENT

Use Case	Patient Search	
Code	CU02	
Target	Allow the user to search, consult patients registered in the system.	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	Patient data is displayed.	
Actors	User	
Main Flow	1.	User logs in to the system.
	2.	The user accesses the patient module.
	3.	The user clicks on the search button and enters the DNI of the patient to search for.
	4.	The system displays the form to register the patient.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

TABLE VI. USE CASE ADD TRIAGE VARIABLE

Use Case	Add triage variables	
Code	CU03	
Target	Allow user to add variables for the triage process.	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	Data was saved correctly.	
Actors	User	
Main Flow	1.	User logs in to the system.
	2.	The user accesses the triage module and presses the add button.
	3.	The system displays a form to add more triage variables.
	4.	The user fills in the data and clicks on save.
	5.	The system stores the information.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

TABLE VII. USE CASE REGISTER SYMPTOM

Use Case	Record symptom	
Code	CU04	
Target	Allow user to register new symptoms (knowledge base).	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	The symptom is recorded.	
Actors	User	
Main Flow	1.	The user accesses the symptoms module.
	2.	The user clicks on the add new symptom button.
	3.	The user enters the symptom code, name, description, then clicks the add button.
	4.	The symptom is saved in the system.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

TABLE VIII. USE CASE SEARCH FOR SYMPTOM

Use Case	Record symptom	
Code	CU05	
Target	Allow the user to search for the symptoms registered in the system.	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	The searched symptom is displayed in a table.	
Actors	User	
Main Flow	1.	The user accesses the symptoms module.
	2.	The user clicks on the search symptom button and enters the symptom to search for.
	3.	The system displays a table with the searched symptom and its respective characteristics.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

TABLE IX. USE CASE GET DIAGNOSTIC

Use Case	Obtain diagnosis	
Code	CU06	
Target	Allow the user to make the diagnosis to the patient, by means of the choices of the symptoms presented by the patient.	
Preconditions	The user must be logged in.The user must have permission to access this module.	
Post conditions	Not applicable.	
Actors	User, system.	
Main Flow	1.	The user enters the diagnostic module.
	2.	The user clicks on the test button.
	3.	The system redirects to a page where the patient triage process is performed first.
	4.	The system initiates the diagnostic process by means of a test.
	5.	The user enters the patient's symptoms.
	6.	The system displays the results of the diagnosis.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

TABLE X. USE CASE CONSULT DIAGNOSTICS

Use Case	Consult diagnosis	
Code	CU07	
Target	Allow the user to query the patient's diagnosis by displaying a treatment.	
Preconditions	Perform CU10.	
Post conditions	Not applicable.	
Actors	User	
Main Flow	1.	The user enters the diagnostic module.
	2.	The user clicks on the search button and enters the patient's ID number.
	3.	The system displays the results of the diagnosis performed, providing a possible treatment.
Performance	High.	
Frequency	Infrequent.	
Priority	High.	

- Triage table.
- Table Triage detail.
- Symptom table.
- Diagnostic table.
- Test Table.
- Table DetailsTest.
- Disease table.

2) Test: In this stage of the methodology the expert system developed in prolog was tested based on the rules established in the Bayesian network, starting with the statement of facts as shown in Fig. 7 where the first step was to enter the possible symptoms that a patient may have for the triage process, as well as the symptoms that the expert system will handle for this stage of the development process, this part is very important because they are the statements regarding the knowledge base, then the rules were typed as shown in Fig. 8. The rules are the knowledge representation of the expert system that is expressed in natural language and through a conditional, in this case to determine a diagnosis the rules evaluate the symptoms of the knowledge base together with the symptoms that the patient is going to perform and compares them with a priority of attention, which allowed to obtain a good diagnosis. Subsequently, the web application interface was developed, which can be seen in Fig. 9. In order to make the system

```
%triage evaluation
%immediate attention
prioridadI(vitalsings,difficultyinbreathing).
%urgent attention
prioridadII(dizziness,nausea,memotyloss,blurredvision).
%urgent care
prioridadIII(obeysorders,diabetes,prevalentdiseases,highpressure).

priorityIV(speaks,normalCare).
priorityIV(speaks,nofever,normalCare).
priorityIV(speaks,walk,musclepain,normalCare).
priorityIV(speaks,nofever,musclepain,hypertension,normalCare).

priorityV(walk,nonUrgentCare).
priorityV(walk,allergy,nonUrgentCare).
priorityV(walk,smoke,nonUrgentCare).
priorityV(walk,cholesterol,nonUrgentCare).
priorityV(walk,allergy,smoke,nonUrgentCare).
priorityV(walk,allergy,smoke,cholesterol,nonUrgentCare).


%REGLAS PRIORIDADV
rule(P,R):-patient(P),reply(R),priorityV(P,R).
rule(P,S,R):-patient(P),symptom(S),reply(R),priorityV(P,S,R).
rule(P1,P2,R):- patient(P1),patient(P2),reply(R),priorityV(P1,P2,R).
rule(P,S,R):-patient(P),symptom(S),reply(R),priorityV(P,S,R).
rule(P1,S,P2,R):-patient(P1),symptom(S),patient(P2),reply(R),priorityV(P1,S,P2,R).
rule(A,B,C,D,E):-patient(A),symptom(B),patient(C),symptom(D),reply(E),priorityV(A,B,C,D,E).

%REGLAS PRIORIDADIV
ruleIV(P,R):-patient(P),reply(R),priorityIV(P,R).
ruleIV(P1,P2,R):- patient(P1),patient(P2),reply(R),priorityIV(P1,P2,R).
ruleIV(P1,P2,S,R):-patient(P1),patient(P2),symptom(S),reply(R),priorityIV(P1,P2,S,R).
ruleIV():-patient(P1),patient(P2),symptom(S1),symptom(S2),reply(R),priorityIV(P1,P2,S1,S2,R).
```

Fig. 8. Expert System Rules.

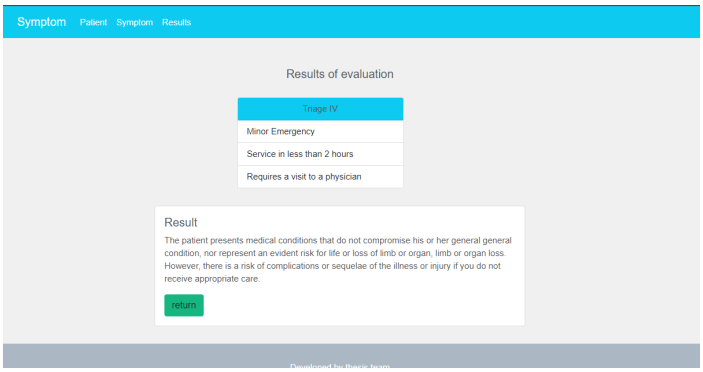


Fig. 9. Test Result.

usable, intuitive and generate a comfortable experience for the user, this interface displays the results obtained from the expert system.

3) Symptom Comparisons on the Web: The results in terms of internet search show interesting data such as for example that the words in terms of search for symptoms related to COVID-19, such as fever, dry cough, sore throat, are the most searched on the web to corporation of other diseases or symptoms such as could be allergy or any other, in Fig. 10. shows the trend of internet search with respect to some symptoms, in this graph you can see that among the symptoms or diseases shown, the term COVID-19 has been the most searched during the last 4 months from July to October 2021, it should be noted that these data are only from searches made in Peru, these data are important to know because they indicate that there is a high trend with respect to queries on the web related to health issues.

V. Conclusion and Future Work

As a final part of this research, it is concluded that there is a perpetual need for digital improvement in healthcare, people will only be able to change their way of thinking by obtaining truthful sources of diagnosis, when they are presented quickly.

The implemented system is a good short term solution that helps to put an end to the problem of the lack of information

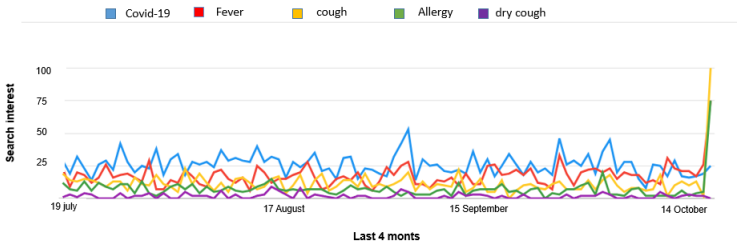


Fig. 10. Comparative Symptom Search.

sources and successfully benefits all types of users, both those who want to search for diagnoses and appointments on the web, as well as those who want to offer their health services, both hospitals and clinics.

For future work, we recommend analyzing well the sequence pattern that needs to be regulated, the logical aspect is one of the most representative parts of the expert system, the rules must comply with a specific and well-structured order. It is recommended to continue with the investigation applying other types of machine learning algorithms.

Acknowledgment

To thank the support of the University of Sciences and Humanities, through its research institute for the support.

References

[1] J. J. E. MacRohon and J. H. Jeng, "A Real-Time COVID-19 Data Visualization and Information Repository in the Philippines," *2021 9th International Conference on Information and Education Technology, ICIET 2021*, pp. 443–447, 2021.

[2] J. Rai, R. C. Tripathi, and N. Gulati, "A comprehensive survey of IT sectors affected by covid-19," *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020*, pp. 52–54, 2020.

[3] S. Ayani, F. Sadoughi, R. Jabari, K. Moulaei, and H. Ashrafi-Rizi, "Evaluation Criteria for Health Websites: Critical Review," *Frontiers in Health Informatics*, vol. 9, no. 1, 2020.

[4] K. József, "Health information on the internet," *Orvosi Hetilap*, vol. 159, no. 22, pp. 855–862, 2018.

[5] A. J. C. Sotomayor and L. Andrade-Arenas, "Mobile application oriented to the attention of blood donors in the medical centers of northern lima," in *2020 IEEE Engineering International Research Conference (EIRCON)*. IEEE, 2020, pp. 1–4.

[6] C. Watters, B. Miller, M. Kelly, V. Burnay, Y. Karagama, and E. Chevretton, "Virtual voice clinics in the COVID-19 era: have they been helpful?" *European Archives of Oto-Rhino-Laryngology*, vol. 278, no. 10, pp. 4113–4118, 2021. [Online]. Available: https://doi.org/10.1007/s00405-021-06643-6

[7] S. A. H. Morales, M. F. M. Antayhua, and L. Andrade-Arenas, "Development of predictions through machine learning for sars-cov-2 forecasting in peru," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0121188

[8] J. J. Kim and G. A. Bekey, "Adaptive abstraction in expert systems for medical diagnosis," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 1992-June, pp. 345–352, 1992.

[9] J. Muhammad, "Web-based Clinic Management System (CMS)," vol. 8, no. 05, pp. 131–135, 2019.

[10] F. Andrade-Chaico and L. Andrade-Arenas, "Projections on insecurity, unemployment and poverty and their consequences in lima's district san juan de lurigancho in the next 10 years," in *2019 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, 2019, pp. 1–4.

[11] J. Imlawi, "Health website success: User engagement in health-related websites," *International Journal of Interactive Mobile Technologies*, vol. 11, no. 6, pp. 49–64, 2017.

[12] S. Boon-Itt, "Quality of health websites and their influence on perceived usefulness, trust and intention to use: An analysis from Thailand," *Journal of Innovation and Entrepreneurship*, vol. 8, no. 1, 2019.

[13] F. Afsana, M. A. Kabir, N. Hassan, and M. Paul, "Automatically Assessing Quality of Online Health Articles," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 591–601, 2021.

[14] N. E. Youngblood, "Digital inclusiveness of health information websites," *Universal Access in the Information Society*, vol. 19, no. 1, 2020.

[15] M. T. Thielsch, C. Thielsch, and G. Hirschfeld, "How Informative is Informative? Benchmarks and Optimal Cut Points for E-Health Websites," *Proceedings of the Mensch und Computer 2019*, 2019.

[16] L. Z. Gao, "Study of health education for university students in the new period," *Proceedings 2011 International Conference on Human Health and Biomedical Engineering, HHBE 2011*, pp. 929–931, 2011.

[17] S. Valizadeh-Haghi, Y. Khazaal, and S. Rahmatizadeh, "Health websites on COVID-19: Are they readable and credible enough to help public self-care?" *Journal of the Medical Library Association*, vol. 109, no. 1, 2021.

[18] V. L. Carmen Milagros, "Universidad Nacional Mayor de San Marcos Analysis of the quality and content of web pages in Spanish about their nutritional advice to lose weight To qualify for the Professional Degree of Bachelor of Nutrition," 2020.

[19] E. J. G. Buitrón, D. C. Corrales, J. Avelino, J. A. Iglesias, and J. C. Corrales, "Rule-based expert system for detection of coffee rust warnings in colombian crops," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 5, pp. 4765–4775, 2019.

[20] D. C. Brock, "Learning from artificial intelligence's previous awakenings: The history of expert systems," *AI Magazine*, vol. 39, no. 3, pp. 3–15, 2018.

Implementation of a Web System to Detect Anemia in Children of Peru

Ricardo Leon Ayala, Noe Vicente Rosas, Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—Now-a-days, anemia is considered a worldwide problem that not only seriously affects our health, but also has economic and social consequences. Therefore, seeks to provide a solution to the problem to detect anemia with a non-invasive method quickly, simple and low-cost way. In this research work, a web system was designed applying the scrum methodology to detect anemia and simplify the detection process of anemia in Peruvian children. In addition, this study shows as a result a technological prototype that helped in the diagnosis of anemia ; at the same time it provides food recommendations to patients to combat anemia efficiently, with a variety of recipes and ingredients that are available in any home, helping in the recovery process. In addition, the analysis carried out on children with anemia in Peru is shown, where it is known that Puno is the most affected department. With respect to the capital Lima, the most affected district is Callao. However, this amount is expected to drop considerably in the coming years.**

***Keywords*—Anemia; diagnosis; health; scrum; web system**

I. Introduction

Anemia is considered today a global problem, which not only seriously affects health, but also has economic and social consequences. Anemia is characterized by a decrease in the level of red blood cells or by an abnormal reduction in hemoglobin of less than 11 g/dl for children under 6 years of age and 12 g/dl for children over 6 years of age [1]. Therefore, it is currently known that 2 billion people suffer from this disease worldwide [2].

A review of studies, conducted in 19 European countries, found that between 2 and 25 percent of children aged 6 to 12 months were iron deficient. In addition [3], preschool children in low- and middle-income countries are estimated to be anemia.

On the other hand, a diagnosis carried out on 112714 children, indicated that the low consumption of iron in food is one of the main causes of infant anemia in India. Since 80% of children between 12 and 23 months had anemia and 69.5% of children under 5 years of age were anemics [4].

Peru, is no exception [5], since anemia in the country is mostly found in the low-income Andean regions, since several studies affirm that the population of approximately 463 districts has this disease, which is a major problem that threatens public health.

The Ministry of Health [6], published a report, indicating that 6 out of 10 children between 6 and 12 months of age have anemia, a worrying situation that requires further follow-up.

As we know, children are more prone to anemia, and this is due to iron deficiency, the same happens with children under 12 months, in premature infants anemia occurs in the first months of life [7]. In addition, it was identified that Puno, Pasco, Cusco, Loreto, Ucayali and Madre de Dios are the places most affected by anemia [8]. For this reason, this research work offers a viable solution to the problem of anemia, using a web system, in order to contribute to the reduction of the high percentage of anemia in Peru.

The proposed research, seeks to provide a solution to the problem, to detect anemia with a non-invasive method, in a fast, simple and low-cost way. It allows us to reduce the process of blood analysis, leaving aside the conventional invasive methods to measure hemoglobin levels and identify anemia. Thanks to the use of technology, anemic children under 8 years of age are easily recognized in Peru, thus reducing the impact of that disease on society and treating it at the earliest possible stage.

The objective of the research, is to design and develop a web-based system, applying the scrum methodology, to non-invasively detect and simplify the process of detecting anemia in children in Peru.

This research work is conformed as follows: Section II presents the Literature Review. Section III defines the methodology to be applied to the project and defines each stage in a theoretical way. Section IV describes the case study, Section V describes and shows the result and discussion obtained. Finally, Section VI defines the conclusion and future work.

II. Literature Review

The problem of anemia in children is common today, both in Lima and in the province for different reasons. The different results, methodologies and limitations found in the different articles on anemia in children are analyzed below.

The application of expert systems is applied in different fields, such as the following article [9]. In it applied expert systems oriented to deep learning. Applying artificial intelligence and taking into account the best known models such as the neural network in the short and long term. They carried out the prediction by collecting information to perform the training of the model, in addition, this information helped them to discover more accurate patterns. To demonstrate the effectiveness of the predictive model, they applied this knowledge in the field of air pollution and power generation. In both cases the prediction had a high success rate when comparing the results with real data.

On the other hand, according to research [10], indicates that the development of a non-invasive mobile application based on the pulse oximetry method and the use of a sensor, based on oxygen saturation values and hemoglobin concentration, can detect anemia in children, facilitating and benefiting the low-income population, since parents who suspect that their children may have anemia, have to submit them to clinical examinations, which are usually expensive and require specialized equipment. Adding to this, the vast majority of children with this disease are in rural areas with few resources, which makes it a bit complicated to have enough money to carry out the corresponding tests.

In other words [11], the creation of a mobile application and the use of additional tools would help to simplify the process of diagnosing anemia in an easy, simple and non-invasive way, making it accessible to anyone with no prior medical knowledge and low resources. However, in another research, another method is described to detect anemia in a non-invasive way by incorporating an algorithm based on neural networks in the mobile application, to detect hemoglobin levels by simply taking pictures of the fingertips.

Similarly, in the following thesis [12], presented an app, that provides recommendations to reduce poor nutrition, in case of anemia, in children in the school "Apostle of Punchauca", the mobile application was developed in order to reduce cases of anemia, source of information for parents, in addition, to prevent future cases of anemia in children, which is why thanks to its application they were made aware of the risks of anemia, how to combat it and know how to identify it.

In the same way, another research, details that used a similar methodology as the previous author, highlighting that the development of the application, with an image analysis algorithm, that analyzes the color data and image metadata allows quantitative and non-invasive self-mea,surement of hemoglobin levels in the patient's blood with high accuracy without the need for any accessories or calibration equipment [13].

On the other hand, [14] also implements a server-side algorithm to process the collected data, employing machine learning algorithms, that are trained using deep learning concepts to enable it to publish accurate results by simply analyzing fingertip color. The author [2] considers that anemia is originated when the concentration of hemoglobin within the red blood cells it is below normal.

In conclusion, different research works have been studied with similar methods that have successfully achieved the proposed objectives. Thanks to this, can highlight that the use of algorithms allows a more accurate diagnosis to detect anemia with a non-invasive method, achieved with the help of technology allowing access to any person.

III. Methodology

The agile scrum methodology, was used for the elaboration of this research work. It is considered agile because it uses incremental and iterative process approaches. On the other hand, it has proven to be more useful than the traditional waterfall model, because it improves the productivity of the processes and helps to reduce the time consumed for its realization. In a traditional waterfall model, planning is done before testing, and the process is managed in phases and once it is done, it is not possible to go back to the previous phase. On the other hand, scrum can make changes at any stage to improve results [15]. The scrum flow is shown in Fig. 1.

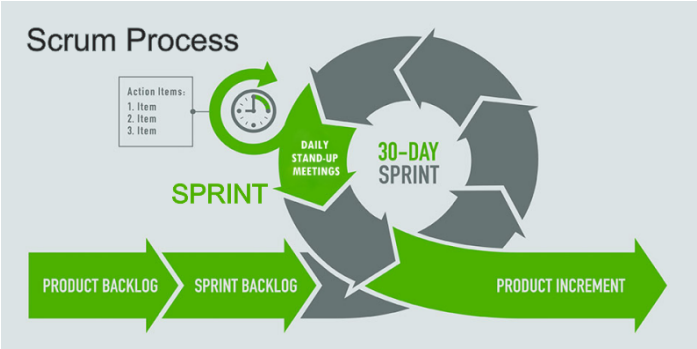


Fig. 1. Scrum Flow.

A. Scrum

The use of scrum involves integrating roles between teams, an advantage of using scrum is that it allows teams to adapt quickly to manage and plan their work, since each step of scrum allows you to plan, design, develop and test the code, all these activities are divided between the following roles [15]

B. Scrum Practices

Scrum has practices, also called Scrum formalities: Daily Meeting, Sprint Review, Sprint Planning Meetings and Sprint Retrospective [16]:

1) Daily Meeting: The meeting is held every day, where the team discusses the problems and the process of the project

2) Sprint Review: A meeting is held in which the team presents the results at the end of each Sprint to the owner.

3) Sprint Planning Meetings: A meeting is held to define the activities to be carried out during a Sprint.

4) Sprint Retrospective: A meeting is held to evaluate the team's unemployment during the Sprint and the practices to be carried out in order to improve team productivity.

C. Scrum Development Structure

1) Requirements Stage: It is established to obtain the planning requirements of the project.

2) Definition of User Stories: It is the description of the functionality of the system, for this the requirements are analyzed, in collaboration between the client and the team, the user stories will be improved throughout the life of the project.

3) User Stories Prioritization: The priority of each of the user stories [17], is established and the order in which they be developed is determined, taking into account which is the most fundamental for the project and thus following a hierarchy.

4) Analogous Estimation: When the user stories are obtained, they go through an examination, where the use of a tool or an estimation method is required [18], this consists of a study of each of the user stories, referring to the time that the scrum team believes can develop it, in this phase involves both the time and the resources and expenses that may be needed.

5) Creation of the Product Backlog: User stories, priorities and required tasks are included in a Product Backlog. It includes the features or short and long term requirements and functionalities, that have been defined jointly by the development team and Product Owner, and also establishes the points or level of effort for each task. During the sprint process [19], the assigned tasks are movable depending on the requirements of the Product Owner.

6) Definition of Sprint Speed and History Points: The speed and time of development of each Sprint, is analyzed through an analysis that varies according to the experience of the Scrum team [20].

7) Sprint Presentation: The established Sprint is presented according to the requirements and the determined times, in this phase tests are performed on what is presented by the scrum team, being accepted or denied, this is repeated according to the number of sprint that contains a project.

8) Feedback: At the end of the project or each sprint, the scrum team, holds a feedback meeting, where they analyze everything developed in order to know their successes and failures in all aspects, in order to improve in the next project or sprint.

D. Development Tools

1) C (Sharp) programming language: C is a modern programming language, based on the c++ language and taking functions from other languages such as Java. The programming is mainly used object oriented. Currently it is still a language that is eating in the computer world and is mainly used for web development, software development and mobile applications [21].

2) ASP.NET Technology: ASP.NET technology, is a program architecture created on the basis of several programming languages, developed by Microsoft Corporation. In comparison with other developed models [22], ASP.NET stands out for having tools that help the programmer. In addition, has a high-level interactive graphical interface, a large customization layer, easy operation and high efficiency, when executing development processes, such as: creation of dynamic web pages. In addition ASP.NET incorporates servers, request, responsive pages, etc. In this way, it facilitates the development of a web. That is why, in the present research work ASP.NET was used for the development of a web site.

3) Microsoft Visual Studio: It is a modern development environment, which can be used to create programs, applications and software components for Microsoft Windows. Using object-oriented promotion, the program text consists of a set of descriptions of components when certain events occur. It uses the principles of component inheritance, which allows the creation of properties, events and methods [23]. The concept of Visual programming is implemented in the Visual Studio environment with the help of Windows.

TABLE I. USER STORIES

N°	User Stories
H1	I as a user want the Web System to allow me to log in with my Gmail, Outlook or other alternatives to have a more convenient access.
H2	As a user, I want the Web System to allow me to register my personal data to have access to all the functionalities.
H3	As a user, I want the Web System to inform me how to perform the test in order to have an assertive diagnosis.
H4	I as a user want the Web System to allow me to perform a test to diagnose my status.
H5	As a user, I want the Web System to allow me to perform the diagnosis through my own photos in order to obtain a more efficient result.
H6	As a user I want the Web System to provide me with the statistics and results according to my diagnosis to know why the diagnosis has been positive or negative.
H7	As a user, I want the Web System to provide me with dietary recommendations, diets and information based on my diagnosis to improve my current condition.
H8	I as a user want the Web System to have a selective food section according to my preferences in order to get personalized recommendations.
H9	I as a user want the Web System to show a section of recipes with high nutritional level to strengthen my hemoglobin level.
H10	I as a user would like the Web System to allow me to provide feedback and comments on the recipes to inform the effectiveness of each recipe.

4) Balsamiq: Balsamiq mockup,is one of the software tools used for design or prototyping, also to develop the user interface of an application. Balsamiq Mockup [24],[25], makes it easy to create the user interface, by providing you with the necessary tools to facilitate the creation of the application design. That is why Balsamiq was used, for the creation of the research prototypes.

5) Development architecture: For the development of the application, both front-end and back-end tools are needed, that is why in Fig. 2, the development architecture is shown, in each section the tools to be used are listed.

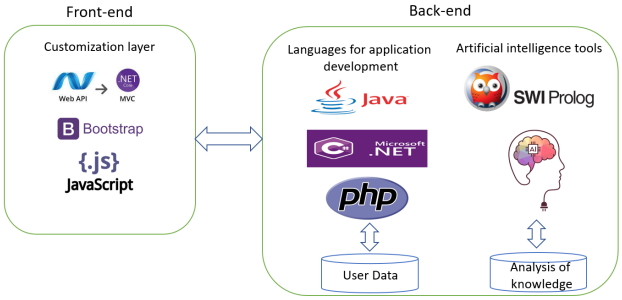


Fig. 2. Development Architecture Diagram.

IV. CASE STUDY

A. Start-up Stage

1) User Stories: After knowing the functional requirements, user stories were created, which were defined based on the requirements, as shown in Table I.

B. PLanning Stage

1) Analogous Estimation: In this stage, the complexity estimation was made with respect to the development of each

story, for this, the qualification range goes from 1 to 13, 1 is the lowest complexity level and 13 is the highest. To make the estimation, H1 and H2 are taken as a base, since this is where the "Login" is developed, so as an example, H3 is located in column 2, which indicates that H3 is 2 times more complex to develop than H1 and H2, the same logic applies to the other stories. Thanks to the Analogous Estimation, it is possible to present an idea about the functional development of each story for a future project. To arrive at the following result, the team discusses the estimation of each story and classifies them according to criteria. As shown in Table II.

TABLE II. ANALOGOUS ESTIMATE

	1	2	3	5	8	13
H1	Login					
H2	Login					
H3		2 Login				
H4					8 Login	
H5						13 Login
H6				5 Login		
H7			3 Login			
H8				5 Login		
H9				5 Login		
H10		2 Login				

2) Creation of the Product Backlog: For the creation of the complete backlog, it was ordered according to its history, priority and estimation. This information is obtained from the analog estimation table and the prioritization of the user stories. Table III, shows the Product Backlog.

3) Sprint Speed Definition and History Points: o carry out the following chart, it was decided together with the development team, to divide the sprint into four, in addition to detailing which user stories belong to each sprint. For example, H1, H2 and H3 belong to sprint 1. As shown in Fig. 3, sprint 1 has 4 points, this result was obtained thanks to the information of the analog estimation table, where the effort points obtained by the stories were added, H1 and H2 obtained 1 point each and H3 obtained 2 points, the sums of these stories belonging to the first sprint gives us as a result 4 points. sprint velocity refers to the estimated time for the completion of a sprint, the velocity is determined by the development team. In total, there are 45 points of history and an average of 30 points of speed.

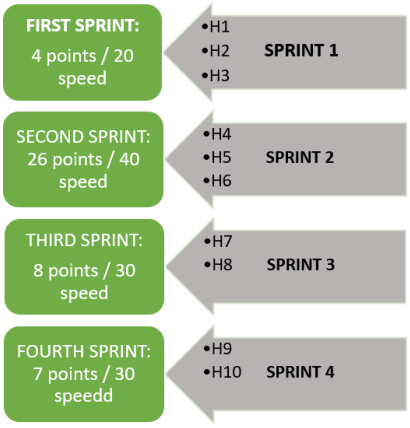


Fig. 3. History and Speed Points.

The estimated duration of the project is 4 months, Table IV shows the duration of each Sprint, as well as the estimated time to develop each functionality.

TABLE III. PRODUCT BACKLOG

User Stories	Priority	Estimation
H1: As a user, I want the Web System to allow me to log in with my Gmail, Outlook or other alternative accounts to have a more convenient access..	1	1
H2: As a user, I want the Web System to allow me to register my personal data to have access to all the functionalities.	2	1
H3: As a user, I want the Web System to inform me how to perform the test in order to have an assertive diagnosis.	3	2
H4: I, as a user, want the Web System to allow me to perform a test to diagnose my status.	4	8
H5: As a user, I want the Web System to allow me to perform the diagnosis through my own photos in order to obtain a more efficient result.	5	13
H6: As a user, I want the Web System to provide me with statistics and results according to my diagnosis to know why the diagnosis has been positive or negative.	6	5
H7: As a user, I want the Web System to provide me with dietary recommendations, diets and information based on my diagnosis to improve my current condition.	7	3
H8: I, as a user, want the Web System to have a selective food section according to my preferences in order to get personalized recommendations.	8	5
H9: I, as a user, want the Web System to show a section of recipes with high nutritional level to strengthen my hemoglobin level.	9	5
H10: I, as a user, want the Web System to allow me to provide feedback and comments on the recipes to inform the effectiveness of each recipe.	10	2

TABLE IV. SPRINT BACKLOG

Interface	Duration
Web system to detect anemia	4 months
Sprint 1: Login Interface	1 week
Sprint1: Registration Interface	1 week
Sprint 1: Home Interface	2 week
Sprint 2: Input evaluation	2 weeks
Sprint 2: Creation of activities	2 weeks
sprint 2: Creation of interactive games	2 weeks
Sprint 3: Activity customization features	2 weeks
Sprint 3: Interactive game customization features	2 weeks
Sprint 4: Results reports	2 weeks

V. RESULT AND DISCUSSION

A. Development of Prototypes by user Stories

1) Sprint Presentation: In this stage of development, the prototypes of the web page are shown, ordered according to the sprint and its estimation. Each image specifies which story it belongs to and the criteria they have had for the creation of the design.

- First Sprint: As shown in Fig. 3, the first sprint has a

total of 4 points, which include H1, H2 and H3, the design of each of them will be shown below.

H1: As a user, I want the Web System to allow me to log in with my Gmail account, Outlook or other alternatives to have a more convenient access.

For Story 1 we show the page where the user can log in, plus you have the option to register using your Gmail or Facebook account. It was decided to split the structure of the page in order to focus the user's attention to the form. this is shown in Fig. 4.

Fig. 4. Start of Session.

H2: As a user, I want the Web System to allow me to register my personal data to have access to all functionalities. Fig. 5 shows the user registration form in case you do not have a user name and password. The data requested are basic but necessary to access the system's functionalities.

Fig. 5. Registration.

Fig. 6 shows, the main page, in the menu there is a button that takes you to the diagnostic test. Continuing with the content of the page, it has a section where it includes the "Informative Guide", this button will take you to the previous guide to perform the diagnosis.
H3: As a user, I want the Web System to inform me how to take the test in order to have an assertive diagnosis. A guide is shown in Fig. 7, which was designed for the user to follow and take into account the recommendations before performing the test, in order to avoid errors during the test.

Second Sprint:

It consists of a total of 26 story points, which is a more complex indication for system development and prototyping. The design of each story is shown below.

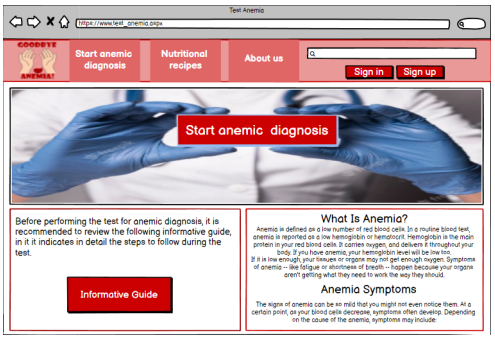


Fig. 6. Home Page.

Fig. 7. Informative Guide.

H4: I, as a user, want the Web System to allow me to perform a test to diagnose my status. The test design is shown in Fig. 8. The test is developed in two stages, the first stage consists of answering various questions necessary for the diagnosis.

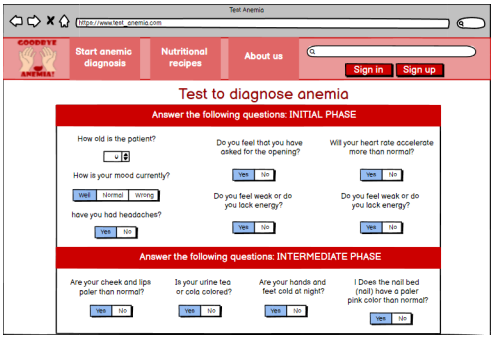


Fig. 8. Diagnostic Test.

H5: As a user, I want the Web System to allow me to perform the diagnosis through my own photos in order to obtain a more efficient result. After having carried out a previous questionnaire, the next section allows us to upload a photo of the patient's nails. upload a photo of the patient's nails, in this way with the help of artificial intelligence the application will detect, based on the color of the nails, whether anemia is present, this analysis directly influences the final diagnosis. All

this is shown in Fig. 9.

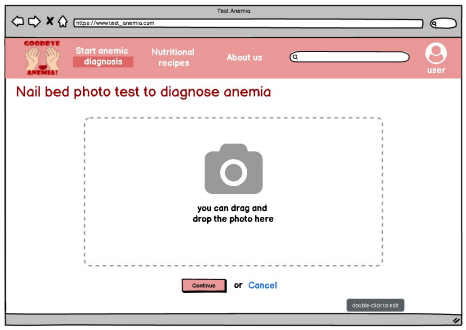


Fig. 9. Photo Diagnostic Test

H6: As a user, I want the Web System to provide me with statistics and results according to my diagnosis to know why the diagnosis has been positive or negative. As shown in Fig. 10, the results of the diagnosis are shown in the following bar graph: on the x-axis is the result of the questions and on the y-axis is the maximum score. It should be added that it shows the result "Negative" or "Positive", this result varies depending on the patient's diagnosis. The Recommendations section will be discussed later.

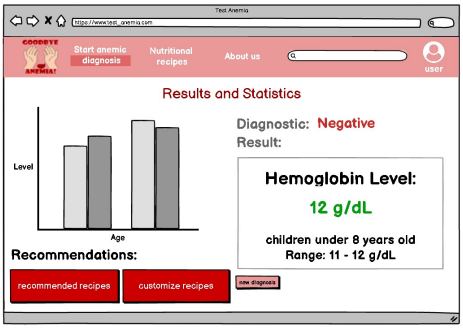


Fig. 10. Diagnostic Results Table.

Third Sprint:
H7: As a user, I want the Web System to provide me with dietary recommendations, diets and information based on my diagnosis to improve my current condition. The recommendations section is shown in Fig. 10, however this section is shown in Fig. 11.

H8: As a user, I want the Web System to have a selective food section according to my preferences to obtain personalized recommendations. As shown in Fig. 11, in the recommendations section, the patient has the possibility to select the foods of his or her preference, this information will later be displayed in a section of recipes, which include these previously selected foods in order to personalize the patient's food in a certain way.
Fourth Sprint:

H9: I, as a user, want the Web System to show a section of recipes with high nutritional level to strengthen my hemoglobin level. As shown in Fig.

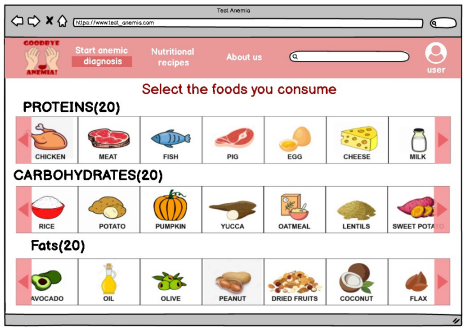


Fig. 11. Selective Food Section.

12, the web system has a recipe section, where the patient has the possibility of choosing the recipes of preference, and also where the recipes are displayed with the foods chosen in Fig. 11.

H10: I, as a user, want the Web System to allow me to give feedback and comments on the recipes to inform about the effectiveness of each recipe. Finally, in Fig. 12, the patient can rate the recipe using star points ranging from 1 to 5, as well as make a comment on the recipe in order to help or motivate future patients to try the recipes.

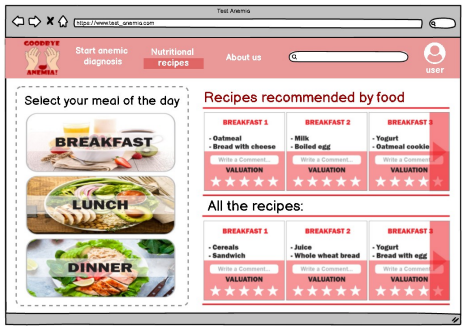


Fig. 12. List of Recipes to Combat Anemia.

B. Survey Conducted for the Creation of user Stories

For the development of the requirements, a survey was conducted among 40 people, including fathers and mothers in Lima-Peru. The survey was conducted with the help of the Google Forms tool. The questions were formulated with the purpose of knowing the user's preferences. Likewise, the answers were limited to the conditional "true" and "false". Table V, below shows the questions asked to the respondents.

1) Survey Analysis: According to the survey, it was obtained, the following results, which helped us to define the list of requirements, for the development of the web system, that will allow us to detect anemia and include some additional options that will help in the improvement of the patient as shown in Fig. 13.

- Of the 40 people surveyed, 32 opted to register before, accessing the website because it would give them a personalized profile and access to additional functions.

TABLE V. QUESTIONS ASKED TO RESPONDENTS

N°	Survey Questions
1	Have you suffered from anemia?
2	Have you ever used a website that detects anemia?
3	Would you like to enter directly to the website or do you want to register?
4	Would you like the page to allow login with your account from Gmail?
5	Would you like to carry out a previous survey related to your current state of health?
6	Do you consider the implementation of AI (Artificial Intelligence) necessary for the diagnosis?
7	Would you like to show the results with graphs and tables?
8	Would you like the website to give you advice and recommendation at the end of the diagnosis?
9	Do you know foods that avoid suffering from iron deficiency?
10	Do you know recipes to combat anemia?
11	Would you like recipes to fight anemia included?

- On the other, hand 36 people would like to log in quickly, using their Gmail account for immediate access.
- 30 of them want to carry out a previous survey regarding, their current health status in order to have a more assertive diagnosis.
- In addition, 28 respondents who were aware of artificial intelligence, considered it necessary to incorporate it for a better diagnosis, offering them greater reliability of the results.
- 35 of the participants would like to be able to graphically, display the results for better understanding.
- On the other hand, respondents were asked if they would like to see the incorporation of advice and recommendations after a diagnosis, with the result that all respondents were in favor of incorporating those functionalities. Likewise, 38 of them would like to include recipes to combat anemia according to the results obtained in their diagnosis.

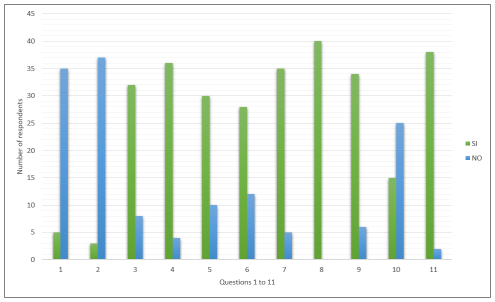


Fig. 13. Survey Results

C. Analysis of each Sprint by Burndown Chart

The use of Burndown chart, also called burndown diagram, represents graphically the effort and the follow-up of the tasks, during the established time of each sprint [26]. Generating the data graphically, the x-axis of the Burndown chart shows the time set per sprint and the y-axis shows the history points. Moreover, the blue line represents the expected values and the orange line reflects the progress of the history points, the further away from the blue line the less incremental value has been added. In other words, user stories are not completed or advanced.

1) First Sprint: According to the results obtained, the first sprint is made up of 4 story points with a set time of 4 weeks. As can be seen in Fig. 14, during the first few days, the points were successfully advanced, but between the end of August and the beginning of September, this changed, delaying the proposed advances.

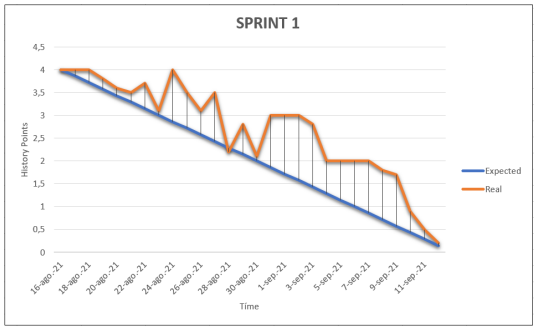


Fig. 14. Analysis of the First Sprint.

2) Second Sprint: It is made up of 26 story points, with a time frame of 6 weeks. During the development of the sprint we can observe that the advances in some days were completed on time and there were even early story deliverables as shown in Fig. 15.

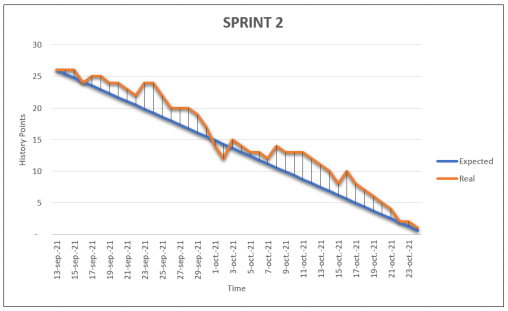


Fig. 15. Second Sprint Analysis.

3) Third Sprint: Composed of 8 story points, with a set time of 4 weeks, the development of the sprint was very good with all story points being completed starting with an early delivery and then with minimal delay with progress, but manageable as shown in Fig. 16.

4) Fourth Sprint: It has 7 story points, with a planned time of 2 weeks. Likewise, as in Sprint 3 the results were very parallel to the expected values demonstrating, a great development and performance of the sprint, culminating with the development of the web page for anemia diagnosis as shown in Fig. 17.

D. Analysis of Children with Anemia in Peru

1) Analysis in Peruvian Children: As shown in Fig. 18, the departments with the highest prevalence of anemia in children are the following: Puno, Loreto, Pasco, Huancavelica

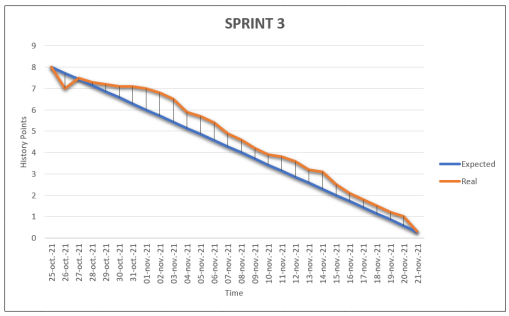


Fig. 16. Third Sprint Analysis.

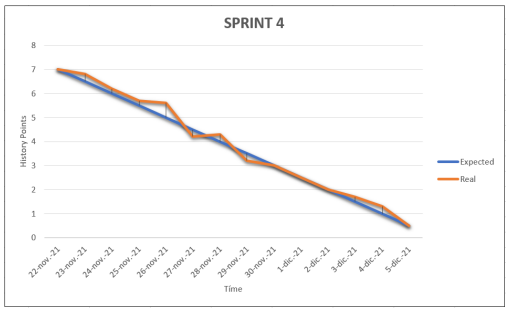


Fig. 17. Fourth Sprint Analysis.

and Ucayali. Thanks to this information, it was concluded that the first test carried out by the web system, be performed in these departments, since they are the most affected by this disease, and therefore, they should be attended as soon as possible. On the other hand, the departments with the lowest prevalence are: Lambayeque, Moquegua, La Libertad, Tacna, Lima and Cajamarca. Similarly, the Web System will be beneficial. However, it will not have the same impact as in the most affected departments. In Peru, practically 50% of the child population does not have a good control of their diet and this is reflected in the following graph. In short, our web system will present a solution to this problem and reduce these results to improve the quality of life of our Peruvian brothers and sisters.

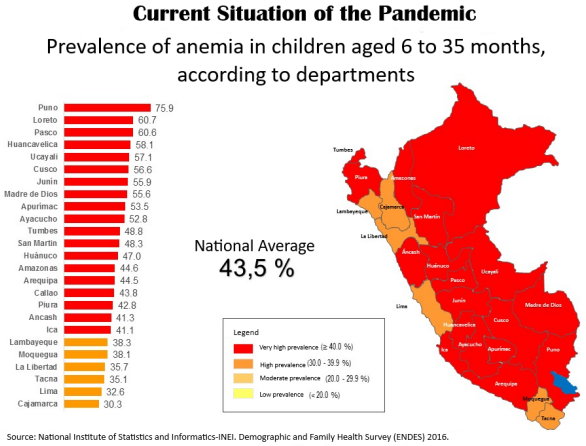


Fig. 18. Situation of Children with Anemia in Peru [27].

2) Analysis of Children in Callao: The district of Callao, has a high prevalence rate of anemia in children under 3 years of age. children under 3 years of age, which is why an in-depth analysis was carried out. As Fig. 19, the blue line represents the infection rate in Peru and the orange line represents Callao. and the orange line represents Callao. Thanks to the following analysis, it is known that, in 2021, only 2 out of every 10 children under 3 years of age will be affected by anemia, which is great news, given that years of This is great news, since years ago this result was alarmingly higher.

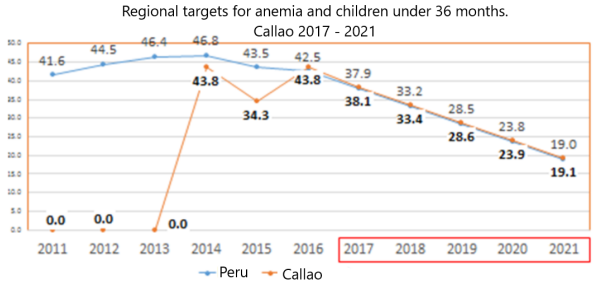


Fig. 19. Situation of Children with Anemia in Callao [27].

VI. Conclusion and Future Work

In conclusion, in this research work it was possible to design a web system to detect anemia in children, with the help of the design tool Balsamiq and applying the Scrum methodology. Thanks to Balsamiq it was possible to present the creative idea and the functionalities requested by the users through eye-catching and intuitive graphics.

With respect to the methodology, the Scrum framework was of vital importance, since it allowed a sequential development with high value, thus improving the quality of the product. In addition, the various charts and graphs that Scrum presents are key to presenting information clearly and accurately.

In the future, it is expected to complement this research by developing the web system, applying all the steps of this article to make this web system a reality. The development of the tool will be useful for people who do not have an adequate diet or feel they need an anemic diagnosis. This is why the implementation of the web system is expected as soon as possible, in order to contribute to the health sector and help people who require a diagnosis.

Acknowledgment

Recognize the University of Sciences and Humanities, and its research institute, for their support in research.

References

[1] R. Prasanth, "Prevalence of Anemia in both Developing and Developed Countries around the World," *World Journal of Anemia*, vol. 1, no. 2, pp. 40–43, 2017.

[2] M. Cabanillas-Carbonell, H. Ñahuiña-Balbuena, J. Soto-Justiniano, and O. Casazola-Cruz, "Mobile application for the monitoring and control of the diet in people with anemia," in *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, 2020, pp. 1–4.

[3] E. Mantadakis, E. Chatzimichael, and P. Zikidou, "Iron deficiency anemia in children residing in high and low-income countries: risk

factors, prevention, diagnosis and therapy," *Mediterranean Journal of Hematology and Infectious Diseases*, vol. 12, no. 1, 2020.

[4] N. G. Onyeneho, B. C. Ozumba, and S. Subramanian, "Determinants of childhood anemia in india," *Scientific reports*, vol. 9, no. 1, pp. 1–7, 2019.

[5] A. Hernández-Vásquez, D. Azañedo, D. A. Antiporta, and S. Cortés, "Spatial analysis of gestational anemia in Peru, 2015," *Revista Peruana de Medicina Experimental y Salud Publica*, vol. 34, no. 1, pp. 43–51, 2015.

[6] A. Delgado and N. P. Ccancce, "Where is the highest rate of children with anemia in Peru? An answer using grey systems," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 313–317, 2020.

[7] J. Velásquez, Y. Rodríguez, M. Gonzáles, L. Astete, J. Loyola, W. Vigo, and Á. Rosas, "Factores asociados con la anemia en niños menores de tres años en Perú: análisis de la Encuesta Demográfica y de Salud Familiar, 2007-2013.revistaBiomédica[revista en inetrnet]2016[acceso 30 de septiembre];36;[220-9].http://dx.doi.org/10.7705/biomedica.v3," *Biomédica*, vol. 36, pp. 220–229, 2016.

[8] D. I. Garrido-Salazar, S. M. Garrido-Salazar, and G. Vivas-Armas, "Anemia frequency in children living at Andean high altitude in Ecuador, Peru, and Bolivia," *Acta Pediatrica de Mexico*, vol. 40, no. 6, pp. 305–317, 2019.

[9] R. Ul Islam, M. S. Hossain, and K. Andersson, "A deep learning inspired belief rule-based expert system," *IEEE Access*, vol. 8, pp. 190 637–190 651, 2020.

[10] R. P. Córdova Cárdenas, "Diseño e implementación de una aplicación móvil basada en android para la evaluación de anemia ferropénica en personas de acuerdo al nivel de hemoglobina," Ph.D. dissertation, 2018. [Online]. Available: http://dspace.uazuay.edu.ec/handle/datos/8173

[11] J. Jayakody, E. Edirisinghe, and S. Lokuliyana, "HemoSmart: A Non-Invasive Device and Mobile App for Anemia Detection," *Cognitive Engineering for Next Generation Computing*, pp. 93–119, 2021.

[12] R. Bendezu and A. Ysla, "App de recomendaciones alimentarias para reducir la mala alimentación en casos de anemia en niños del colegio "apoóstol de Punchauca"," Tesis de pregrado, Universidad San Martin de Porres, 2020. [Online]. Available: https://hdl.handle.net/20.500.12727/6824

[13] R. G. Mannino, "A noninvasive, image-based smartphone app for diagnosing anemia," no. May, 2018.

[14] A. Hafeel, H. S. Fernando, M. Pravienth, S. Lokuliyana, N. Kayanthan, and A. Jayakody, "IoT Device to Detect Anemia: A Non-Invasive Approach with Multiple Inputs," *2019 International Conference on Advancements in Computing, ICAC 2019*, pp. 392–397, 2019.

[15] W. Mahmood, N. Usmani, M. Ali, and S. Farooqui, "Benefits to organizations after migrating to Scrum," *Proceedings of the 29th International Business Information Management Association Conference - Education Excellence and Innovation Management through Vision 2020: From Regional Development Sustainability to Global Economic Growth*, no. May, pp. 3815–3828, 2017.

[16] M. Kumar and R. Dwivedi, "Applicability of Scrum Methods in Software Development Process," *SSRN Electronic Journal*, 2020.

[17] 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

[18] 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

[19] M. Morandini, T. A. Coleti, E. Oliveira, and P. L. P. Corrêa, "Considerations about the efficiency and sufficiency of the utilization of the Scrum methodology: A survey for analyzing results for development teams," *Computer Science Review*, vol. 39, p. 100314, feb 2021.

[20] B. Singh, "Comparative Study and Analysis of Scrum and Lean Methodology," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, no. 3, pp. 3441–3448, 2018.

[21] P. C. Van Oorschot, "Toward Unseating the Unsafe C Programming Language," *IEEE Security and Privacy*, vol. 19, no. 2, pp. 4–6, 2021.

[22] J. Yang and Z. Zhao, "Development and implementation of computer assisted instruction system in physical education based on ASP.NET Technology," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 13, pp. 145–156, 2019.

[23] A. Fedorov and A. Lukyanchikov, *Recent Achievements and Prospects of Innovations and Technologies*, 2019.

[24] Mubassiran, "Jurnal Ilmiah Manajemen Informatika," *Ilmiah Manajemen Informatika*, vol. 12, no. 2, pp. 1–70, 2020.

[25] 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

[26] L. R. Tuanama, J. Q. Gutarra, and L. Andrade-Arenas, "Design of a mobile application for the automation of the census process in peru," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111184

[27] INEI, "La anemia en el perú," 2018. [Online]. Available: https://proyectos.inei.gob.pe/endes/resultados.asp

Relationship between Stress and Academic Performance: An Analysis in Virtual Mode

Janet Corzo Zavaleta[1]
Departamento de Estudios Generales
Universidad de Ciencias y Humanidades
Lima, Perú

Roberto Yon Alva[2]
Departamento de Estudios Generales
Universidad de Ciencias y Humanidades
Lima, Perú

Samuel Vargas Vargas[3]
Departamento de Estudios Generales
Universidad de Ciencias y Humanidades
Lima, Perú

Eleazar Flores Medina[4]
Departamento de Estudios Generales
Universidad de Ciencias y Humanidades
Lima, Perú

Yrma Principe Somoza[5]
Departamento de Estudios Generales
Universidad de Ciencias y Humanidades
Lima, Perú

Laberiano Andrade-Arenas[6]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—This research work analyzes the relationship between stress and academic performance of engineering students at the University of Sciences and Humanities, in Peru, in the context of the pandemic. During this period, classes at the university were held virtually; and the difficulties that students present to carry out their classes were identified, such as lack of connectivity, family and financial problems and anxiety. The objective of the research is to analyze the relationship between stress and academic performance of engineering students. The work is part of a mixed approach, and data collection was carried out through interviews and surveys of engineering students based on the two variables identified: stress and academic performance. It began with a descriptive analysis, then moving on to inferential analysis to perform the hypothesis test. The SPSS was used for the reliability analysis using Cronbach's alpha with 0.84 as a result and validation by expert judgment with 84.5% acceptance. It was obtained as a result that between the Stress variable based on its three dimensions and academic performance there is not relationship since it was obtained that its P-value is greater than 0,05; it is concluded that stress is not only academic but also should consider others as labor, social. In addition, the positive stress that drives academic performance emerged. The beneficiaries with the research are students, and the university. It is concluded that there is no relationship between stress and academic performance.**

***Keywords*—Academic performance; anxiety; stress; teaching-learning; virtual mode**

I. Introduction

The health crisis that is being experienced around the world, due to the COVID-19 pandemic, has led various states to partially or totally close schools and universities to prevent the spread of the virus. In this context, the new educational challenge of virtual or distance education is imposed in almost all the countries of the world, bringing with it effects that not only represent a massive impact on productivity; but also, for social life and student learning [1]. Worldwide, it can be seen that the effects of social isolation and virtual education affect all students emotionally. According to the various studies carried out on this subject, it is evident that the majority of students have shown an increase in their levels of anxiety and stress, due to the drastic change around learning and future demands in the professional field [2]; to concern for their academic progress, performance, and their adaptation to distance learning. That is why the need for greater psychological counseling of students is emphasized [3].

In the new context of virtual education, the University of Sciences and Humanities (UCH) has implemented sundry measures that guarantee the achievement of the objectives proposed in the study plan. For this, it has implemented divers pedagogical resources available to teachers and students through the virtual platform that show good results; However, it has been shown that many of our students have expressed feelings of anxiety and stress that, in some cases, have led them to the determination to abandon their studies. This has been evidenced in the increase in the number of students who drop out; and although the university does not have reliable studies on this subject, teachers perceive this problem every day. The research work has as a unit of analysis, a student of the Faculty of Science and Engineering. In this Faculty, students manifest difficulties with connectivity, the internet, family problems due to the pandemic, as well as psychological problems. In this sense, it investigates the different manifestations of stress that may be affecting students in their academic performance.

Besides, the importance of the study is to be able to analyze the relationship between stress and academic performance, of our students from the Faculty of Sciences and Engineering, determining the levels of stress experienced by students and academic performance, in the virtual modality. For this, support mechanisms can be established for our students with comprehensive work with teachers and the educational community in general, based on the comprehensive training proposal. In this way, students will benefit, to improve their academic performance, and cope with stress through proposals that will come out of the research. Also, the university can avoid dropping out, which can be one of the causes of stress. Faced with the problem, formulate the following research question: What is the relationship between stress and academic performance of the students of the Faculty of Sciences and Engineering in times of pandemic?

The objective of the research is to analyze the relationship between stress and academic performance, of students of the Faculty of Sciences and Engineering of the UCH, in the virtual

modality, in times of pandemic. For this, the degree of stress of university students is described in the virtual mode in times of pandemic; in addition, academic performance and its effects related to stress are analyzed.

The investigation is made up of six sections. In Section II, In the literature review; in Section III, the methodology was carried out; also in Section IV are the results; in Section V, the discussion, ending in Section VI with the conclusions and future work.

II. Literature Review

A. Theoretical Basis

1) Stress: As for the definition of stress, it is quite diverse and, until now, it has focused on a psychological approach, conceiving it as a disease that is associated with certain alterations of the organism [4]. However, stress can also be positive to face certain difficulties. The author [5] distinguished between pleasant stress, which he called stress; and stress or distress, when it comes to unpleasant stress. When it comes to stress, in general, it refers to anguish, but not everyone can relate to pleasant situations or events that have caused us stress: weddings, births, promotions, receiving awards, meeting old friends and many more. In that sense, any change, positive or negative, requires a response from our bodies to adapt and bring us back to our relatively peaceful state.

Stress is also seen as a state of imbalance between demands, be they from within or from external sources, and our perceived capacities to meet those demands. This is most acutely experienced when the expectation of meeting demands is not met, which will have consequences for the person.

In the academic field, the study of stress focuses mainly on a positivist stance, which leaves aside a phenomenological and qualitative look; that is, a holistic pedagogical approach that provides a comprehensive view of the phenomenon is not assumed. In order [4] to conceive the definition of stress in the academic field, it is necessary to take into account: both the psychobiological and the psychosocial components. According to what was stated by [6], academic stress is the process of change in the components of the teaching and learning process, through a set of individual and institutional adaptive mechanisms, a product of the overwhelming and demanded demand in the teaching and learning experiences that are developed in higher education institutions, with the purpose of maintaining a steady state in the educational process.

2) Academic Performance: Academic performance is a value associated with the quality of higher education institutions; since, it allows to have a better panorama of the educational reality; however, there are various criteria when measuring it. The concept of academic performance [7], must consider the diverse and complex factors that are revealed in learning; in this sense, academic performance is considered when there is a contribution in the achievement of student results, through evaluation. Likewise, it is pointed out that it is a multicausal phenomenon, which makes its assessment complex. Among the factors that intervene, they point to the components of internal and external order to the individual, which can be of a social, cognitive and emotional order.

In the case of stress and its relationship with academic development, studies indicate stress rates in university students, which are observed with greater preponderance in the first years of the degree and in evaluation periods. In addition, it has been shown that university students can present various stressful situations that affect their performance [8]. Regarding the relationship between stress levels and academic performance, there are numerous studies, such as those of [9] who carried out a study with 162 university students from the University of Balarias Islands (UIB), for which They used the Systemic Cognitive Inventory (SISCO) of academic stress as a measurement instrument. Among the main findings, it was reported that teacher evaluations and the jobs that teachers ask for are the main cause of stress in students; however, no direct relationship was found between stress and academic performance.

B. Related Work

This research work seeks to establish the relationship between stress and academic performance. For this, a search of the works related to this topic has been carried out, based on divers methodologies and results, which allows a broader overview of the problem studied.

In recent years, numerous studies have been conducted with various methodologies. For example [10], they identified the causes of stress and its effects on the academic performance of Malaysian Higher Education students. In this study, it is determined that life and interactions on campus are related to stress and academic performance. It is for this reason that the researchers suggest that university authorities guarantee suitable environments for learning and advisory centers [11].

The study carried out by [12] on stress in Astana Medical University, satisfaction and academic performance and online learning during the pandemic, carried out in a nursing school, reveals that for students learning during this period, in virtual mode, was stressful; and the perception of satisfaction was low and moderate. Regarding the academic performance of the students, they indicate that they were affected by the pandemic. Likewise, a significant and inverse correlation is shown between academic performance and stress during online learning [13].

On the other hand, the study carried out with students from the University of Nairobi, revealed that stress was related to academic performance; therefore, the majority of students stated that they had moderate to high levels of stress [14]. In the study carried out on the relationship between stress and academic performance, a high percentage of stressed students was obtained: about 65%. Performance was statistically significant at ages ranging from 19 to 23 years. Besides, the linear regression was inverse; that is, the higher the stress, the lower the academic performance. To reduce this problem, it is recommended to carry out programs with academic workshops [15].

Although it has been proven in various studies, to establish the relationship between stress and academic achievement, there are other factors to take into account: physical and psychological aspects. In the study carried out by [16], on the effects of COVID-19 on the mental health of university students in the United States, an increase in stress and anxiety

levels was reported. This increase was associated with multiple factors, such as fear and concern for their health, difficulty concentrating, disorders with sleep patterns, decreased social interactions and greater concern about their grades.

In the study carried out at the University of Pakistan, it is observed that the body mass index, stress and academic performance in students had a negative correlation of stress. Likewise, in order to calculate the process, the global system for data analysis (SPSS) was used [17]. According to the study carried out with Albanian students, on the averages of the students and their relationship with stress, the direct relationship between optimal academic performance and student tranquility was evidenced. Also, they used already elaborated questionnaires based on the stress scale and on attachment. The authors suggest that more in-depth research should be carried out, since the study was exploratory in its first phase [11].

The effects of the pandemic may increase dropout rates in educational institutions. Distance education, although not new, has been introduced on a large scale by trial and error to alleviate these effects; however, it has evidenced problems in learning and levels of understanding of students. In this sense, the recovery of this sector is essential to avoid a major generational catastrophe [18]; that is why higher education institutions have been forced to implement alternative methods to guarantee the continuity of studies and minimize the gap in this new context. For this purpose, the use of virtual platforms has been arranged with the objective that the actors of the educational process interact with each other, mainly teachers and students, and achieve the established curricular learning outcomes. This has meant a challenge to become familiar with the new teaching-learning methods; unfortunately many students state that they are not satisfied with this type of learning, increasing the rate of anxiety and depression [3].

It should be noted that the stress experienced by students is generated by academic and non-academic aspects such as: environmental, sociocultural and psychological factors. This means that students feel the pressure to perform well in order to meet the expectations of their parents and thus obtain benefits at work level in the future [19]. Specialists point out that academic stress can be decreased over time, and only a small portion of students will develop chronic disorders, if they do not receive medical help on time; this implies that the severity of the initial reactions to stress can predict their continuity and treatment [20].

It cannot be denied that stress influences academic performance; however, this can be reduced. On the one hand, through the decisions made by teachers and institutions, to facilitate the transition to the virtual modality, influencing the reduction of class hours, qualification requirements and support in evaluations. On the other hand, students also put in place various mechanisms to deal with it. They can be adaptive behaviors: acceptance and adaptive coping, or maladaptive ones like denial and disengagement [21]. On the other hand, students also put in place various mechanisms to deal with it. They can be adaptive behaviors: acceptance and adaptive coping, or maladaptive ones like denial and disengagement [2].

Likewise, students in this period have been forced, in many cases, to work to pay for their studies, which means a higher level of concern that affects their academic performance. In recent years, stress has been increasing moderately every year. According to the research between stress and performance, can observe that those who are exposed to work and academic hours their stress level increases considerably. Likewise, it has been shown that students who do not work their stress level is moderate. In summary, the authors found in the research that stress has different forms of manifestations; as well as the presence of positive stress is evidenced; the relationship between stress and academic performance was also found; however, more innovation is lacking in the uses of instruments to measure stress. In addition, positive stress must be related to academic performance, with an in-depth study to know the consequences of it.

III. Methodology

A. Methodological Design

The research work has a qualitative-quantitative approach, that is, mixed. This approach is often presented as two paradigms that complement each other to carry out a deeper investigation, with the perception and interpretation of the research data [22]. The research design is non-experimental, which is characterized by not manipulating the variable in the research [23]. Likewise, the research is descriptive, correlational, causal and cross-sectional in scope.

B. Development of the Methodology

The following steps were performed:

1) Modeling: A hypothetical model was carried out, between the variable stress and academic performance. According to Fig. 1, the relationships between them are observed, as an assumption, where the positive sign means that there is a direct relationship between the variables; and the negative sign, that there is an inverse relationship. That allows to be able to contrast, if there really is a correlation and if there is, what type it is, inverse or direct.

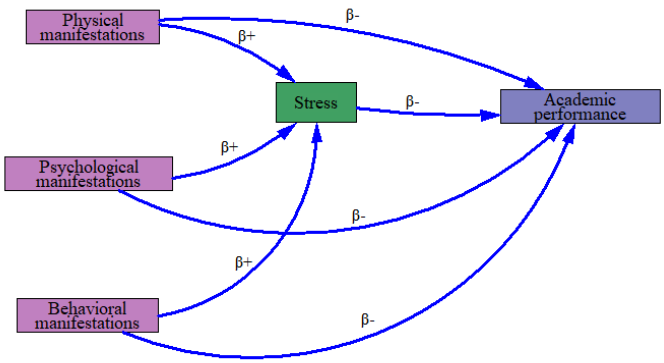


Fig. 1. Hypothetical Model.

2) Interview: 8 students from the 1st to 4th cycle of the Faculty of Engineering were interviewed. For this, the following steps were taken into account (see Fig. 2).

- Elaboration of the questions.

- Selection of students.
- Recording the interview.
- Analysis of the interview.

A semi-structured interview with 6 questions was conducted, as shown in Table I.

TABLE I. INTERVIEW QUESTIONS GUIDE

Questions	
1	During your academic preparation at the university, have you felt stress? If you have felt stress, how do you think it affects your studies?
2	When you are stressed for academic reasons, do you have health problems such as pain, insomnia, indigestion or others? If yes, can you specify which ones and how they are manifested?
3	In periods of stress, do you usually have problems on an emotional level, that is, in your mood, your feelings, lack of concentration or others? If yes, can you specify which ones and how they are manifested?
4	During the last semester, you were seized with feelings of depression and / or sadness In what newspapers specifically? How are you coping?
5	How do you perceive stress affects your participation in class?
6	Do you have problems of aggressiveness, conflicts and of social relationships with your colleagues? Why?

3) Survey: For this study, a survey has been prepared, based on the SISCO questionnaire on stress. In Peru, in recent years, several studies have been carried out on stress in students based on the SISCO SV-21 inventory. Among these works, have [24] who analyzed the psychometric evidence of the SISCO SV-21 inventory in 560 students from public and private universities in Metropolitan Lima; reaching the conclusion that said inventory gathers the evidence of validity and reliability for its use [25].

In the present work, the survey was applied using the Google form questionnaire as an instrument. The student population under study is 150 students from the first to the fourth cycle. Through a simple random sample, 109 students were obtained from the morning and night shift.

In addition, the academic performance variable was analyzed by placing intervals to the average grades from 0 to 20, considering the levels very low, low, regular, and high and very high. This information was obtained from academic records; This consists of their average grades from the first stage of the 2021-2 semester (see Table II).

TABLE II. AVERAGE NOTE RANGE

Scale	Academic performance	Average note range
1	Very low	[0;5[
2	Low	[5;11[
3	Regular	[11;15[
4	High	[15;18[
5	Very high	[18;20]

IV. RESULTS

A. Instrument Reliability and Validation

- Reliability with Cronbach's Alpha
  The instrument's reliability result must be greater than 0.70 for it to be approved (see Table III) [26]; since the value must range between 0 and 1 [27] [28]. The reliability of the instrument was determined with 10% of the population. For this, the data were collected and processed with the SPSS program. The result was 0.84. The analysis of questions 1 to 17 of Table IV shows that if a question is deleted from the survey, the Cronbach's alpha varies between 0.82 and 0.85. That is, it is within the mean of Cronbach's alpha, which is 0.84. From the penultimate column of Table IV, its values must be greater than 0.200. It is observed that questions 2 and 5 do not comply; for this the modification was made.

TABLE III. INTERVAL FOR RELIABILITY

Interval	Reliability assessment
[0 ; 0,5[	Unacceptable
[0,5 ; 0,6[	Poor
[0,6 ; 0,7[	Weak
[0,7 ; 0,8[	Acceptable
[0,8 ; 0,9[	Well
[0,9 ; 1]	Excellent

TABLE IV. ALFA DE CRONBACH

Questions	Mean scale if item has been suppressed	Scale variance if item has been suppressed	Total correlation of corrected elements	Cronbach's alpha if the item has been suppressed
1	32,00	57,78	0,31	0,83
2	33,40	61,52	0,09	0,84
3	31,75	56,30	0,45	0,83
4	32,15	53,61	0,50	0,83
5	33,15	61,61	0,04	0,85
6	32,25	51,67	0,75	0,81
7	33,35	56,13	0,54	0,83
8	32,75	57,57	0,30	0,84
9	32,50	53,32	0,60	0,82
10	32,15	53,50	0,73	0,82
11	32,65	55,82	0,48	0,83
12	32,65	55,29	0,33	0,84
13	33,30	57,38	0,42	0,83
14	33,55	58,37	0,62	0,83
15	32,80	53,75	0,61	0,82
16	33,60	58,67	0,63	0,83
17	32,80	52,91	0,48	0,83

- Validation by Expert Judgment
  Expert judgment was carried out to validate the contents of the instrument [29], (see Table V). The profile of the experts that was considered for the validation was: specialists in psychology and education who work at the university. The validation by expert judgment was taken as a reference for approval, greater than 75%. As a result, 84.7% acceptance was obtained as the average of all the questions evaluated. For this, the following criteria were taken into account: clarity, updating, coherence, consistency, objectivity, content and sufficiency.

B. Qualitative Analysis

According to the interviews carried out with the students from the first to the fourth cycle of the Faculty of Sciences and Engineering, the following categories could be constructed.

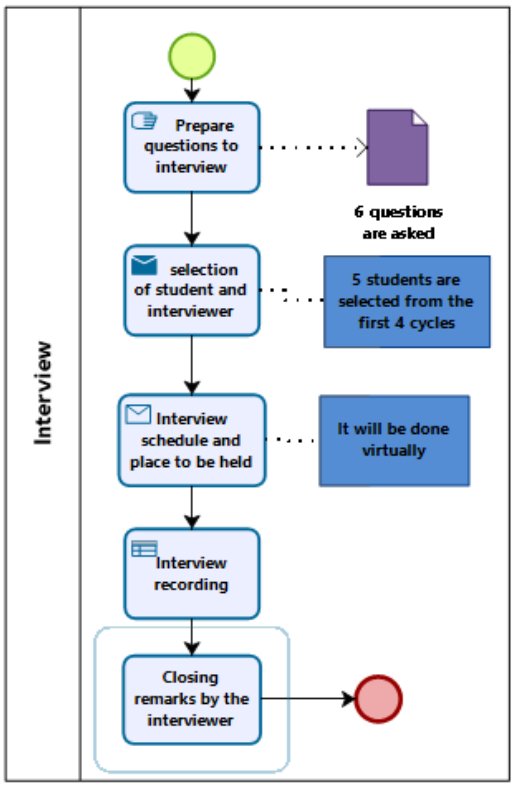


Fig. 2. Steps of an Interview.

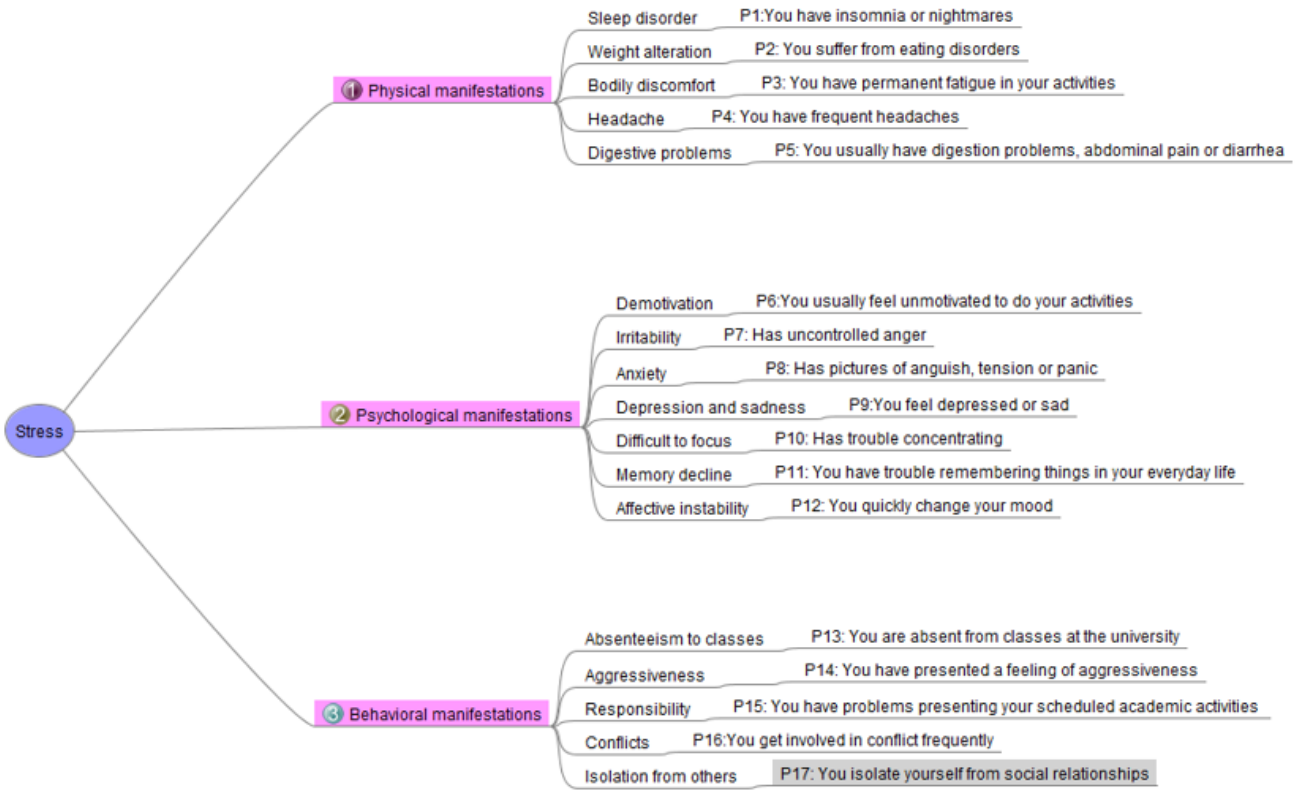


Fig. 3. Dimensions of the Stress Variable.

TABLE V. EXPERT JUDGMENT

Question	Expert1	Expert2	Expert3	Percent
1	80%	85%	90%	85%
2	75%	80%	85%	80%
3	70%	90%	80%	80%
4	90%	90%	90%	90%
5	80%	80%	80%	80%
6	90%	90%	90%	90%
7	70%	90%	80%	80%
8	90%	90%	90%	90%
9	80%	85%	90%	85%
10	75%	80%	85%	80%
11	70%	90%	80%	80%
12	90%	90%	90%	90%
13	80%	80%	80%	80%
14	90%	90%	90%	90%
15	70%	90%	80%	80%
16	90%	90%	90%	90%
17	90%	90%	90%	90%

1) Stress and Academic Performance Relation: All students report feeling stress during their academic preparation; but they point out that this occurs in specific periods when they approach the date of the evaluations; or in some courses they don't understand. At this point they coincide in cases of students who work and study, they more accurately recognize the symptoms of stress. Likewise, students recognize that stress affects their studies, as it causes them to lack concentration and lack of motivation. As stated by [30], academic stress can reduce academic performance, decrease motivation and increase the risk of dropping out of school.

E1: *Yes, definitely especially in these times of the pandemic was when I decided to study and the moments of greatest stress ehh I think are the evaluations and when jobs are put together ... duties that I have to perform at work sometimes I am short on the time ... I think that in part yes because ... I am very stressed or ... I am not sometimes like I lose motivation.*

2) Physical Manifestations: In the student's discourse, it stands out that most have presented physical symptoms due to stress. The main physical manifestations are headache, insomnia, palpitation of the eyes and digestive problems. These findings coincide with that reported by Silva et al. [31], where it was found that the physical manifestations of stress are headaches or migraines and an increase or decrease in food consumption. These symptoms occur in short intervals and can usually be managed without the need for medical treatment.

E3: *Yes, when I am under stress I have headaches, not always ... headaches or insomnia. The headache lasts for hours, not long, but regular. I mostly control it when I start to calm down*

3) Psychological Manifestations: All students present emotional manifestations as a consequence of stress. They usually present anxiety, reluctance, lack of motivation, sadness and anger [32]. These symptoms appear in short periods and may be associated with comprehension problems; or to the development of various activities at an economic and academic level. It should be noted that anxiety at the academic level may be due to the courses that they consider more difficult or the expectation of their grades, since they consider that they are assigned a low grade and that it is not related to the effort made.

E4: *Some anxiety. There comes the issue of anxiety sometimes there are days that I have to stay up late to be able to review or the same issue of knowing what I am going to do the next day as I am going to leave to be able to do the different activities generates stress, generates anxiety more than anything, well it is linked to stress ... and my mind is going elsewhere.*

4) Behavioral Manifestations : Most students participate in various classroom activities, depending on the courses. While some interviewees indicate that they prefer not to participate, as they consider that it is not necessary because they know the issues. Given this, can indicate that there is not direct relationship between stress and participation in activities developed in class. These results are complemented by those presented by Encina Meza et.al. [33] who points out that in stressful situations, the most frequent behavioral symptoms are lack of concentration, decay, isolation and the tendency to feel more conflictive and with a tendency to argue. In this sense, can point out that in terms of aggressiveness, conflict and social relationship problems between students, a direct relationship with stress is not observed. However, some students emphasize that group work can generate stress, due to the fact that not all participate actively and certain communication problems.

E4: *...perhaps that out there there are not all cases of a student who perhaps are the ones who want to do everything according to their interpretation and maybe out there they are a little susceptible to changes, but after I have crashed, I have felt rather like Many of them are also young, as they look for some support from people who are of age, I see that sometimes they lean on me in some things.*

C. Emerging Categories

1) Stressors: It was identified that the most frequent stimuli that generate stress are evaluations, overload of activities and work; added to this is the limited time to perform them. On the other hand, it is observed that students present greater stress in the periods close to the evaluations; since they feel pressured to pass the courses, and their expectations may be offset by their results. These results correspond to those presented by Cordova et al. [34]. Who point out that within the demands of the academic environment, the ones that generate the most stress for the student are overload of tasks, type of work and evaluations. Besides, several of the students indicate that they are not organized properly; therefore, they are joined by various jobs and evaluations that they feel they will not be able to fulfill.

E1: *but the previous cycle if from the midterm exam and until the final exam I had these symptoms.*

Likewise, we observe that the courses that generate the greatest stress are those related to formal sciences: mathematics, physics and those of the specialty of engineering.

E6: *mostly because of the area that was mathematics and that was systems. Then it got a bit complicated and I kind of left your mind blank.*

2) Coping Strategies: Within the interviews, it has been observed that students test various stress coping strategies. These strategies focus on direct actions after personal evaluation and analysis, before seeking support from external elements.

This coincides with the findings obtained by Boullosa [35], who points out that the techniques most used by students to overcome stress are to strive to achieve their academic goals by concentrating on solving the problem and looking for spaces for relaxation. Most of them indicate that they handle the situation personally, without resorting to external help, since they consider that they are prepared to cope with this situation. Some techniques they use are trying to calm down or think about positive aspects of their preparation, doing an activity that they like or resorting to a metacognitive strategy to evaluate their learning process.

E1: *First I try to calm down, I try to see the positive side of that of the situation and for another opportunity to do it much better.*

However, a small number of students indicate that they have resorted to interpersonal relationships as a coping strategy; that is, to family or some belief. These findings are related to the study conducted by Barraza [36], who points out that the strategies most used by students are the elaboration of a plan, execution of tasks or activities, and assuming these stressful situations with humor; while the least used are praising their progress or religiosity.

E3: *Not the other way around, they try to support me (my family) but at the moment I am with all the stress ... and I get defensive but it can be controlled.*

E7: *ehhh, I do it by myself; bone also raised a small prayer also as it is called, I say to God give me strength, please ...*

3) Positive Stress: According to Díaz & Fierro [37], positive stress or eustress is a positive physiological and emotional response to academic stressors, which are perceived by the student as a challenge or challenge and not as a threat. In this sense, it is pointed out that the resources available to the student allow them to respond to academic demands and strengthen their skills. Besides, some students point out that stress can be positive for the development of their academic activities, since it motivates them and provides them with that "pressure" necessary to achieve their academic goals.

E6: *no, because of ... with this pressure I feel that I have to study more and about going out so that next week I have good results.*

D. Descriptive Analysis

The survey is made up of two sections: in the first, the sociodemographic data, such as age, sex, cycle, shift; and in the second section, the questions of the stress variable that have 3 dimensions with 17 questions (see Fig. 3). These were built with the Lickert scale, where 1 is never; 2 hardly ever; 3 sometimes; 4 usually; y 5 always. It is observed in Fig. 4, the gender and age of the student with its frequency, being the male gender the one that predominates over the female; likewise, the most frequent age is 21 years and 1 student over 50 years of age. Regarding the female students surveyed, 9 are 20 years old more frequently and one student is 40 years old. An asymmetric bell with accumulation between 17 and 21 years old is evidenced in both graphs, both for male and female, with greater frequency and the others are scattered. Likewise, in Fig. 6 it is observed that the average of the male and female gender are close to 2,10. The one with the greatest amplitude of length of the stress average is in females, with a range of 1.90 to 2,28 and in the male gender, the length of their stress averages is 2 to 2,15.

In Fig. 5 the dimensions of the stress variable were analyzed, where the three obtained a median close to 2, with respect to the average stress score. Además en la Fig. 7, it is observed that 16,85% have high stress; analyzing the moderate percentage, it is wide with 57,30% where the majority of students remain neither very high nor very low. Relating to the percentage of academic performance, 45,96% is high and what is shown in Table VI the percentages of the stress dimensions is moderate, concluding that stress and performance in both is on the moderate average. Besides, the analysis of academic performance was carried out by means of the average grades of the students from the first to the fourth cycle of the students of the Faculty of Engineering. The mean of 3,90 indicates that it is above the average, it means that the students have a high average close to 46%, the low average is little significance, since it is approximately 5% (ver Fig. 8).

TABLE VI. PERCENTAGE OF STRESS

Dimensions	Low	Moderate	High
Physical manifestations	30,34%	50,56%	19,10%
Psychological manifestations	25,84%	55,06%	19,10%
Behavioral manifestations	25,84%	51,69%	22,47%

E. Comparison Analysis of Gender Means by Dimensions

Through statistical analysis, the Mann-Whitney U test was performed for sample variables, which corresponds to the male and female gender, with respect to the dimensions of the investigation, (see Table VII). It is observed that the P-value is 0,140 of the first dimension. By theory it is known that if the P-value is less than 0,05 there is a difference of the means in the comparison and if it is greater than 0,05 there is no difference between the averages.In this sense, in the 3 dimensions, the P-value is greater than 0,05; therefore, it is concluded that in the three dimensions the means are not different since their P-value for the 3 is greater than 0,05, with a margin of error above 5%.

F. Relationship between Variables

The relationship between all the variables of the Stress dimensions was made, using Spearman since it is non-parametric. From Table VIII it is observed ** means that the correlation is significant at the 0,01 level (bilateral); where is interpreted as 1% margin of error. In the Table IX, the scale of the correlation coefficient is observed; comparing with the results obtained, the dimension of psychological manifestations and with physical manifestations is 0,512 **; that is, moderate positive correlation. Likewise, the dimension behavioral manifestations with physical manifestations is obtained 0,647 ** which is in the range of correlation, moderate positive and finally the dimension behavioral manifestations with psychological manifestations is 0,816 **, being in the range of high positive correlation. It is observed in the Table IX highlighted in blue, the range where the results obtained are found. On the other hand, the relationship between the academic performance variable and the 3 dimensions of stress is observed where the result is very low. In Table VIII it is

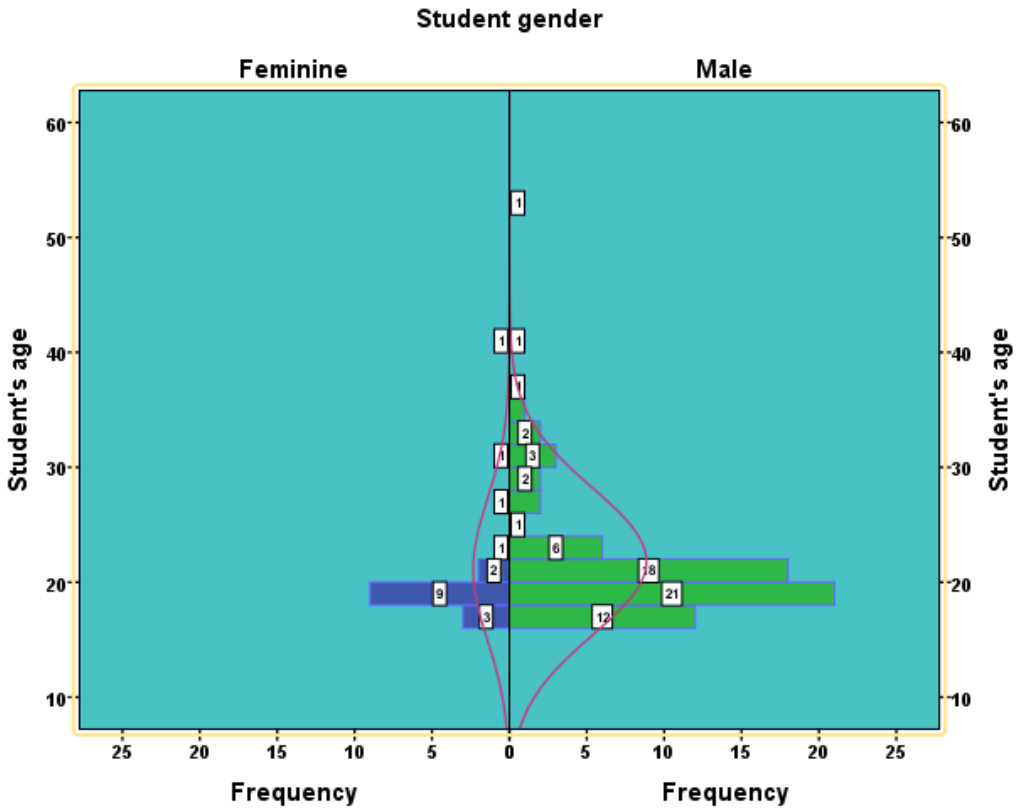


Fig. 4. Analysis of Gender and Age.

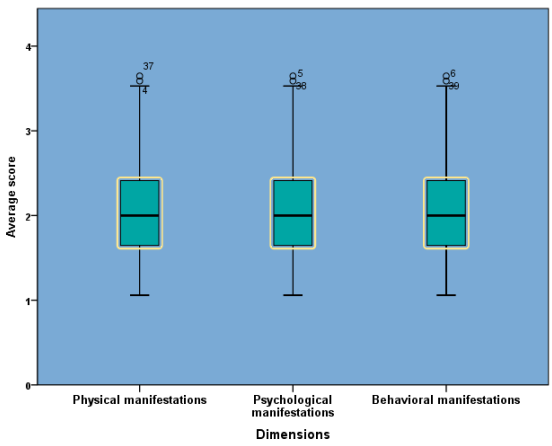


Fig. 5. Analysis by Dimensions.

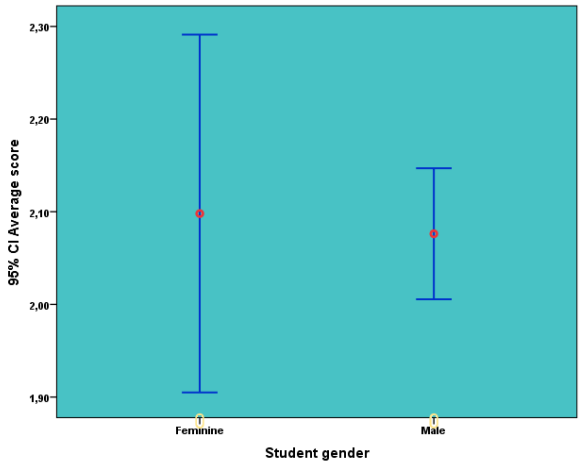


Fig. 6. Confidence Interval Analysis.

observed that the psychological and behavioral manifestations are inverse but very low, that is, the greater the stress, the less academic performance, but it is very low; this can be verified with Fig. 8 where 45,96 % academic performance is high and in Fig. 5 where the 3 dimensions of stress are not higher than the average.

TABLE VIII. RELATIONSHIP BETWEEN VARIABLES

Variables	1	2	3	4
1.Academic performance	1			
2. Physical manifestations	0,076	1		
3. Psychological manifestations	-0,106	0,512**	1	
4. Behavioral manifestations	-0,108	0,647**	0,816**	1

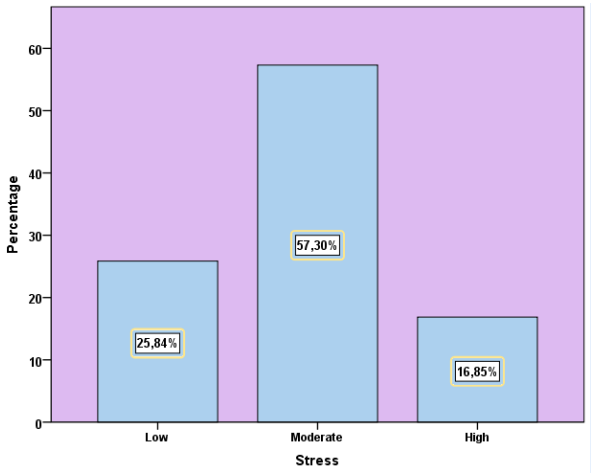


Fig. 7. Student Stress.

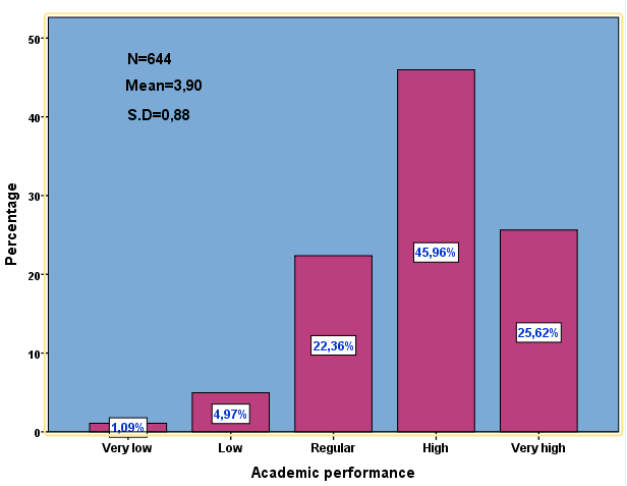


Fig. 8. Academic Performance of Students.

G. Multiple Linear Regression

It is observed in the Table X the predictor variables that are the dimensions of the stress variable and where academic performance is the dependent; it is observed that the significance is greater than 0.05 in the three dimensions of stress. It is stated then that none of the three dimensions of the stress variable is related to academic performance. Besides, $R^2 = 0,040$ indicates that 4% of the academic performance variable is explained by the variables of the Stress dimensions. Analyzing the value of Beta for the prediction of the variables under study, it was found that the beta of the physical and psychological manifestations are positive, that is, but very low, observing the significance greater than 0,05, concluding that there is not relationship; that of psychological manifestations is close to one its significance value, that is, there is not relationship in a determining way.

TABLE VII. COMPARISON OF MEANS OF 2 INDEPENDENT SAMPLES

Dimensions	Gender	Z	P
Physical manifestations	Male	-1,475	0,140
	Femenine		
Psychological manifestations	Male	-0,341	0,733
	Femenine		
Behavioral manifestations	Male	-0,906	0,365
	Femenine		

TABLE IX. SCALE OF VALUES OF THE CORRELATION COEFFICIENT

Value	Meaning
-1	Large and perfect negative correlation.
-0,9 a -0,99	Very high negative correlation.
-0,7 a -0,89	High negative correlation.
-0,4 a -0,69	Moderate negative correlation.
-0,2 a -0,39	Low negative correlation.
-0,01 a -0,19	Very low negative correlation.
0	Null correlation.
0,01 a 0.19	Very low positive correlation.
0,2 a 0,39	Low positive correlation.
0,4 a 0,69	**Moderate positive correlation.**
0,7 a 0,89	**High positive correlation.**
0,9 a 0,99	Very high positive correlation.
1	Large and perfect positive correlation.

TABLE X. MULTIPLE LINEAR REGRESSION

Predictor variable: The 3 dimensions of the stress variable	**Dependent variable: Academic perfomance** $R^2 = 0,040$		
	Beta	**t**	**Sig.**
Physical manifestations	0,235	1,686	0,095
Psychological manifestations	0,012	0,067	0,947
Behavioral manifestations	-0,249	-1,199	0,234

V. DISCUSSION

In the research it was obtained that the relationship between stress and academic performance is very low. This may have an explanation, that stress is not unique to academics; in other words, there are other forms of stress, such as work stress, psychosocial stress and others, product of the pandemic.The academic performance in times of pandemic, with distance classes, in the students of the Faculty of Engineering is high and very high with a percentage of 45,92% and the stress dimensioned in the physical, psychological and behavioral aspects has an average result regular. If compare with the study carried out by [38] on stress and academic performance, they found that if there is an inverse relationship between these 2 variables under study. However, his study was in another context, where there was no presence of the COVID-19 pandemic. They obtained a moderate stress, approximately 72% of students; on the other hand, the research carried out in the article is approximately 50%. In the research, students from the morning shift but also the night shift were surveyed where almost all students work, and may also have work stress, since they do it remotely. This last variable was not considered in the research.

VI. Conclusion and Future Work

According to the quantitative study, it is observed that there is not relationship between stress and academic performance; where the average of the students is high and very high with 45,96% and the stress of the 3 grouped dimensions is moderate with 57,30%. However, students report feeling stress in their academic preparation. Stress manifests itself on a physical level, mainly with headaches and insomnia, with 50,56% of the moderate form; psychologically, with anxiety, demotivation and sadness; but this does not directly affect their interpersonal relationships; obtaining a 55,06% in the moderate way and in the behavioral manifestations a 51,69%. The students indicate that the periods close to the evaluations generate more stress at the academic level; as well as the pressure to pass the subjects, mainly those of science that are considered the most complex. To cope with stress and anxiety, they resort to direct strategies, based on a personal analysis without resorting to external elements. On the other hand, it should be noted that stress is positive to the extent that it allows us to respond correctly to demands or difficulties. In addition, it is observed in students that stress is used as a kind of fuel to achieve their academic goals. There is a relationship between the three dimensions of the stress variable.

The methodology carried out with the techniques of surveys and interviews, allowed to have a holistic panorama of being able to analyze and cross information in the investigation. New studies are suggested in relation to positive stress and academic performance in university students. Also, delve into the situation generated in the teacher as part of teaching in the virtual mode. On the other hand, analyze the subject from a qualitative methodology that allows us to know the perception of those involved in this process.

References

[1] A. Aristovnik, D. Keržič, D. Ravšelj, N. Tomaževič, and L. Umek, "Impacts of the COVID-19 pandemic on life of higher education students: A global perspective," *Sustainability (Switzerland)*, vol. 12, no. 20, 2020.

[2] X. Wang and S. Hegde, "Changwon son, bruce keller, alec smith, and farzan sasangohar. 2020.","," *Investigating Mental Health of US College Students During the COVID-19 Pandemic: Cross-Sectional Survey Study." Journal of Medical Internet Research*, vol. 22, no. 9, p. e22817.

[3] M. Fawaz and A. Samaha, "E-learning: Depression, anxiety, and stress symptomatology among lebanese university students during covid-19 quarantine," in *Nursing forum*, vol. 56, no. 1. Wiley Online Library, 2021, pp. 52–57.

[4] C. A. R. Collazo and R. Hernández, "El estrés académico: una revisión crítica del concepto desde las ciencias de la educación," *Revista electrónica de psicología Iztacala*, vol. 14, no. 2, pp. 1–14, 2011.

[5] H. Selye, "What is stress," *Metabolism*, vol. 5, no. 5, pp. 525–530, 1956.

[6] B. R. García, M. d. P. G. Arrieta, and A. L. E. B. Montagut, "Estresores académicos percibidos por estudiantes pertenecientes a la escuela de enfermería de ávila, centro adscrito a la universidad de salamanca," *RevistaEnfermeríaCyL*, vol. 6, no. 2, pp. 98–105, 2014.

[7] G. M. G. Vargas, "Factores asociados al rendimiento académico en estudiantes univer-sitarios, una reflexión desde la calidad de la educación superior pública," *Revista educación*, vol. 31, no. 1, pp. 43–63, 2007.

[8] R. Castillo, G. J. G. Walker, and J. G. D. Castillo, "Influencia del estrés en el rendimiento académico de un grupo de estudiantes universitarios," *Educación y ciencia*, vol. 4, no. 43, pp. 31–40, 2015.

[9] J. V. Conti, A. M. Mas, and P. P. Sampol, "Diferencias de estrés y afrontamiento del mismo según el género y cómo afecta al rendimiento académico en estudiantes universitarios," *Contextos educativos: Revista de educación*, no. 22, pp. 181–195, 2018.

[10] J. W. Oketch-Oboth and L. O. Okunya, "The Relationship Between Levels of Stress and Academic Performance Among University of Nairobi Students," *International Journal of Learning and Development*, vol. 8, no. 4, 2018.

[11] L. Prifti and E. Rapti, "The relationship between attachment, stress and academic success in Albanian students," *Journal of Educational and Social Research*, vol. 8, no. 2, 2018.

[12] A. K. Bolatov, T. Z. Seisembekov, A. Z. Askarova, R. K. Baikanova, D. S. Smailova, and E. Fabbro, "Online-learning due to covid-19 improved mental health among medical students," *Medical science educator*, vol. 31, no. 1, pp. 183–192, 2021.

[13] S. P. Burudi, J. Wasike, and L. Ndegwa, "Desafíos que enfrentan las bibliotecas académicas en la utilización de dispositivos móviles en el acceso y uso de la información en la universidad de kenyatta y la universidad de nairobi en kenia," *African Journal of Education, Science and Technology*.

[14] R. M. Oducado and H. Estoque, "Online learning in nursing education during the covid-19 pandemic: Stress, satisfaction, and academic performance," *Journal Of Nursing Practice*, vol. 4, no. 2, pp. 143–153, 2021.

[15] Dawood Ahmad, Iftikhar Ahmad Baig, and Namra Munir. , "Relación del rendimiento académico con el estrés percibido y el índice de masa corporal." *Dilemas contemporáneos: Educación, Política y Valores*, aug 2019.

[16] C. Son, S. Hegde, A. Smith, X. Wang, and F. Sasangohar, "Effects of covid-19 on college students' mental health in the united states: Interview survey study," *Journal of medical internet research*, vol. 22, no. 9, p. e21279, 2020.

[17] S. F. Abdullah, N. A. Shah, and R. M. Idaris, "Stress and its relationship with the academic performance of higher institution students," *International Journal of Advanced Research in Education and Society*, vol. 2, no. 1, pp. 61–73, 2020.

[18] S. Tejedor, L. Cervi, F. Tusa, and A. Parola, "Education in times of pandemic: Reflections of students and teachers on virtual university education in Spain, Italy and Ecuador," *Revista Latina de Comunicacion Social*, vol. 2020, no. 78, 2020.

[19] Y. Chandra, "Online education during covid-19: perception of academic stress and emotional intelligence coping strategies among college students," *Asian education and development studies*, 2020.

[20] X. Li, P. Fu, C. Fan, M. Zhu, and M. Li, "Covid-19 stress and mental health of students in locked-down colleges," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, p. 771, 2021.

[21] S. S. Changwon, "H. & et al. 2020. effects of covid-19 on college students' mental health in the united states: Interview survey study," *Journal of medical internet research*, vol. 22.

[22] J. Brannen, "Mixing methods: The entry of qualitative and quantitative approaches into the research process," *International Journal of Social Research Methodology: Theory and Practice*, vol. 8, no. 3, 2005.

[23] A. Jarde, J. M. Losilla, and J. Vives, "Methodological quality assessment tools of non-experimental studies: A systematic review," *Anales de Psicología/Annals of Psychology*, vol. 28, no. 2, 2012.

[24] D. Manrique-Millones, R. Millones-Rivalles, and O. Manrique-Pino, "The sisco inventory of academic stress: Examination of its psychometric properties in a peruvian sample," *Ansiedad y Estrés*, vol. 25, no. 1, pp. 28–34, 2019.

[25] L. O. O. Ugarte, S. F. Morales-Hernández, and M. K. Solano-Jáuregui, "Evidencias psicométricas de inventario sisco sv-21 para el estudio del estrés académico en universitarios peruanos," *Propósitos y Representaciones*, vol. 9, no. 2, p. 647, 2021.

[26] E. C. Barboza and L. R. Miranda, "Análisis de confiabilidad y validez de un cuestionario sobre entornos personales de aprendizaje (ple)," *Ensayos Pedagógicos*, vol. 13, no. 1, pp. 71–106, 2018.

[27] I. Iraola-Real, L. K. H. Sarmiento, C. M. Sanchez, and C. Andersson, "Mathematical self-efficacy and collaborative learning strategies in engineering career aspirants," in *The International Conference on Advances in Emerging Trends and Technologies*. Springer, 2020, pp. 127–136.

[28] I. Iraola-Real and M. Gonzales-Macavilca, "Anxiety and poor performance facing exams and studying mathematics: Predictive research on applicants to the universidad nacional de ingenieria of lima (peru)," in *2020 IEEE World Conference on Engineering Education (EDUNINE)*. IEEE, 2020, pp. 1–4.

[29] E. F. Medina, Y. P. Somoza, L. Andrade-Arenas, J. C. Zavaleta, R. Y. Alva, and S. V. Vargas, "Analysis of distance learning in the professional school of systems engineering and informatics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120763

[30] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 104–112, 2020.

[31] M. E. Silva-Ramos, María Fernanda; López-Cocotle, José Juan; Columba Meza-Zamora, "Estrés académico en estudiantes universitarios Investigación y Ciencia - 2020," *Investigación y Ciencia*, pp. vol.28,num.79,2020,pp75–83, 2020. [Online]. Available: https://www.redalyc.org/articulo.oa?id=67462875008

[32] S. E. Piemontesi, D. E. Heredia, L. A. Furlan, J. S. Rosas, and M. Martínez, "Ansiedad ante los exámenes y estilos de afrontamiento ante el estrés académico en estudiantes universitarios," *Anales de Psicología/Annals of Psychology*, vol. 28, no. 1, pp. 89–96, 2012.

[33] R. E. Encina, L. B. Meza, and M. Auchter, "Estres academico percibido por los estudiantes que finalizan el primer año de licenciatura en enfermeria de la UNNE TT - Academic stres perceived by students that finish their first year of bachelor degree in nursing from the UNNE," *Notas enferm. (Córdoba)*, vol. 18, no. 32, pp. 27–32, 2018. [Online]. Available: https://revistas.unc.edu.ar/index.php/notasenf/article/view/22744/22355

[34] D. Córdova and E. Irigoyen, "Estrés y su Asociación en Rendimiento Académico en los Estudiantes de la Facultad de Medicina desde primero a octavo nivel de la PUCE, sede Quito en el periodo correspondiente de Enero a Mayo 2015," Ph.D. dissertation, 2015.

[35] G. Boullosa Galarza, "Facultad de letras y ciencias humanas," Ph.D. dissertation, 2013. [Online]. Available: https://n9.cl/4l87a

[36] A. B. MacÍas, "El estrés académico en alumnos de maestría y sus variables moduladoras: Un diseño de diferencia de grupos," *Avances en Psicologia Latinoamericana*, vol. 26, no. 2, pp. 270–289, 2008.

[37] S. A. Díaz Azuara and C. R. Fierro Santillán, "Aplicación de un e-cuestionario de eustrés y distrés académicos socioformativos en estudiantes de educación media superio," *Eutopía*, vol. 11, no. 28, pp. 22–28, 2018. [Online]. Available: http://revistas.unam.mx/index.php/eutopia/article/view/65904

[38] N. Sohail, "Stress and academic performance among medical students," *J Coll Physicians Surg Pak*, vol. 23, no. 1, pp. 67–71, 2013.

Implementation of a Web System to Improve the Evaluation System of an Institute in Lima

Franco Manrique jaime, Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—In Peru and the world, millions of students saw their education interrupted caused by the problems brought by the COVID-19 virus, due to this many educational entities began to adopt learning platforms and web systems so that the teaching process is not affected, having to comply with all the guidelines and requirements of the institution to solve any academic difficulty. That is why in the present work the implementation of a web system was proposed for the improvement of the qualification and evaluation processes of an institution using the scrum methodology since it is an agile framework that is based on empiricism and offers adaptation and flexibility in the projects. For the software development, the open source language PHP was used since it is more adapted to these web systems, Mysql was also used, which is a database manager for relational databases. The results of this research was the correct implementation of this system to the educational institution, verifying the absence of errors and the improvement of the processes involved so that the institution can provide students with an adequate learning process.**

Keywords*—COVID-19; evaluation; learning platform; scrum; web system

I. Introduction

Today the percentage of educational institutions that adopt a virtual teaching platform is increasing, this due in large part to confinement due to pandemic, creating a change in the global teaching process, Currently, virtual teaching is defined as the development of the learning process through the use of digital and information tools, where the student and the teacher are not physically present in the same room. Nowadays, many students went from the face-to-face to the virtual, bringing with them new problems to solve. This new reality brought new methods and ways to have a better understanding of the educational process in universities and institutions [1]. The change in the way of teaching in institutions due to the COVID-19 pandemic forced many educational entities to adopt virtual teaching systems since, due to the circumstances, classes and online interaction are considered the best option that can be adapted to this new form of teaching and that complies with the confinement regulations issued by the governments of different parts of the planet. These web systems provide all kinds of tools so that students can learn from home without having to return to face-to-face classes or be in contact with other people [2]. Although most institutions have some type or form of software to process, retrieve, store information, today these systems need to be increasingly accurate with a minimum error percentage because these systems must provide students and teachers tools and resources quickly and safely [3]. Some platforms have certain limitations that prevent covering all the needs of the institutions, these institutions, by not having the required modules and by not satisfying all the needs of the different areas involved, are not able to provide the information necessary for decision-making in an academic setting. This affects students because the student, guided by the teacher, makes use of the different tools and resources offered by these web systems to access important information and develop activities that help the student consolidate knowledge and develop skills.

For all the aforementioned and with the purpose of solving these academic difficulties of the Institute of Sciences and Humanities (ICH), through technology, specifically in the design and development of a web system, the following question was asked: What measure will a web system of qualification and evaluation improve the process of monitoring and control of information in the Institution of Sciences and Humanities?

Due to the lack of adherence of some processes with the requirements of the institution, this can harm its decision-making, affecting the learning and teaching process. This work offers a proposal, solution to the problems of the educational institution through the web system that will optimize resources by improving the level of satisfaction of those involved as teachers and students, in addition, the web system will accelerate the evaluation and qualification processes of the institution making them more efficient and with fewer errors.

The objective of this project is to seek to improve through a web system the qualification process and results of mock admission exams managed by the directorate of systems and communications of the ICH.

In Section II, a review of the literature was carried out; in Section III the methodology; as well as in Section IV the case study; in Section V the result and discussion; and finally in Section VI Conclusion and future work.

II. Literature Review

The project consists of designing and implementing a web system for the improvement of the qualification process and results of the mock admission exam of an educational institution to provide a better experience for students in addition to processing and analyzing the grades of students to obtain information performance and progress. A large percentage of educational institutions in Peru use learning platforms for virtual teaching and with the presence of COVID-19 and

quarantine this percentage is increasing even more, so knowing these tools today is essential for institutions educational.

Regarding virtual education due to pandemic, there are different positions such as that of the author Eduardo Norman Acevedo [4] which says that due to the current situation referring to the health situation caused by COVID-19, educational institutions had to change the teaching processes to adapt to this new way of teaching and learning, being mainly evaluative, ethical, sociological, pedagogical changes. On the other hand, María Laura Picón [5] indicates that millions of students were affected by this problem, having to adapt to these new tools in order to receive online classes, however, how prepared were not only the institutions but also the students? This change in teaching brought with it new problems for students such as lack of technological resources, lack of internet in rural areas, students with disabilities, etc. This need to apply virtual teaching revealed the inequality of training for the use of these technologies and the lack of knowledge by both teachers and students. Based on the above, Javier Arturo Hall López [6] mentions that virtual teaching also brought problems on the physical health of students, due to sedentary lifestyle and lack of exercise make students more likely to suffer from weight problems, this is a challenge pedagogical in teaching and in the way of teaching classes, which has to propose better teaching methodologies so that the student can be guided in the development of their motor skills. At the same time, these education platforms are increasingly being adopted by more education centers, as explained by D. Benta [7] in his report, where he mentions that tools such as Moodle help motivate students and those involved in solving collaborative and individual tasks.

About traditional teaching, Jacinto Joaquín Vertíz Osores [8] indicates, that the traditional form of evaluation used in universities has been attributed almost entirely to the criteria of the professors, who guided the process towards the quantification of the indicators expressed in their links and that they are supposed to reflect the level of progress of the students. In retrospect, this traditional way of evaluating students to measure learning and level of knowledge has remained frozen in time and does not appear to have changed in the last 50 years [9]. However, at the current stage, where the educational scheme is not the same, with means of remote interaction, in this new scheme, teachers ask university authorities about evaluation methods, consult strategies and propose methods, being fully aware of the shortcomings of the new interaction scenario. On the other hand, Olga V. Bondarenko [10] points out that, in a virtual educational environment, the nature of the interaction of those involved in learning (a student, a group of students or a teacher) changes fundamentally. The term study is not used because it is interpreted as a cognitive activity of students who, under the guidance of a teacher, master skills, knowledge. The involvement of students in a virtual education environment means that the teacher must change his role from mentor and director to tutor, facilitator and moderator [11]. The teacher strives to help students find an individual educational path. Currently and due to the pandemic, there are more and more modern students living in a media environment where the use of computers, internet resources and the use of mobile devices is part of their day to day. It is for this point that virtual educational environments are now a reality that basically deals with an information space for the interaction of students in an educational process generated by communication and information.

The advantages of a virtual educational environment include: flexibility, economic efficiency, interactivity, mobility. In this context, M'Balia Thomas [12] indicates that the change to virtual classes due to the pandemic forced teacher educators to rethink them, they do not normally teach or design the content of online courses. Thus, as the global pandemic grows and accelerates a continued rise in e-learning, teacher educators must reexamine what it means to be equitable, responsive and inclusive to the individual needs of a diverse set of pre-service teachers and take into account consideration of collective professional needs. With regard to systems related to virtual education, there are various authors such as Cristian Enrique Mejía and Mariano Enrique Álava [13] who developed a grade registration system for a school with 60 years of service that lacked a virtual education environment, it is for this reason that there was a deficient management of the file with important content such as qualifications, enrollment, etc., and that they were solved by this software that optimized said processes. It also allows to reduce the response time of processes and improve the teaching process and thus increase the satisfaction of the students. It is clear that these web systems helped learning from home, institutions such as those mentioned had to adapt to these new platforms so that this training process is not interrupted and although not all have the means to take this type of non-face-to-face classes, a large percentage were able to continue with their training process.

In summary, the authors analyzed in their research show that there is still a need to carry out digital transformation in institutions to make the implementation of the evaluation system more effective and efficient. The investigation that will be carried out is to see how to optimize with the implementation of the web page in an effective way in its processes.

III. Methodology

A. Methodology SCRUM:

The Scrum methodology is a framework that addresses complex problems (ver Fig. 1), where a lot of ambiguity can be observed; in addition, it helps to promote creativity and productivity of teams because events tend to generate creative tension in the team [14]. In addition to this always tends to seek to deliver products of the highest value. both scrum and other methodologies allow the team to make the most of resources and time, however this does not mean that this framework does not have disadvantages. Scrum is based on empiricism, rather, learning as you go, it is for this reason that the incremental interactive life cycle is usually perfect for this framework [15]. In Scrum there are roles which are the Scrum Master, the Scrum team and the Product Owner [16]. This agile framework or framework created in the 90's by Schwaber and Sutherland. It follows an approach that is based on evidence, starting from the premise that problems can never be defined or understood in their entirety, which is why organizations and companies have to focus on getting trained teams to respond to change.

The characteristics that most identify this framework [17], are the following:

- Flexibility and adaptation.

- Return on investment.
- Risk mitigation.
- An always motivated team.
- An alignment between the client and the team Productivity and quality.
- This intensified work entails a high prediction of times since the speed and performance of the equipment is known.
- The method of work and continuous review produces a higher quality of the software.

B. Scrum Activities

1) Sprint Planning: This is a meeting that has to be held at the beginning of a sprint and it is necessary for the entire team to participate [15]. This task planning is divided into two main parts:

First part of the meeting. This has a Timebox of 4 hours.

- This is where the client shows the list of requirements to the manager, and specifies the priority requirements.
- Then the list is analyzed, doubts are resolved and it is agreed which requirements will be a priority in each iteration.

Second part of the meeting: As in the first part, here a timebox is made that can last a maximum of 4 hours. Here the iteration is analyzed and planned, and proceeds to create tactics that will help you meet the objective.

- Team members can assign themselves tasks to perform.
- The effort to perform each task will be calculated.

2) Sprint: In Scrum we have iterations, these iterations in Scrum are called sprint, each one of the sprint delivers value to the end users or to those who are solving a particular problem, this value grows with each iteration.

3) Scrum Daily Meeting: In these meetings, what is sought is the transfer of relevant value together with the collaboration of team members to improve productivity. Here each member reviews everyone's work to be able to make the necessary changes to finish and comply with the sprint.

4) Sprint Retrospective: This is the event where all members of the Team evaluate and inspect themselves, this in order to improve during the next Sprint. Here we analyze how they are working, why they achieve or not the objective in which they are failing.

5) Sprint Review: This event takes place at the end of each sprint. During this informal meeting the increase is reviewed, rather, what was done during the sprint and if there were changes, the Product Backlog is analyzed.

6) Team Roles: This methodology is made up of three main roles. Each of these roles has different responsibilities and must act differently [18].

- **Scrum Master:** He is an expert in scrum, he is a facilitator coach, he trains and removes the impediments that can be seen in daily meetings.
- **Product Owner:** It is the empowered person, who establishes the product backlog, is responsible for the return on investment, is represented by a single person and works on the product vision.
- **Development team:** All the members are called developments and they are full time in what they are doing, it is a multidisciplinary team that is self-managed and estimates its own time, it is a small team and it is recommended that it be from 3 to 9 people.

7) Product Backlog: This is a list with elements, these elements are called PBI (Product backlog item), this artifact is not only composed of user stories, in the scrum guide they mention that other values such as needs, requirements, cases of uses, defects etc. This stack is ordered, being the first in the stack the ones that generate the greatest value to the business or the highest value to the end users and therefore the last in the stack is what generates the least value or little value compared to the first pbi. Scrum, being iterative and agile, allows new PBIs to be entered into the backlog that may arise as we advance in the project, that is why it is said that the product backlog is emergent, which means that we can be adding new things according to the product or according to the project I need it.

8) Burndown Chart: This is a graph that shows over time the speed and the time in which the objectives and requirements are being completed. Provides a better overview of the team's pace of progress in order to predict whether the team will complete the work in the estimated times.

9) ScrumTaskboard: This is a board with the tasks to be done. And be organized in such a way that the objectives are assigned the necessary tasks to finish it, usually post-its are used, and they change position according to their status. To differentiate the tasks that each member is performing, smaller colored stickers are used.

10) Disadvantages of Scrum: Scrum is not perfect. Therefore, we can mention some of the most important disadvantages of Scrum:

At the time of starting with Scrum, the whole team must know very well its principles and theoretical framework, also that there be knowledge of the Scrum roles or this could alter the functions and tasks. To a large extent, much of the success of the project will depend on the level of knowledge of the scrum master (rather than on the innovation, creativity or quality of the inputs). If for some reason some task is not completed, other tasks may be compromised due to that sometimes these can be related, this can generate delays in the sprints or the delivery of an unsatisfactory value.

The following figure shows the stages of this methodology graphically, showing the order of the activities [14].

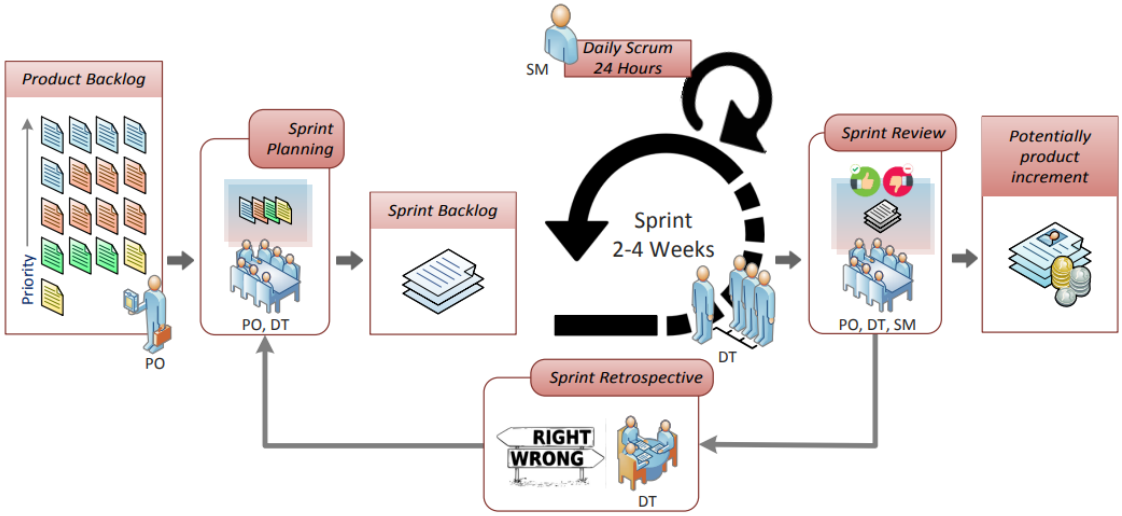


Fig. 1. The Phases of the Scrum Methodology.

IV. Case Study

A. Initiation Phase

1) User Story: In Table I, was create user stories, that are general and informal explanations of the characteristics of the software, these are from the point of view of the end user, in our case we identified 22.

2) Product Backlog: In the product backlog basically organize the list of tasks was carried out for the project. In Table II, we show the priority or rather the degree of importance and to which module it belongs, in this work we were able to identify 4 modules.

B. Planning and Estimation Phase

1) Sprint Backlog: In Table III, Table IV, Table V are the objectives, the elements chosen in the product backlog to be completed in each sprint, in our project we identify three sprints.

In Table VI, the weighting that was used is detailed in the following.

2) User Story Sprint 1:

a) Exam management module 1: In this module 7 user stories were created for the creation, deletion, modification and list of exams, likewise a time can be assigned for taking the test, also for greater convenience the exam can be copied.Table VII and Table VIII show the windows of this module.

b) Exam management module 2: In this module 4 user stories were created for the management of sections, questions and courses, this window can only be used by administrator type users.In the part of the questions because they are associated with an exam, the questions cannot be eliminated directly, what is done is a logical elimination. Table IX and Table X show the windows of this module.

TABLE I. User Story

ID	USER STORY
1	As an administrator I need to view the windows to configure the exams.
2	As Administrator I need to register an answer for each question.
3	As Administrator I need to register the sections to place the questions.
4	As Administrator I need to record the time for each area.
5	As Administrator I need to register the number of questions for each area.
6	As Administrator I need to record the total time for all sections.
7	As Administrator I need to register the total number of questions to set limits.
8	As Administrator I need to register the courses for each section.
9	As Administrator I need to have reports by dates to indicate the day the exam was taken.
10	As Administrator I need a merit report by specialty to see your information.
11	As Administrator I need a general merit report to see your information.
12	As an administrator I need to edit the number of questions per area to update the data.
13	As Administrator list the sections to view the existing information.
14	As Administrator I need to have a course report to see your information
15	As Administrator I need to view the exam model to see the changes made.
16	As a student, I need to take the exam for the evaluation.
17	As a Student I need to visualize the windows for taking exams.
18	As a Student I need a report card to see my results.
19	As a Student I need to finish the test to choose a new section.
20	As a Student I need to see the detailed result, showing which ones were incorrect, which ones were correct and which ones I did not answer.
21	As a Student I need to finish the exam even if time is not over yet.
22	As a Student I need to be able to enter an exam to view their schedules.

3) User Story Sprint 2: In this module the administrators will be able to observe the progress of the students in the courses as well as view the reports of the students by date, year, course and type of exam. These data can later be analyzed to obtain statistical tables as a result. Table XI and Table XII show the windows of this module.

TABLE II. PRODUCT BACKLOG

ID	USER STORY	Priority	Sprint	MODULE
1	As an administrator I need to view the windows to configure the exams.	High	1	Exam management 1
2	As Administrator I need to register an answer for each question.	High		
3	As Administrator I need to register the sections to place the questions.	High		
4	As Administrator I need to record the time for each area.	High		Exam management 2
5	As Administrator I need to register the number of questions for each area.	Half		
6	As Administrator I need to record the total time for all sections.	High		
7	As Administrator I need to register the total number of questions to set limits.	Half		
8	As Administrator I need to register the courses for each section.	High	2	Results
9	As Administrator I need to have reports by dates to indicate the day the exam was taken.	Low		
10	As Administrator I need a merit report by specialty to see your information.	High		
11	As Administrator I need a general merit report to see your information.	High		
12	As a Student I need to take the exam for the evaluation.	High	3	Student
13	As a Student I need to visualize the windows for taking exams.	High		
14	As a Student I need a report card to see my results.	Half		
15	As a Student I need to finish the test to choose a new section.	High		
16	As a Student I need to be able to enter an exam to view their schedules.	High		

TABLE III. SPRINT BACKLOG FROM SPRINT 1

PRIORITY	CHORES	DIFFICULTY	HOURS
	SPRINT 1		11
1	As an administrator I need to view the windows to configure the exams.	5	3
2	As Administrator I need to register an answer for each question.	3	2
3	As Administrator I need to register the sections to place the questions.	3	1
4	As Administrator I need to record the time for each area.	2	1
5	As Administrator I need to register the number of questions for each area.	2	2
6	As Administrator I need to record the total time for all sections.	2	1
7	As Administrator I need to register the total number of questions to set limits.	2	1

4) User Story Sprint 3: In this module, students will be able to take the exams and see their results.To take an exam, the date and time have to be appropriate in addition to having previously chosen an exam.After the student completes the test, the scores and the exam are calculated It ends without the option of taking it again. To enter this module the student must be previously registered in the system.Students who did not solve the test are scheduled another date and their score

TABLE IV. SPRINT BACKLOG FROM SPRINT 2

PRIORITY	CHORES	DIFFICULTY	HOURS
	SPRINT 2		10
1	As Administrator I need to register the courses for each section.	4	3
2	As Administrator I need to have reports by dates to indicate the day the exam was taken.	4	3
3	As Administrator I need a merit report by specialty to see your information.	3	2
4	As Administrator I need a general merit report to see your information.	3	1
5	As Administrator I need to register the number of questions for each area.	2	1

TABLE V. SPRINT BACKLOG FROM SPRINT 3

PRIORITY	CHORES	DIFFICULTY	HOURS
	SPRINT 3		12
1	As an administrator I need to edit the number of questions per area to update the data.	3	3
2	As Administrator list the sections to view the existing information.	3	3
3	As Administrator I need to have a course report to see your information.	3	2
4	As Administrator I need to view the exam model to see the changes made.	3	1
5	As Administrator I need to manage the scores by section.	4	3

is modified, in addition the student's information is saved and processed for future reports Table XIII and Table XIV show the windows of this module.

5) Design Pattern MVC: The model view controller pattern (mvc) is a pattern that separates the data into three different modules, each fulfilling a function different. This software architecture pattern applies code reuse, facilitating software development and maintenance as shown in Fig. 2 [14]. In Fig. 3, we can see the order of the folders following this pattern. This model can be used to develop desktop applications or mobile applications in this way creating the mobile version of the system will be much easier.

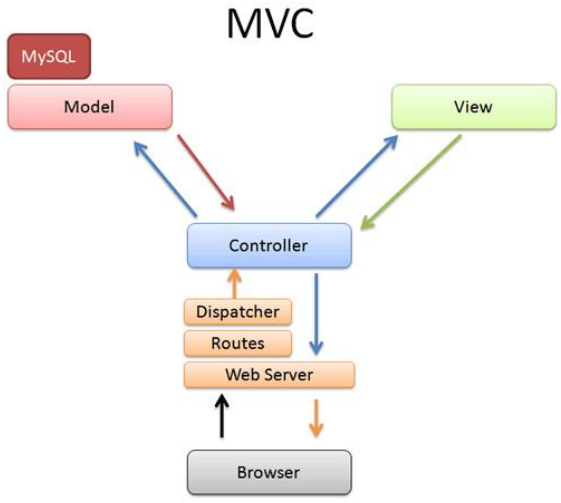


Fig. 2. MVC Operation.

TABLE VI. WEIGHTING CRITERION

DIFFICULTY	WEIGHT
Easy	1
Little Easy	2
Regular	3
Little difficult	4
Hard	5

TABLE VII. USER STORY-EXAM MANAGEMENT

USER STORY	
Number 1	User: Administrator
Story name: Exam Settings	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:1
Responsible developer: Manrique jaime franco	
Description: In this window the administrator will be able to manage the exams for later activation.	
Observation: The user must be previously logged in.the user must be an administrator.	
CRITERIA OF ACCEPTANCE	
The system generates exams	
The system displays recorded exams	
PROTOTYPE	

6) Database Model: Fig. 4 shows the diagram of the database.The diagram has 15 tables; the main ones being the exam table where the exams with the corresponding questions are recorded and the configuration table which relates the students with the grades.
The DB was developed in MYSQL since it is a database manager that best adapts to web systems as well as being free and easy to use. This manager is compatible with Windows and Linux operating systems.
The database was developed using Navicat which is a tool for the development of BDs and that allows managing multiple databases.This tool was chosen for its compatibility with the manager and for its easy handling.

V. RESULT AND DISCUSSION

A. Sprint 1 Testing and Review

1) Exam: In Fig. 5, the exam window is shown where the exams will be shown, registered, deleted and modified, also it will be possible to create schedules and register questions. For the administrator's convenience, it will also be possible to duplicate the exam including the questions.

2) Section: Fig. 6 shows the section window where sections will be displayed, registered, deleted and modified. To register a section it is necessary to specify its name.

TABLE VIII. USER STORY-SECTION MANAGEMENT

USER STORY	
Number: 2	User: Administrator
Story name: section management	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:1
Responsible developer: Manrique jaime franco	
Description: In this window the administrator will be able to manage the sections for later use in exams.	
Observation: The user must be previously logged in.The user must be an administrator.	
CRITERIA OF ACCEPTANCE	
The system generates sections	
The system displays registered sections	
PROTOTYPE	

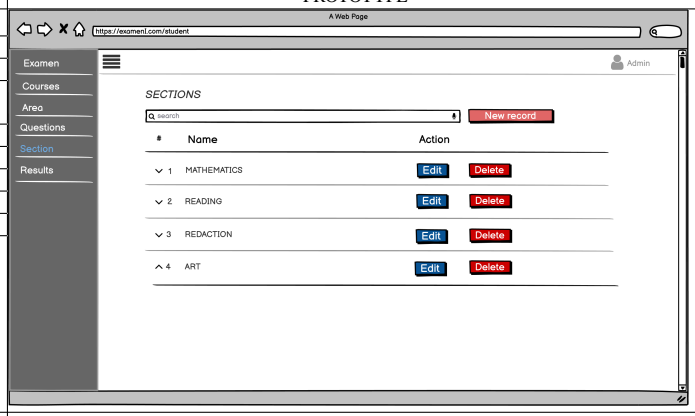
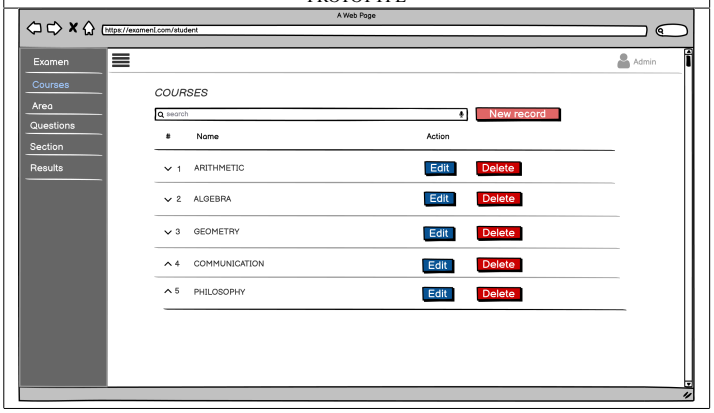


TABLE IX. USER STORY-COURSE MANAGEMENT

USER STORY	
Number: 3	User: Administrator
Story name: course management	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:1
Responsible developer: Manrique jaime franco	
Description: In this window the administrator will be able to manage the courses for later use in exams.	
Observation: The user must be previously logged in.The user must be of administrator.	
CRITERIA OF ACCEPTANCE	
The system generates courses	
The system displays the registered courses	
PROTOTYPE	



3) Course: Fig. 7 shows the course window where the courses will be displayed, registered, deleted and modified.To register a course it is necessary to specify the section to which it belongs.

4) Area: In Fig. 8, the window of areas is shown which for security reasons cannot be modified, for now the only way

TABLE X. USER STORY-QUESTION MANAGEMENT

USER STORY	
Number: 4	User: Administrator
Story name: question management	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:1
Responsible developer: Manrique jaime franco	
Description: In this window the administrator can manage the questions for later use in exams.	
Observation: The user must be previously logged in.The user must be of type administrator.	
CRITERIA OF ACCEPTANCE	
The system generates questions	
The system displays the registered questions	
PROTOTYPE	

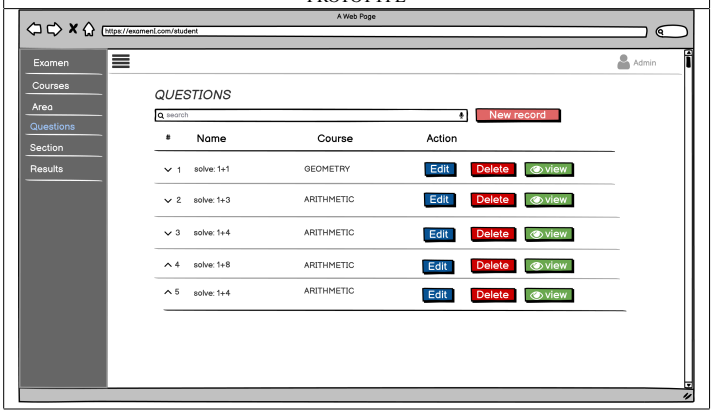


TABLE XI. USER STORY-STUDENT REPORT

USER STORY	
Number: 5	User: Administrator
Story name: question management	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:2
Responsible developer: Manrique jaime franco	
Description: In this window the administrator will be able to see the reports for later use in the analyzes.	
Observation: The user must be previously logged in.	
CRITERIA OF ACCEPTANCE	
The system generates the reports	
The system filters by data	
PROTOTYPE	

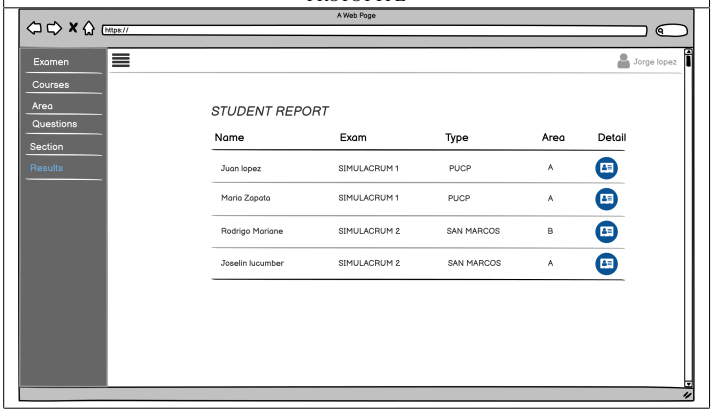


TABLE XII. USER STORY-REPORT OF MERITS

USER STORY	
Number: 6	User: Administrator
Story name: report of merits	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:2
Responsible developer: Manrique jaime franco	
Description: In this window the administrator can view the merits of the students in order of specialty and merit.	
Observation: The user must be previously logged in as administrator type.	
CRITERIA OF ACCEPTANCE	
The system generates reports by merit	
The system filters according to the data	
PROTOTYPE	

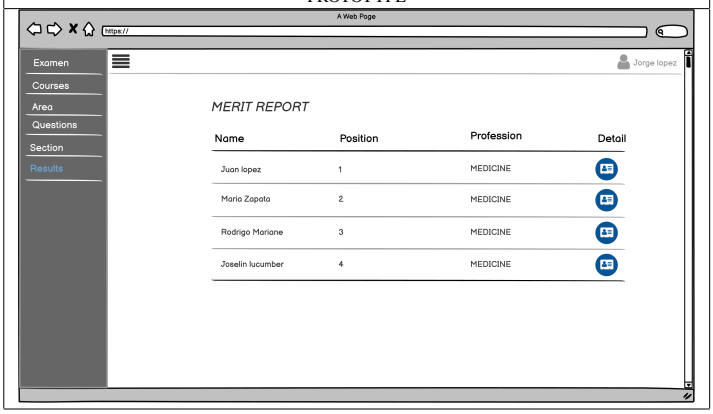
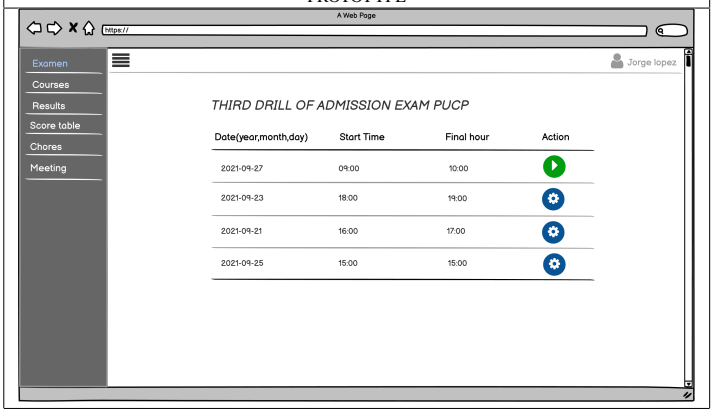


TABLE XIII. USER STORY-EXAM TIME TABLES

USER STORY	
Number: 7	Uer: Administrator
Story name: exam schedules	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:3
Responsible developer: Manrique jaime franco	
Description: In this window the student will be able to view the times of the selected exam.	
Observation: The user must be previously logged in as an administrator.	
CRITERIA OF ACCEPTANCE	
The system lists the schedules	
The system displays the available schedules	
PROTOTYPE	



to add or delete areas is through the database.

5) Question: Fig. 9 shows the question window where the entered questions will be displayed, registered, deleted and modified.

6) Alternatives: Fig. 10 shows the alternatives window where the alternatives entered per question will be shown, registered, eliminated and modified.

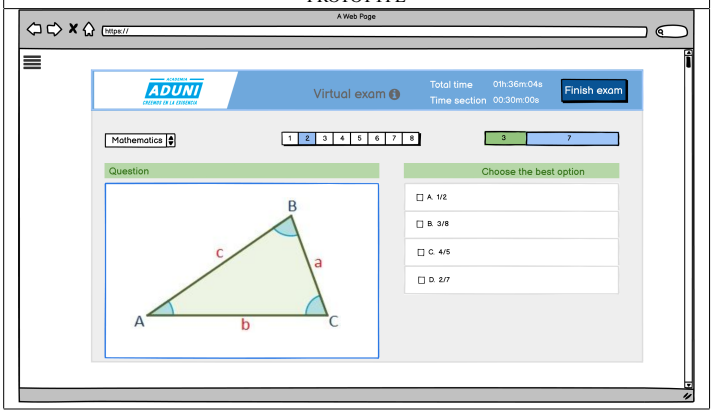
B. Sprint 2 Testing and Review

1) Exam Results: Fig. 11 shows the results exam window where the list of exams taken will be shown, clicking on one will show the students who took said exam.

2) Student results: Fig. 12 shows the results window by exam where the list of students who took the selected exam will be shown, showing their grade and date, also as the

TABLE XIV. USER STORY-VIEW TEST

USER STORY	
Number: 8	User: Administrator
Story name: view test	
Business priority: High	Developing risk: Low
Estimated points:5	Assigned Iteration:3
Responsible developer: Manrique jaime franco	
Description: In this window the student will be able to view the respective test and do it.	
Observation: The user must be previously logged in as an administrator.	
CRITERIA OF ACCEPTANCE	
The system displays the exam	
The system calculates the scores	
PROTOTYPE	



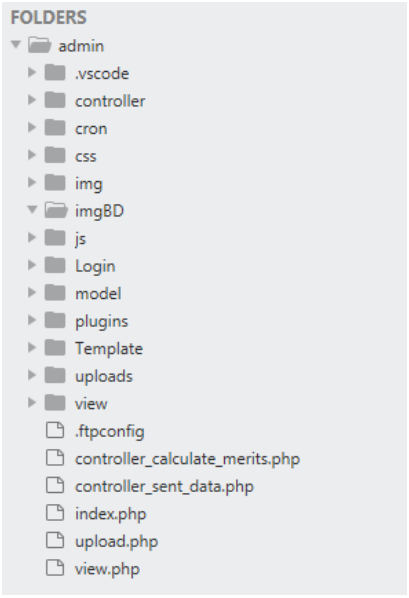


Fig. 3. MVC Order.

database handles a large number of records, a search engine was added to find the student quickly.

C. Sprint 3 Testing and Review

1) Exam: Fig. 13 shows the time list window where the student will view the available schedules, if the student has already completed the test, the buttons will be blocked preventing entry.

2) Test: Fig. 14 shows the exam window where the student will take the virtual practice, the practice has a general time limit and by section, if the student finishes before they can press the button finish exam to finish it, once finished a message will be displayed thanking you for your participation and will redirect you to the schedule list window.

3) Student results: Fig. 15 shows the results window where the student will view the scores obtained from the test in addition to the position by general merit and by specialty, he will also be able to view all the test questions along with the correct answer and the alternative marked by the student.If the student does not take the test, a new date is scheduled, however the merits will not be shown.

The notes are calculated from time to time through files called CRON which are responsible for executing the code every 40 minutes, modifying the time in which the CRON will be executed is done manually. It is due to this file that if a student takes the exam after the scheduled date, it will have no merit and only their score will be recorded.

According to the author [2], his studies were based on the use of web systems by students remotely, where the advantages of using the platform in times of pandemic stand out, which is consistent with our research since it was also carried out in the same pandemic context where the Authorities show support for the investigation as there is satisfaction on their part.

VI. CONCLUSION AND FUTURE WORK

According to the requirements and needs of the Institution of Sciences and Humanities, a web system was implemented for the improvement of evaluation and qualification processes, which consists of three modules which are the exam, results and student management module. The response time of the system was also reduced when displaying the exam and scoring the tests, in the same way it was possible to create a friendly interface for both students and administrative staff, thus improving the creation and taking of virtual evaluations.It is suggested that future work be implemented a mobile application that allows complementing the institute's evaluation system.

REFERENCES

[1] N. Bocanegra and M. Navarro, "Evaluación Virtual: Un recurso para potenciar la Autorregulación y el Aprendizaje," *Redie.Mx*, p. 220, 2019. [Online]. Available: http://www.redie.mx/librosyrevistas/libros/evaluacion virtual.pdf

[2] F. d. R. A. Gordón, "From face-to-face learning to virtual learning in pandemic times," *Estudios Pedagogicos*, vol. 46, no. 3, pp. 213–223, 2020.

[3] I. Gómez-Arteta and F. Escobar-Mamani, "Educaciòn virtual en tiempos de pandemia: incremento de la desigualdad social en el Perú Virtual education in times of Pandemic : increasing social inequality in Perú," *Chakiñan, Revista De Ciencias Sociales Y Humanidades*, 2021. [Online]. Available: https://chakinan.unach.edu.ec/index.php/chakinan/article/view/553

[4] E. N. Acevedo and C. E. Daza-orozco, "construction of content for virtual edication : lockdown ' s content for virtual education : lockdown ' s," pp. 0–4, 2020.

[5] P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta, "Agile software development methods: Review and analysis," *CoRR*, vol. abs/1709.0, 2017.

[6] P. Srinivasan, *Editor*, 2020, no. June.

[7] R. J. Wilcha, "Effectiveness of virtual medical teaching during the COVID-19 crisis: Systematic review," *JMIR Medical Education*, vol. 6, no. 2, pp. 1–16, 2020.

[8] J. J. V. Osores, R. R. C. Flores, R. I. Vértiz-Osores, G. L. V. Ochoa, and A. A. Romero, "Virtual university education in the context of the health

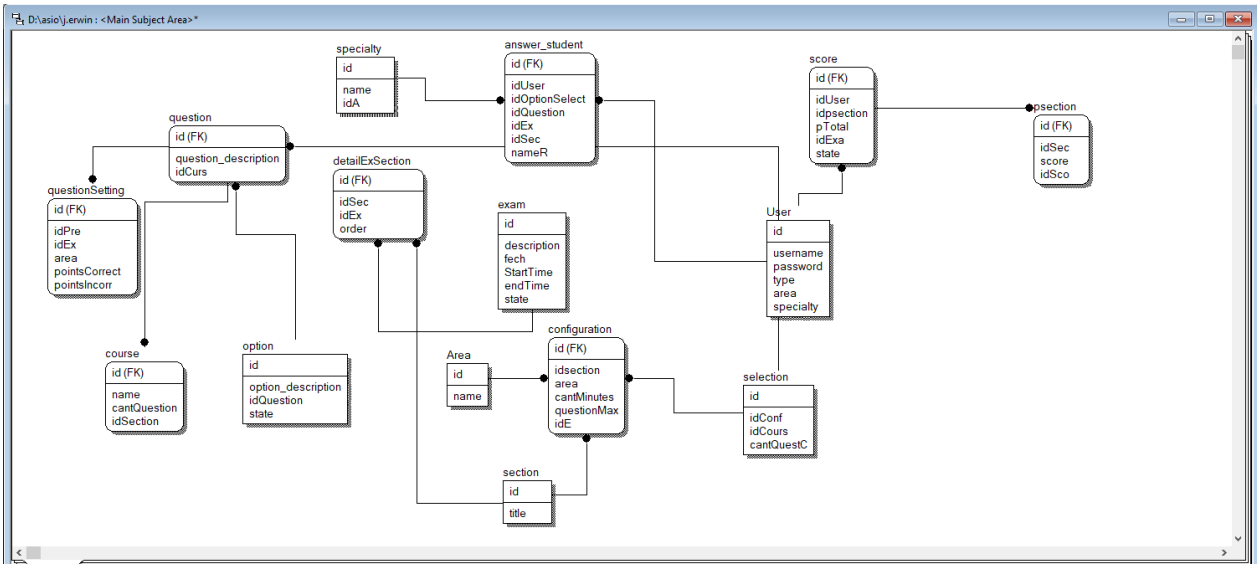


Fig. 4. Database Model.

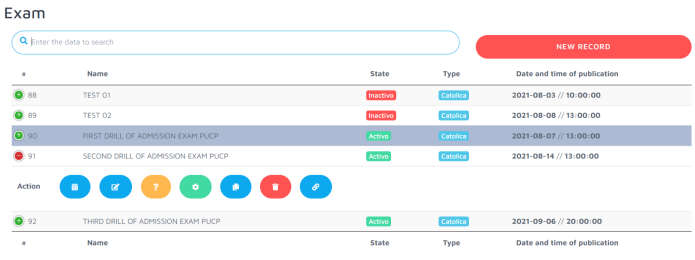


Fig. 5. Exam Management.

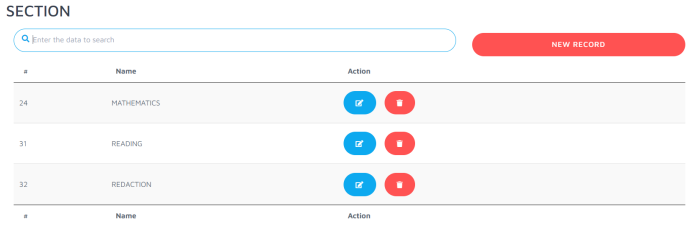


Fig. 6. Section Management.

Fig. 7. Course Management.

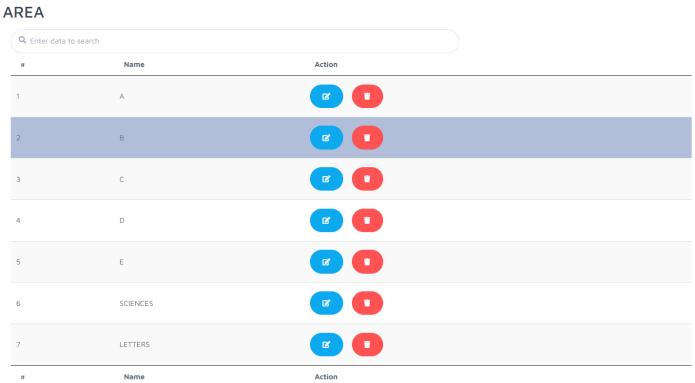


Fig. 8. Area.

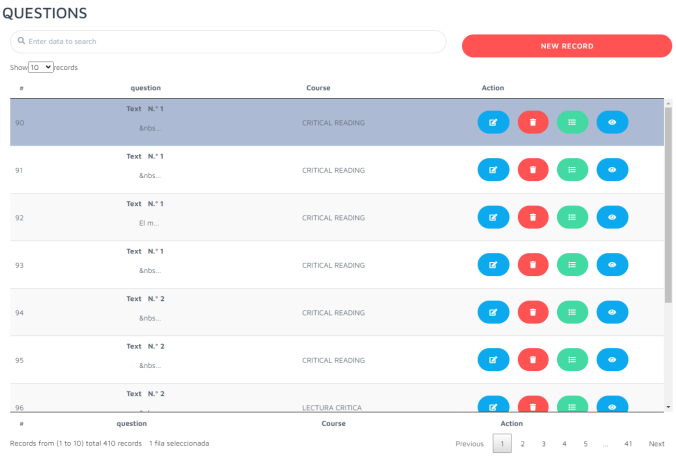


Fig. 9. Question Management.

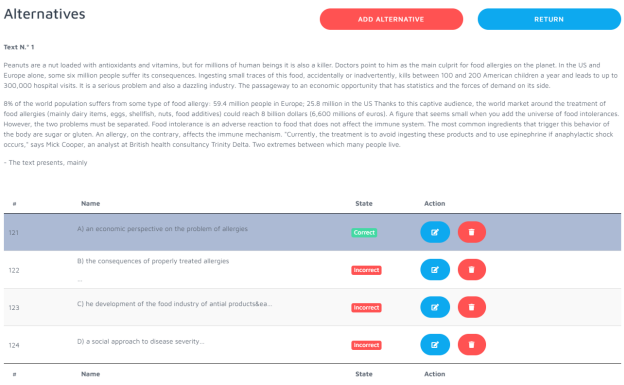


Fig. 10. Alternative Management.

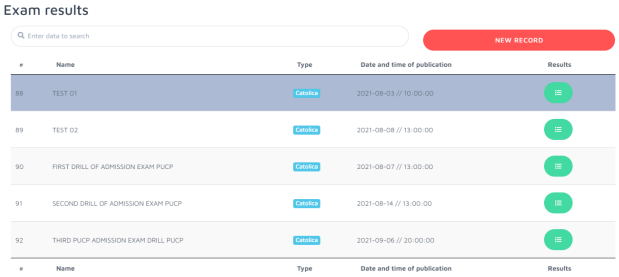


Fig. 11. Exam Results Window.

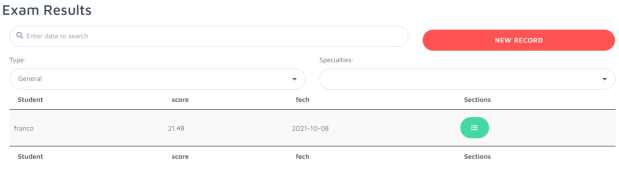


Fig. 12. Student Results.

Fig. 13. Hours Available.

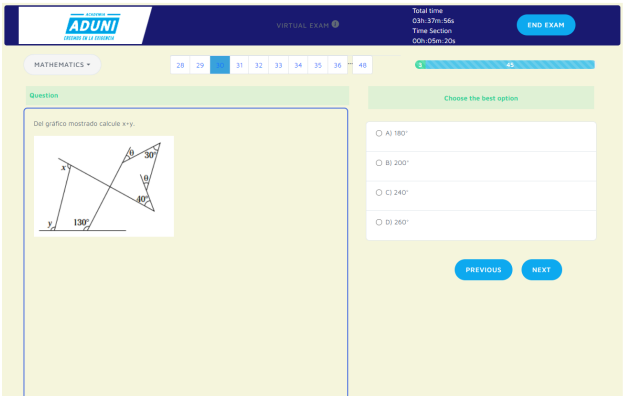


Fig. 14. Virtual Test.

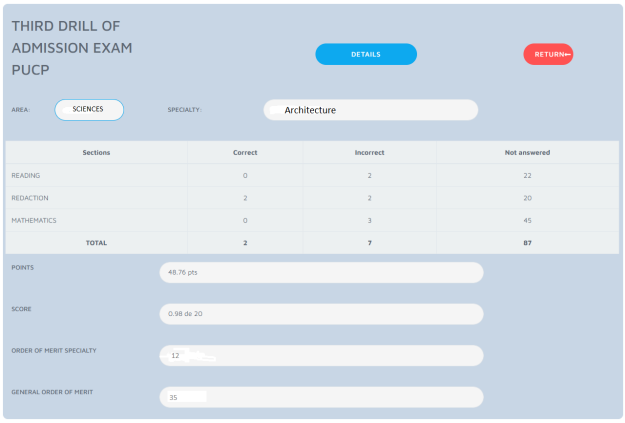


Fig. 15. Results.

emergency due to COVID-19: Challenges in the evaluation processes," *International Journal of Early Childhood Special Education*, vol. 12, no. 1, pp. 467–477, 2020.

[9] Z. I. Almarzooq, M. Lopes, and A. Kochar, "Virtual Learning During the COVID-19 Pandemic: A Disruptive Technology in Graduate Medical Education," *Journal of the American College of Cardiology*, vol. 75, no. 20, pp. 2635–2638, 2020.

[10] O. V. Bondarenko, O. V. Pakhomova, and W. Lewoniewski, "The didactic potential of virtual information educational environment as a tool of geography students training," *CEUR Workshop Proceedings*, vol. 2547, pp. 13–23, 2020.

[11] A. N. Grigorev, V. V. Fadeeva, A. C. Kodoeva, N. V. Bichan, and I. V. Dzyadevich, "Using the grade-rating system assessing the knowledge when teaching informational and legal disciplines in universities of law enforcement agencies," *SHS Web of Conferences*, vol. 108, p. 05004, 2021.

[12] M. Thomas, "Virtual Teaching in the Time of COVID-19: Rethinking Our weird Pedagogical Commitments to Teacher Education," *Frontiers in Education*, vol. 5, no. December, 2020.

[13] S. K. Shahzad, J. Hussain, N. Sadaf, S. Sarwat, U. Ghani, and R. Saleem, "Impact of Virtual Teaching on ESL Learners39; Attitudes under Covid-19 Circumstances at Post Graduate Level in Pakistan," *English Language Teaching*, vol. 13, no. 9, p. 1, 2020.

[14] D. A. Barcelos Bica and C. A. G. D. Silva, "Learning process of agile scrum methodology with lego blocks in interactive academic games: Viewpoint of students," *Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 15, no. 2, pp. 95–104, 2020.

[15] J. Vogelzang, W. F. Admiraal, and J. H. Van Driel, "Effects of Scrum methodology on students' critical scientific literacy: The case of Green Chemistry," *Chemistry Education Research and Practice*, vol. 21, no. 3, pp. 940–952, 2020.

[16] 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

[17] P. N. Astya, Galgotias University. School of Computing Science and Engineering, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section. SP/C Joint Chapter, and Institute of Electrical and Electronics Engineers, "Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016) : 29-30 April, 2016," pp. 867–872, 2016.

[18] 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

Design of an Anti-theft Alarm System for Vehicles using IoT

Jorge Arellano-Zubiate, Jheyson Izquierdo-Calongos, Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—Automobiles have become one of the most sought-after targets for criminals due to their worldwide popularity. Crime is reflected in the statistics, which show that over the years, the crime rate of vehicle theft has been on the rise. As part of the fight against this crime, the vehicles come with certain systems incorporated to avoid this type of situations; obtaining many outstanding results. In this research project, a system was developed that allows through the application of the Internet of Things (IoT), the management of software and hardware technologies that allow the user to have access to various actions, such as vehicle location through the global positioning system (GPS), and identification of the offender, through radio frequency identification (RFID), as well as the global system of mobile communications (GSM). The objective of the research is to design a mobile and IoT application to reduce robberies in the department of Lima-Peru, using the scrum methodology. The result obtained is the design of the mobile application, with its anti-theft system, vehicle blocking and notification of unauthorized ignition.**

***Keywords*—Global mobile communications system; global positioning system; internet of things; radio frequency identification; scrum**

I. Introduction

Since its inception, the automobile has played a very important role in the industrial field, which has been demonstrated by a steady increase in its production and sales. Thus, in India, it is observed that the number of vehicles purchased has reached 100 million units, and it is expected that this figure could reach 450 million by 2021. Likewise, in the United States (US) [1], the picture is not much different as the number of people with driver's licenses is much lower than the number of cars.

However, this boom in the production of motor vehicles has brought with it the appearance of people who seek to obtain them illegally, regardless of the fact that some of them have protection systems, which so far have not been sufficient to stop this type of behavior [2]. Now, this increase can be seen reflected in the country of Mexico, where the percentage of vehicle theft has increased by 27.5% during the first half of 2020 [3]. Similarly, in our country the situation does not vary [4], since in the department of Lima the increase in vehicle theft complaints has been increasing from 3% to 27% in the period of time between 2017 and 2019 according to figures provided by the National Institute of Statistics and Informatics (INEI).

Likewise, this type of activity has intensified in the search for ways to avoid the systems with which the vehicles are equipped, since most of them only have GPS location systems [5], so these criminal gangs have implemented a system whereby they block these signals in certain strategic locations, thus gaining an advantage in their actions.

Now, the systems that use IoT technologies provide us with a series of options such as being able to visualize the status of the peripherals that the vehicles have, in this sense,According to the author [6] can know when the headlights are turned on or that the vehicle's engine has been started; all this through notifications to mobile device. Likewise, it is possible to implement systems that, after the vehicle is started, initiate a monitoring process by means of GPS signals, which allow the owner to know the location of the vehicle in real time. Similarly [7], the IoT allows the owner to interact with this system by remotely turning off the engine.

Indeed, the IoT allows systems to provide many more options to suit the needs that users have, for example, the implementation of vehicle security systems that have the ability to monitor by silent video surveillance unauthorized persons inside the vehicle, which not only prevents the theft of the vehicle, but can also identify the person who did it [8]. On the other hand, there are vehicle security systems with complex identification methods that raise the cost of their development and acquisition. However, the IoT allows different devices, such as RFID tags that are low cost [9], can be used in the identification and operation of these systems with which it becomes a viable option for those who have the need to create such systems and have low economic resources.

This research has been developed taking into account the various factors that IoT technology has, in this sense, an anti-vehicle theft system that manages the different actions of identification and action within a vehicle by means of IoT was carried out.

The objective of the research is to design, through a mobile application, an innovative vehicle security system through which it is expected to reduce the percentage of cases of people who suffer the theft of their vehicles in the department of Lima-Peru.

Then, the following points was developed in the paper: in Section II the literature review where analyze those works that served as a basis for the development of the research, in Section III on the methodology used during the development of the project and the steps that were performed, followed by Section IV the results that have been obtained at the end of the development of the project, then in Section V we develop the

theme of discussions where we analyze the similarities with other projects, finally in Section VI the conclusions and future work.

II. Literature Review

In this research a vehicle anti-theft system has been developed by applying IoT technology to control the different actions of the system, in order to reduce the rate of vehicle theft in the area of the department of Lima - Peru. In this sense, it has been taken into account other research works that have implemented technologies, which have been used in this work.

Thus, we see how [6] implements light and sound sensors inside a vehicle in order to know the state in which it is; this information is centralized in an Arduino Mega 2560 device. Then, by means of a 3G-Shield, this information is transmitted to the user for its respective visualization. As a result, the user can remotely visualize the status of the headlights (on, off), the doors (open, closed), the saloon light (on, off) and the vehicle's engine (on, off). In short, the user can become aware of his habits with respect to the state in which he leaves his vehicle after use.

However, [7] addresses the problem of vehicle theft through the implementation of GSM and GPS technologies. Consequently, IoT is used to communicate with the user by notifying him that his vehicle has been forcibly started and also initiates a constant report of the vehicle's location. Then, the user can interact with the system allowing him to turn off the vehicle by sending a text message which will be received by the system and will trigger the vehicle to stop. In conclusion, the user can access this way to a low-cost anti-theft system that allows him to locate his vehicle in real time, without the need to have specialized knowledge in the field of smartphones.

In addition, the author [8] implements a vehicle security system with video surveillance of the driver. Consequently, IoT is applied together with a biometric driver identification system, which aims to prevent unauthorized persons from operating the vehicle. However, the system adds the functionality to identify the person who tries to perform such action by means of real-time video surveillance of the driver's cabin. In short, the vehicle owner can identify through silent video surveillance any person driving his vehicle, all this through a low-cost system.

However, in [9] the author's problem is that security systems are expensive and cannot be implemented in two-wheeled vehicles. For this reason, a low-cost system is developed that implements the recognition of the driver by means of an RFID tag, which allows the system to turn on the vehicle. A GPS location system is also used, which is visualized by means of an application installed in the owner's cell phone. As a result, the system performs the identification of the RFID tag which allows the ignition of the vehicle by means of the key, then the vehicle can be monitored by means of GPS signals sent by the GSM device connected to the system's Arduino. In summary, it is possible to protect two-wheeled vehicles (motorcycles) with a low-cost security system that gives the location and the ability to identify the driver through RFID tags.

On the other hand, [10] mentions that in recent years with the growth of the global and national economy, vehicles have become a necessity for people and at the same time a great loss of money because of theft, criminals use the method of hot wiring to disable the anti-theft system that comes as part of the vehicle, so they designed a system that turns off the engine when the thief starts the vehicle and can capture the image of the offender when starting the vehicle, so they had to use Arduino, relay, GPS and a camera, to develop the system. As part of the results, all the tests were 100% effective. Finally the results showed that the system worked correctly and efficiently.

Finally, the authors provided very interesting research, helping to complement the research conducted, however, in the research implemented a more innovative system, which is helping drivers of vehicles in Lima.

III. Methodology

For the development of this research project, made use of an agile methodology in conjunction with various tools both in the field of software and hardware. In this sense, the project was developed based on the Scrum methodology, which was detailed later, and also mention which are the technological tools that allowed to implement the system in conjunction with the various electronic devices.

A. Scrum

This is known as one of the best methodologies, but it is actually an agile project development framework [11], as shown in Fig. 1 presents each of the stages and the order they take during the application of the same in a project. In addition, it has principles which allow satisfying customer needs, being accessible to changes in requirements, having a collaborative environment between the development team and customers, regular delivery of progress, reflection on how to improve the errors that can be found [12]. The following is a brief description of each of the stages that make up this framework:

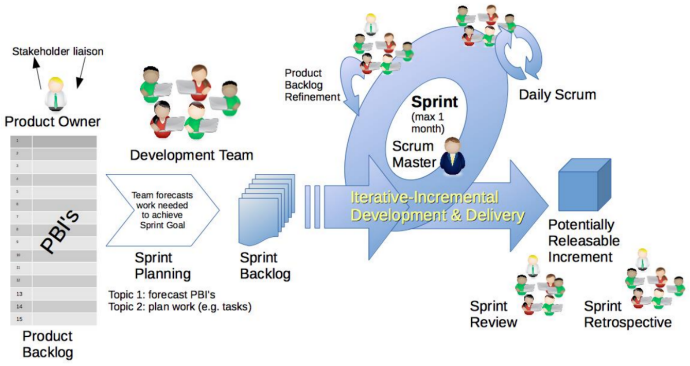


Fig. 1. Scrum Methodology Process [13].

1) Determination of Roles: This is the first stage in which the team as a whole must assign roles within the project development, such as the Product Owner, the Scrum Master and those who was part of the Development Team [14].

2) Planning: At this stage, the team must determine which are the tasks to be developed, in this sense, several meetings are held to discuss each one of them [15]. In this way, a list of the tasks to be performed is obtained, which allows to estimate

each one of them and therefore to have knowledge of the period of completion of the project [16]. At this stage, each of the sprints to be developed is determined, which are considered as mini-projects within the overall project [11].

3) Development: Also, known as the implementation stage, its purpose is to develop each of the Sprints established according to their order of prioritization [14]. It must be taken into account that each of these Sprints must be carried out within the time frame that has been previously established [15].

4) Review and Retrospective: Upon completion of each Sprint, the project development team conducts an analysis to determine whether the project meets the requirements for approval [14]. Then, the team should determine which were the highlights and also those in which difficulties or failures were found, in order to help the team to improve.

B. Software Tools

As part of the development of the research project, the software technologies necessary for its correct development were taken into account.

1) Kotlin: To perform the programming of the application we make use of this system that is compatible with Android, in addition [17], this technology since it was created by JetBrains and Kotlin in 2010 has been characterized by its compatibility, performance and its learning curve.

2) Firebase: It is a service provided by Google through the cloud through which you can perform instant messaging, user authentication, real-time database, and many other functions [18]. However, it must be taken into account that the initial configuration of this service must be carefully carried out, so that all its functions can be used normally [18]. Fig. 2 shows how this medium allows data to be centralized and then distributed to different devices, as well as allowing several actions to be performed simultaneously [16].

3) Moqups: In order to obtain high quality prototypes that are understandable to the naked eye, the Moqups tool is used, which uses some of the services provided by Firebase, which is extremely convenient for compatibility in the development of the project.

C. Hardware Tools

In this section we mention which devices and hardware technologies have been used in the development of this research project.

1) RFID Reader: Within the concept of communication that is established in RFID technology we find the tag reader which fulfills the functions of transmitter and receiver, in such a way [18], the transmitter radiates electromagnetic waves which allow to feed the tags at the time of wanting to detect them and emits the information that was required.

2) RFID Tag: It is one of the emblematic elements of RFID technology due to its production characteristics, which include low production cost, compatibility in manufacturing materials and electromagnetic resistance to the environment in which it is located. In addition [18], there are different types and sizes, which are chosen primarily for their detection capability, which can range from a few centimeters to 25 meters using microwave or UHF antennas.

3) Arduino: The Arduino is an electronic device in which various modules (GSM, GPS, RFID reader) and sensors (light, sound) can be connected to collect information and send on/off signals as needed [6].

4) GPS Module: This is a module whose main functionality is to provide the coordinates in which it is located, this can be sent to a system with Arduino device with which you can track it.

5) GSM Module: This GSM module has the characteristic of providing a means by which communication can be established with it, so, as this GSM technology has evolved, it has a greater coverage which allows a better availability in communication [1].

6) Camera Module: The camera module used for this project is the ESP-32 model, since it has features that fit the needs of the project such as: low acquisition cost, live video transmission, and ease of adaptation to the use of android applications for video transmission via the Internet.

D. Methodology Development

1) Determination of Roles: In this first meeting the roles were determined according to the capabilities of each of the members, Table I shows the different roles and the persons responsible for them.

TABLE I. DETERMINATION OF ROLES

Role	Responsibility
Scrum Master	Laberiano Andrade Arenas
Product Owner	Laberiano Andrade Arenas
Development Team	Jorge Arellano Zubiate Jheyson Izquierdo Calongos

2) Planning: Now, at this stage, the project development team determined which Sprints are required according to the project needs:

- Sprint 1: Implementation of the GSM system that allows the connection between the device and the mobile application. Table II shows the user stories that was developed in this Sprint.
- Sprint 2: Implementation of the RFID system that allows the detection and notification of forced starts in the mobile application. Table III shows the user stories that was developed in this Sprint.
- Sprint 3: Implementation of the GPS system that allows the location of the vehicle through the mobile application. Table IV shows the user stories that was developed in this Sprint.
- Sprint 4: Implementation of the camera system in the vehicle cabin for driver identification through the mobile application. Table V shows the user stories that was developed in this Sprint.

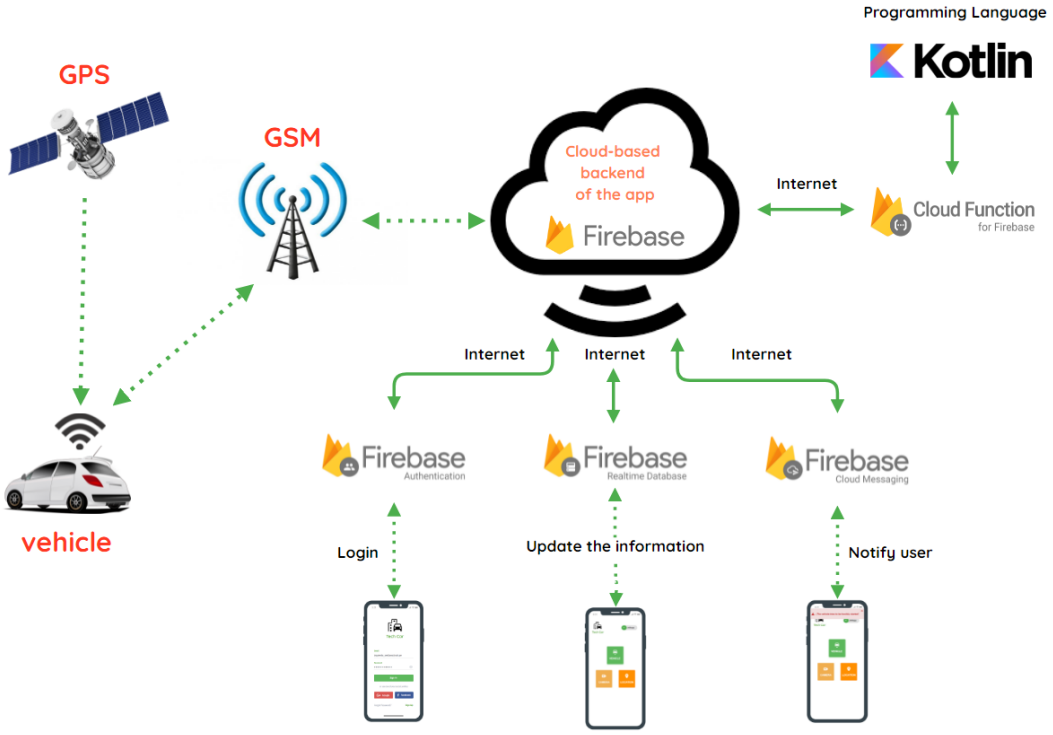


Fig. 2. Diagram of Firebase Operation.

TABLE II. SPRINT 1 USER STORIES

User Stories
I as an administrator want to integrate a GSM device to obtain remote vehicle information.
I as administrator want to implement a server for remote connection to the vehicle.
I as an administrator want a status notification in the app to know the status of the vehicle.
I as an administrator want a registration module for the user to register within the system.
I as an administrator want a login module for the user to validate his login to the system.
I as an administrator want a linking module for the user to get the vehicle information.

TABLE III. SPRINT 2 USER STORIES

User Stories
I as the administrator want to integrate an RFID tag reader to identify the tags.
I as an administrator want an electrical fluid blocking system to block the ignition of the vehicle.
I as the administrator want to register the owner's RFID tag to unlock the vehicle's ignition system.
I as the administrator want an RFID notification to let the user know when the vehicle was forced started.
I as an administrator want a main module for the user to have access to the main options.

TABLE IV. SPRINT 3 USER STORIES

User Stories
I as the administrator want to integrate a GPS device to detect the location of the vehicle.
I as the administrator want to implement Google Maps to facilitate the location of the vehicle.
I as an administrator want a record of the locations so that the user can know the last location of the vehicle.
I as an administrator want a location module in the application to display the vehicle location information.

TABLE V. SPRINT 4 USER STORIES

User Stories
I as an administrator want to implement a camera module to visualize the vehicle cabin.
I as the administrator want a video log so that the user can view who entered the vehicle.
I as an administrator want a usage acceptance module for the user to grant permission to use the vehicle.
I as an administrator want a video module in the application so that the user can view the video in real time.

3) Development:

- Sprint 1: As shown in Fig. 3, the different electronic devices that are part of the circuit of the research project were connected. Each of these devices fulfills

a function . as it allows the IoT concept to be applied. In addition, in this Sprint priority is given to obtaining information through the GSM module, which allowed us to make the remote connection between the Arduino and the mobile application. In this Sprint, the

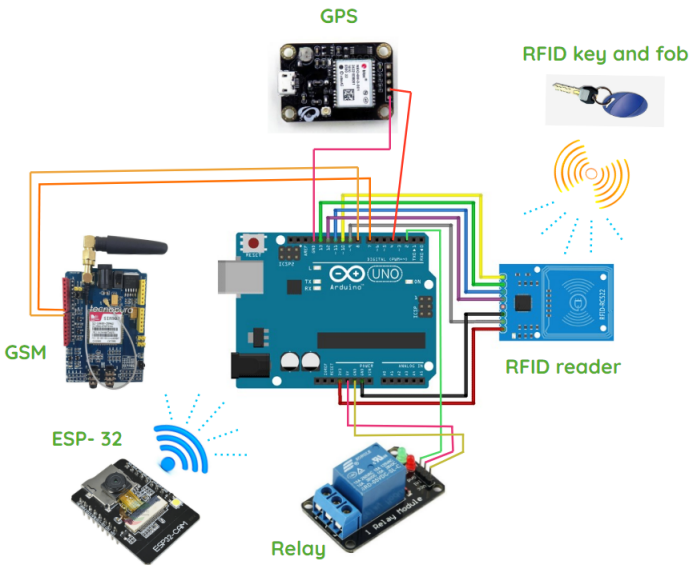


Fig. 3. Physical Diagram of the Anti-theft System.

remote connection between the Arduino circuit and the mobile application was implemented by means of a server that serves as an intermediary between the two. In this sense, the GSM implementation was carried out achieving the remote connection as shown in Fig. 4, where you can see the communication between both points through the server, which allowed us to obtain information on the status of the vehicle and perform the respective actions remotely through the cell phone.

In order for the user to communicate with the system, he/she must be previously registered with a user name and password. Fig. 5 shows the prototype of the form by which the user registers in the application. After registration, the user can use the login form, as shown in Fig. 6, which with the appropriate credentials allows the user to enter the system. Also, Fig. 7 shows the prototype of the main menu through which the user can access the different functions of the system.

- Sprint 2: In this Sprint, the implementation of the RFID tag reading module within the system was carried out. By means of this implementation, the RFID tags will be read and identified in order to unblock or block the electric fluid that allows the ignition and operation of the vehicle's engine. In Fig. 8 the system allows the user to appreciate which is the state (On - Off) in which the vehicle is.
- Sprint 3: In this Sprint, the implementation of the GPS module was carried out, which allows the location of the vehicle. In this sense, as shown in Fig. 9, this module provides the necessary information so that the system can graphically display the location of the

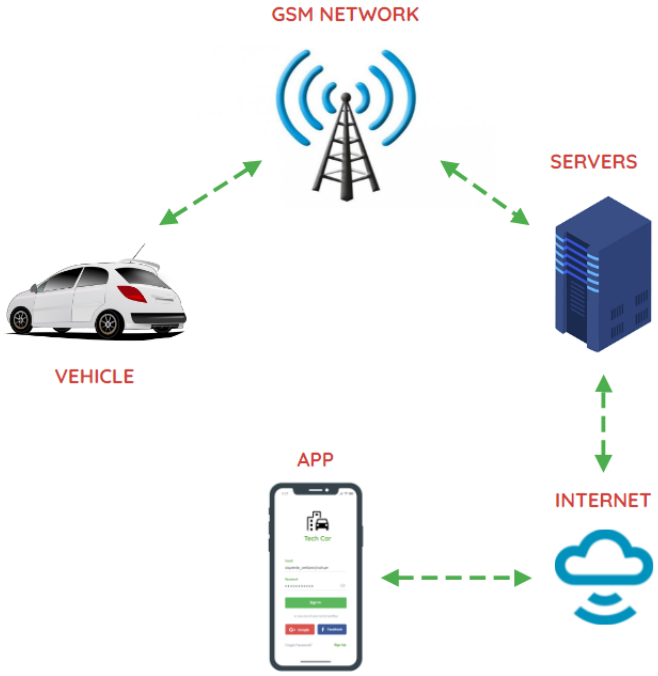


Fig. 4. Diagram of GSM Module Operation.

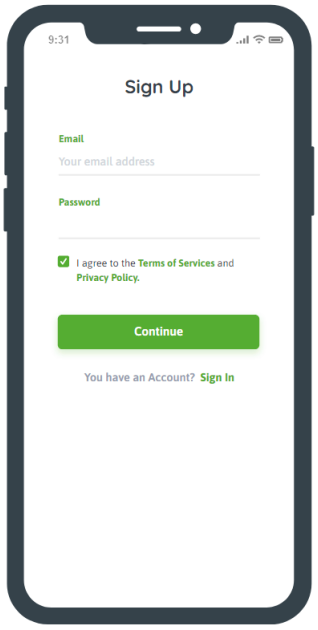


Fig. 5. Registration Prototype.

vehicle in real time, so that the user can know exactly where to locate the vehicle, since it has the coordinates of its location.

- Sprint 4: In this Sprint the implementation and connection of the video camera module is performed, as shown in Fig. 10 it allows the user to visualize the people inside the vehicle in real time. In addition, the options to allow and deny the use of the vehicle

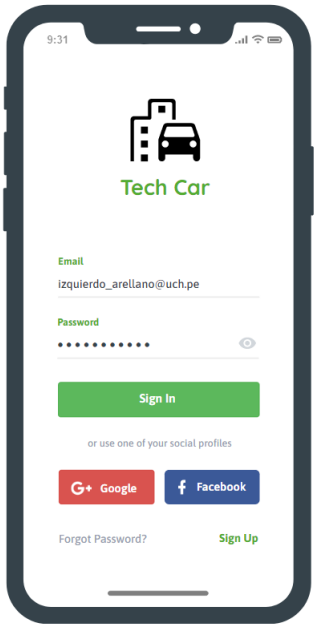


Fig. 6. Login Prototype.

Fig. 7. Prototype of the Main Menu.

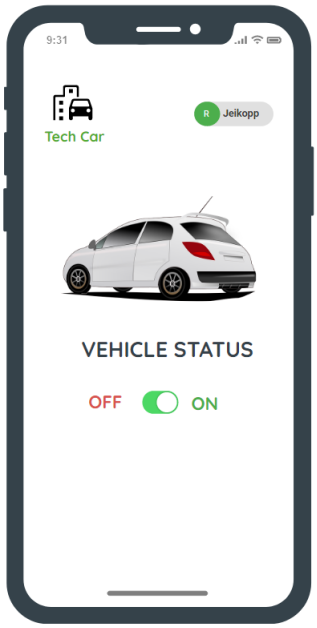


Fig. 8. Vehicle Condition Prototype.

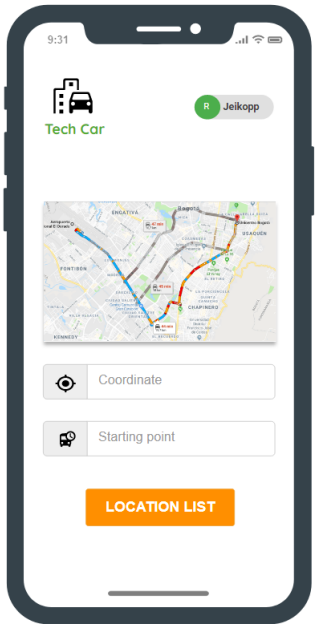


Fig. 9. Vehicle Location Prototype.

according to the user's criteria were implemented in this sprint.

IV. Results

In this section, the analysis of each of the results obtained after the development of each of the Sprints, which are part of the functional structure of this research project, was carried out.

A. System Analysis

Among the results obtained with the development of this research project, we find the flowchart, which is a guide that allows us to know which are the actions taken by the project in various situations.

In this sense, Fig. 11 shows that the system is activated after the vehicle ignition action is performed, then the identification of the RFID tag is performed, which when not identified launches a forced ignition alert. Then, the location of the vehicle is visualized through the use of the GPS device that has

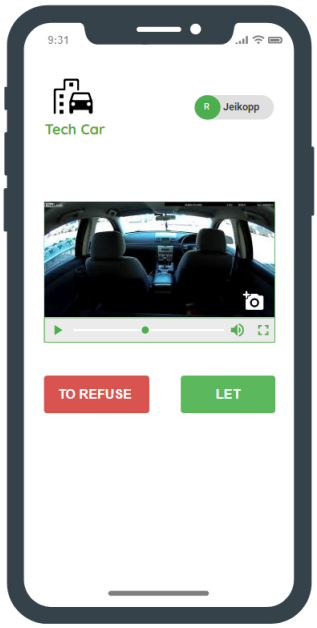


Fig. 10. Prototype of the Visualization of People Inside the Vehicle.

been implemented. In addition, it starts with the transmission of video images, which can be viewed by the user in the application. Finally, there is the option to authorize the user, in case the use of the vehicle is necessary. If the user is not authorized, the vehicle operation is blocked, which is confirmed with a notification.

B. Anti-theft System

1) Notification of Ignition Attempt: As part of the implemented system, RFID tags are used to detect when someone tries to start the vehicle in a forced way. Thus, when someone makes an attempt to start the vehicle without the RFID tag, the system immediately detects the non-existence of the RFID tag reading so it launches a notification which will be displayed on the vehicle owner's computer.

Therefore, after the non-existence of the tag has been detected at the time of the vehicle ignition attempt, an alert is launched in Fig. 12 shows the notification that is displayed by the user from the start menu of the application, which provides access to the functions of GPS location or access to the vehicle's camera module.

2) Notification of Unauthorized Ignition: Now, when the vehicle is started by force or unauthorized by a person who does not have the RFID tag, the system detects that the engine is running and sends an alert to the user in Fig.13 shows the respective notification from the start menu giving the user access to the GPS and Camera module options.

3) Vehicle Locking: When the vehicle has been forcibly started, the user can, as shown in Fig. 14, access the vehicle's camera module and view the real-time video feed from the vehicle's cabin. In addition, it is possible to visualize the person who started the vehicle, so the user has access to the options to allow the use of the vehicle in the case of an authorized person or to block the use of the vehicle.

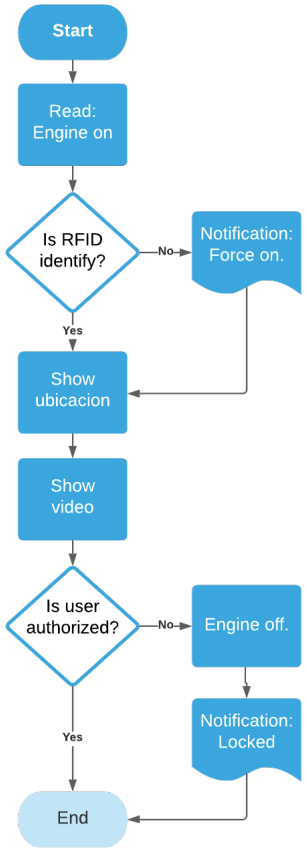


Fig. 11. System Flow Diagram.

Fig. 12. Alert from the Main Menu.

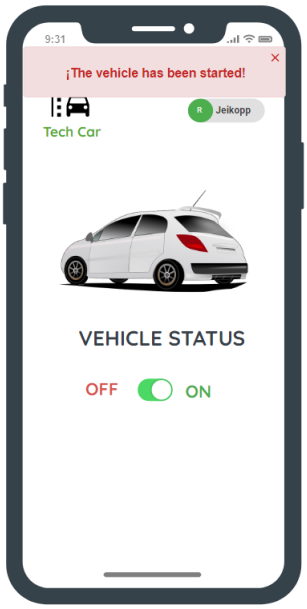


Fig. 13. Vehicle Status Alert.

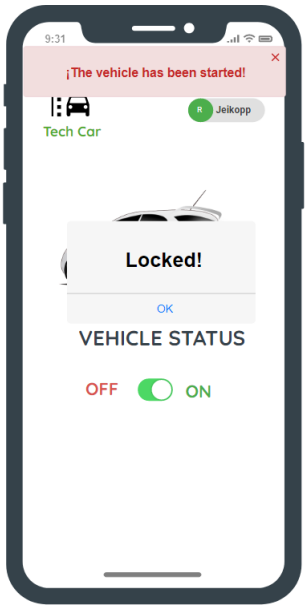


Fig. 15. Locking of the Vehicle.

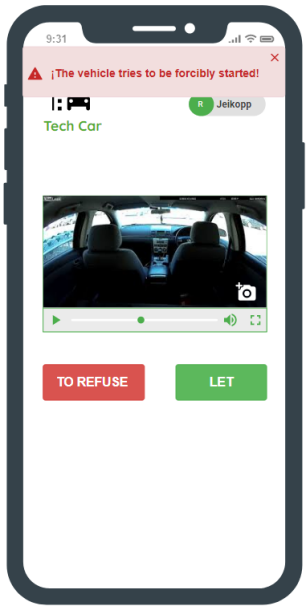


Fig. 14. Alert from the Camera Module.

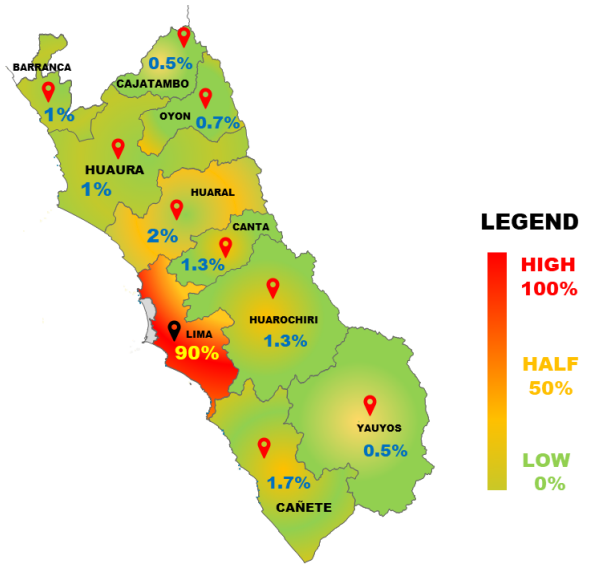


Fig. 16. Heat Map of the Crime Rate in the Department of Lima - Peru.

In fact, when the user determines that the person inside the vehicle is not authorized, he has an option in the system that allows him to block the vehicle. Thus, Fig. 15 shows that when the user blocks the vehicle correctly, he will be notified with a message on the device.

C. Analysis of Robberies in Lima

This research was developed in response to the problem of vehicle theft in the department of Lima, Peru. As shown in Fig. 16, the department of Lima has a vehicle theft rate of no more than 3% in its different provinces; however, the department of Metropolitan Lima has the highest rate, reaching 90%.

The purpose of this research project is to progressively reduce the rate of vehicle theft in the department of Metropolitan Lima. In Fig. 17 we can see that this rate is reduced over time because users have access to the location of their vehicle in real time, the means to visualize who is inside their vehicle, an RFID tag identification system, and the possibility of remotely blocking the vehicle through the IoT application.

This project, being implemented with IoT, allows the user to access the information of those devices with which it has, in addition, it offers the option of interacting with them by giving remote indications, which greatly facilitates the remote control of the vehicle.

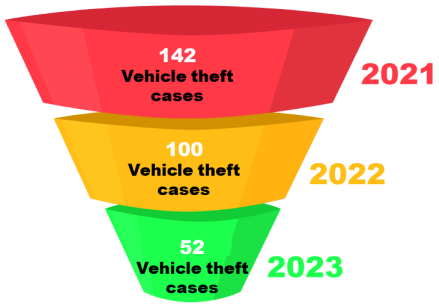


Fig. 17. Crime Projection in Lima - Peru.

V. Discussion

The present research has certain differences in comparison with other research papers dealing with the same technologies.

In fact, he took the trouble to [6] as a reference to the management of the information provided by the sensors to the Arduino device and then transmitted to a mobile application. On the other hand, in our research project the same principle was used to detect the information from the different devices that are connected to the Arduino and to be able to perform actions with them remotely.

In this regard, in comparison with the system developed by [7] which uses the vehicle's coordinates to be displayed by means of text messages on the owner's mobile device, in our system we use these coordinates through an API which allows the visualization by means of a map, allowing a better visualization of the vehicle's location by the user.

Likewise, in the work developed by [1] the system displays the location of the vehicle in the same way with the possibility of sending messages as a response and that these provoke an action in the system. However, in our system the same actions are shown with the difference that they are not a text message, but are commands that are executed after triggering them through the interface which is safer and easier for the user because it can not always have in mind the commands to send them by a text message. It is suggested to carry out research on the topic of RFID implant application such as [19] indicates that this technology can be used for the identification and subsequent operation of devices, thus avoiding the use of tags that can be easily stolen or lost by the user. It is also recommended to study the work of [20] which explains in detail how RFID implant technology works and the different applications that can be performed with it. Therefore, it is recommended to involve a Systems Engineer, an Electronic Engineer, an Automotive Technician and a General Practitioner, in order to carry out a multidisciplinary research.

VI. Conclusion and Future Work

In conclusion, this research allows, through the fusion of different technologies together with the IoT, the reduction of crime levels in vehicle theft, which favors all those who need an innovative system that offers various options in the management of this type of situations. The agile methodology with its scrum framework, allowed to have a more holistic outlook for the design of the prototypes. As a future stage to be developed within our system, it is suggested to develop a history of the locations where the vehicle has been the victim of attacks to obtain data from the places with the highest rate of vehicle theft attacks. There were certain limitations to the analysis that were beyond the scope of systems engineering. In this sense, it is suggested to carry out research in an interdisciplinary, multidisciplinary way, that is, with sociologists, electronic engineers, among others.

Acknowledgment

Thank the University of Sciences and Humanities and its Research Institute for their valuable support and financial contribution in the research carried out.

References

[1] K. Mukherjee, "Anti-theft vehicle tracking and immobilization system," *2014 International Conference on Power, Control and Embedded Systems, ICPCES 2014*, pp. 1–4, 2014.

[2] A. T. Noman, S. Hossain, S. Islam, M. E. Islam, N. Ahmed, and M. A. Mahmud Chowdhury, "Design and implementation of microcontroller based anti-theft vehicle security system using GPS, GSM and RFID," *4th International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2018*, pp. 97–101, 2019.

[3] A. Alonso Berbotto and S. Chainey, "Theft of oil from pipelines: an examination of its crime commission in mexico using crime script analysis," *Global Crime*, pp. 1–23, 2021.

[4] INEI, "Complaints for theft of vehicles, according to department," 2019.

[5] C. Periodical, "Criminal gangs have stolen 88 4x4 trucks, double cab, so far this year." 2021. [Online]. Available: https://elcomercio.pe/lima/policiales/bandas-criminales-han-robado-88-camionetas-4x4-doble-cabina-en-lo-que-va-del-ano-video-nndc-noticia/

[6] S. Ruengittinun, J. Paisalwongcharoen, and C. Watcharajindasakul, "IoT solution for bad habit of car security," *Ubi-Media 2017 - Proceedings of the 10th International Conference on Ubi-Media Computing and Workshops with the 4th International Workshop on Advanced E-Learning and the 1st International Workshop on Multimedia and IoT: Networks, Systems and Applications*, pp. 7–10, 2017.

[7] S. Singh and P. Kumari, "Automatic Car Theft Detection System Based on GPS and GSM Technology," *International Journal of Trend in Scientific Research and Development*, vol. Volume-3, no. Issue-4, pp. 689–692, 2019.

[8] M. R. Pawar and I. Rizvi, "IoT Based Embedded System for Vehicle Security and Driver Surveillance," *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, no. Icicct, pp. 466–470, 2018.

[9] S. Mahendra, M. Sathiyanarayanan, and R. B. Vasu, "Smart security system for businesses using internet of things (iot)," *Proceedings of the 2nd International Conference on Green Computing and Internet of Things, ICGCIoT 2018*, pp. 424–429, 2018.

[10] K. J. P. Ortiz, M. N. T. Calicdan, R. P. Ona, and R. F. H. Torres, "GSM-Based Automobile Ignition Stopping and GPS Tracking with Thief Image Capturing," *Proceedings of 2019 2nd World Symposium on Communication Engineering, WSCE 2019*, pp. 107–111, 2019.

[11] 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, pp. 377–382, 2020.

[12] F. Victor Temitayo, A. BADRU, and N. AJAYI, "Adopting Scrum as an Agile Approach in Distributed Software Development: A Review of Literature," *University of KwaZulu-Natal South Africa*, p. 5, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8016173/

[13] K. Schmitz, "A three cohort study of role-play instruction for agile project management," *Journal of Information Systems Education*, vol. 29, no. 2, pp. 93–104, 2018.

[14] 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, pp. 437–444, 2021.

[15] 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, pp. 672–678, 2021.

[16] 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, 2021.

[17] W. J. Li, C. Yen, Y. S. Lin, S. C. Tung, and S. M. Huang, "JustIoT Internet of Things based on the Firebase real-time database," *Proceedings - 2018 IEEE International Conference on Smart Manufacturing, Industrial and Logistics Engineering, SMILE 2018*, vol. 2018-January, pp. 43–47, 2018.

[18] Y. Duroc and S. Tedjini, "RFID: A key technology for Humanity," *Comptes Rendus Physique*, vol. 19, no. 1-2, pp. 64–71, 2018. [Online]. Available: https://doi.org/10.1016/j.crhy.2018.01.003

[19] E. G. A. Khalil and A. S. A. Osman, "A novel method for patients identification in emergency cases using RFID based Radio technology," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 468–471, 2019.

[20] N. Boella, D. Gîrju, and I. Gurviciute, "To chip or not to chip? determinants of human rfid implant adoption by potential consumers in sweden and the influence of the widespread adoption of rfid implants on the marketing mix," 2019, student Paper.

Framework and Method for Measurement of Particulate Matter Concentration using Low Cost Sensors

Shree Vidya Gurudath*[1]
Ramaiah Institute of Technology
MSR Nagar, Bangalore, India

Krishna Raj P M[2]
Ramaiah Institute of Technology
MSR Nagar, Bangalore, India

Srinivasa K G[3]
NITTTR
Chandigarh, India

***Abstract*—Rapid urbanisation and infrastructure shortcomings leading to heavy traffic, heavy construction activities are major contributors to emission of particulate matter into the ambient atmosphere. This is especially true in developing countries, such as India and China. There have been numerous attempts from government authorities and civic agencies to curtail pollution, but these efforts have been in vain. Cities like Beijing, New Delhi suffer from extremely unhealthy air quality during multiple months of the year. Hence, the onus of keeping oneself safe from extreme affects of air pollution falls on the individual. The following study presents a method and framework to measure particulate matter (PM2.5) concentration using low cost sensors, and infer patterns from the data collected. The study uses a SDS011 high precision laser PM2.5 detector module combined with a raspberry pi, which communicates the measurements through *message queueing telemetry transport* (MQTT) protocol to a ponte server which inturn persists the data into a MongoDB, which can be consumed by algorithms for further analysis. For example, the data obtained from the sensors can be fused with data from measurement stations and geographical land use information to estimate dense spatio-temporal pollution maps which is the basis for computing individual exposure to pollutants.**

***Keywords*—Air pollution; low cost sensor; optical dust sensors; particulate matter; MQTT; ponte**

I. Introduction

Particulate matter is a major source of air pollution across the world. Exposure to particulate matter can cause a multitude of problems to individuals including higher risk of hypertension Prabhakaran et al. [1], lung cancer Ciabattini et al. [2], cardio-vascular disease Jaafari et al. [3]. The affects of pollution are more prominent in the most vulnerable subset of population such as pregnant women Tapia et al., Zhu et al. [4, 5], infants Zhou et al., Rivera et al. [6, 7], and the elderly Wang et al., Han et al. [8, 9].

Recently, there have been major strides in quantifying the exposure of individuals to pollutants. Government agencies report air quality in the form of *Air Quality Index* (AQI), which signifies short term effects of pollutants in the atmosphere. Additionally, there have been numerous studies exploring the use of low cost and mobile sensors to measure individual exposure in an efficient manner. Karagulian et al. [10] provides a comprehensive review of various studies for measurement of different pollutants using low cost sensors.

There have been a few studies where the sensors are mounted on a moving vehicle, and pollution mapping is done along the path of the vehicle. DeSouza et al. [11] mounted Alphasense OPC-N2 sensors on garbage trucks to map out air quality in Cambridge, MA. The study used the collected data to identify clusters signifying pollutant hotspots, and explored techniques to generalise the measurements across the entire traversed route.

Low cost sensors are an efficient way for measuring individual exposure to pollutants. Mahajan and Kumar [12] explored the usage of low cost sensors to quantify individual exposure. The study used a PMS5003 sensor which published data using an ESP8255 wifi module. Similar to this study, calibration of the PMS5003 sensor was performed by the colocation technique with a GRIMM EDM 107 dust monitor. Motlagh et al. [13] provide a vision for the use of low cost sensors for dense air quality monitoring and the study also documents an example implementation in the city of Helsinki, Finland.

Chen et al. [14] used a SDL-607 to measure school students' exposure to PM2.5. The sensor measures particulate matter concentration by means of laser scattering using the principle of nephelometry. The data was stored in internal memory and manually transferred to a computer. Candia et al. [15] propose a system and framework for using low cost sensor networks for air quality monitoring. The study used Nova SDS011, SDS021, and SHINYEI PPD42 sensors connected to an Arduino Mega, with an ESP8266 wifi module and a LoRaWAN module in the absence of a wifi network.

Most of these studies concentrate on the efficacy of the sensors, but there are not many studies which propose a comprehensive framework for acquisition and management of pollution data effectively. There are some platforms available which have been used, for example Schneider et al. [16] used *AQMesh* (AQMesh [17]) platform to map out urban pollution in Oslo, Norway. AQMesh provides a proprietary solution for data management which includes web, API and other modes of access. Lim et al. [18] used *AirCasting* (HabitatMap [19]), an opensource data visualisation platform, to map out pollution in Seoul, South Korea.

In the following sections, a detailed description of the proposed framework is provided. Firstly, the sensors utilized and their internal components are described. Section II-B describes the calibration process for the sensors to avoid drift

and inaccuracies in the reading. Finally, Section II-C describes the overall data acquisition process, including the edge devices, protocol and server infrastructure involved.

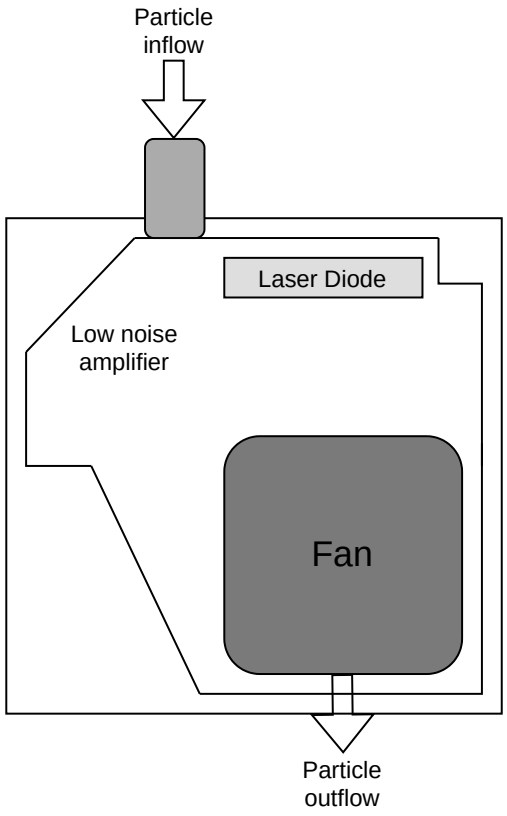
II. Measurement Framework

A. Sensors

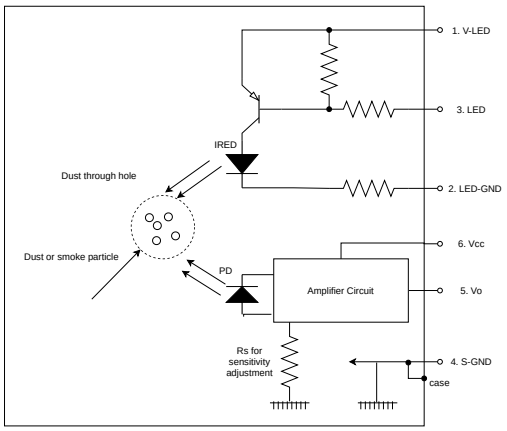
During the study, two sensors were evaluated viz., a Sharp GP2Y1010AU0F optical sensor and a SDS011 laser sensor. Both the GP2Y1010AU0F and the SDS011 sensors use light scattering principle to count the number of particles in the air sample. While the Sharp sensor uses infra-red light, the SDS011 sensor uses laser scattering. Fig. 4 shows the circuit diagram used to connect the sensor to an Arduino Uno in order to get the particulate matter concentration. The SDS011 sensor is easier to use. It has a UART connector that can be directly connected to raspberry Pi using a USB connection. Another advantage of the SDS011 sensor is that it has an inbuilt fan, in order to ensure uniform distribution of the particles in the measurement air sample.

SDS011 is a dust sensor developed by Nova fitness, which uses laser diffraction principle to measure the concentration of particulate matter in the air sample. 1a shows a schematic diagram of the internal components of the SDS011 sensor. It has a sensor output frequency of 1Hz, with a sensitivity range of $0 - 999.9 \mu g m^{-3}$. The sensor has an inbuilt fan and is the most accurate for the size with a relative error of 10%. The sensor has a working range in 0-70% humidity and $-20 - 50^0 C$.

The SHARP GP2Y1010AU0F is an optical dust sensor similar to the SDS011. b shows the inner circuit diagram of the sensor (Fig. 1). It uses an IR LED and phototransistor to measure the amount of dust in the air sample. One of the salient features of the sensor is that it can detect dust from a single pulse, hence works at much higher frequency of upto 100Hz. The sensor measurements are read as voltage which is proportional to the density of particulate matter in the air sample. The sensitivity of the sensor ranges from $0.35 - 0.65 V/(0.1 \mu g m^{-3})$. The main drawback of the sensor is that it does not have an inbuilt fan, hence it needs to be housed in a mixing chamber with a fan in order to get reliable readings. However, when mounted on a moving vehicle, proper placement can eliminate the need for a mixing chamber.



(a) SDS011 Sensor



(b) GP2Y1010AU0F Sensor

Fig. 1. The SDS011 and GP2Y1010AU0F Sensors.

B. Sensor Calibration

Low cost sensors are very capable of detecting pollutant concentrations effectively. With proper maintenance and calibration, they are suitable for long term usage too. Liu et al. [20] studied the performance of a Plantower PMS1003 sensor for PM2.5 measurement at two locations in Australia and China, over a period of 13 months. Zusman et al. [21] provide a comprehensive guide to calibrating low cost sensors using multiple techniques. In this study, sensor calibration was performed by co-location method.

The calibration for the SHARP GP2Y1010AU0F sensor was done by colocation technique. The sensor was colocated near Jayanagar air quality monitoring station maintained by Karnataka state pollution control board (KSPCB) located at 12.920984 LAT, 77.584908 LONG. Simultaneous readings were taken for a duration of two weeks. Fig. 2 shows the outputs of the two sensors. Pearson correlation between the SHARP sensor measurements vs reference sensor is 0.9121332.

C. Data Acquisition Infrastructure

The data measured from the sensors is published through MQTT to ponte server which provides a MQTT broker. The data is persisted in MongoDB, which is used for offline and realtime data analysis. Fig. 3 shows a Schematic representation

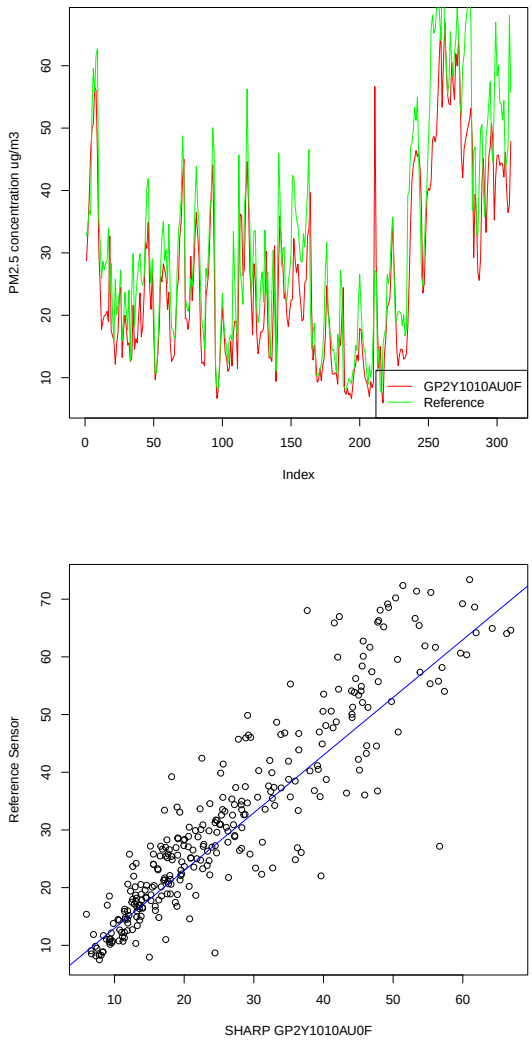


Fig. 2. SHARP Sensor Measurements VS Reference Sensor

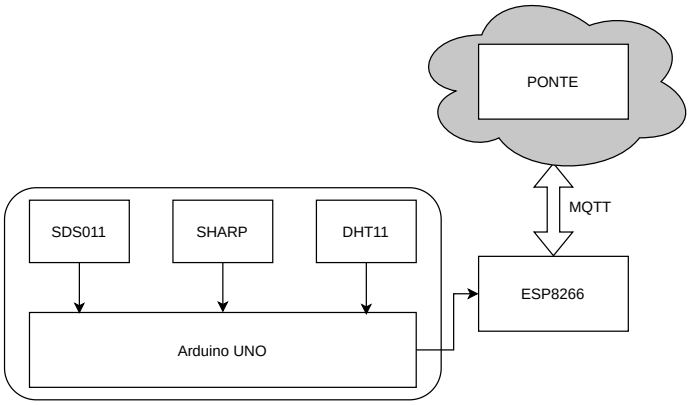


Fig. 3. Schematic Showing the Overall Data Acquisition and Persistence Framework.

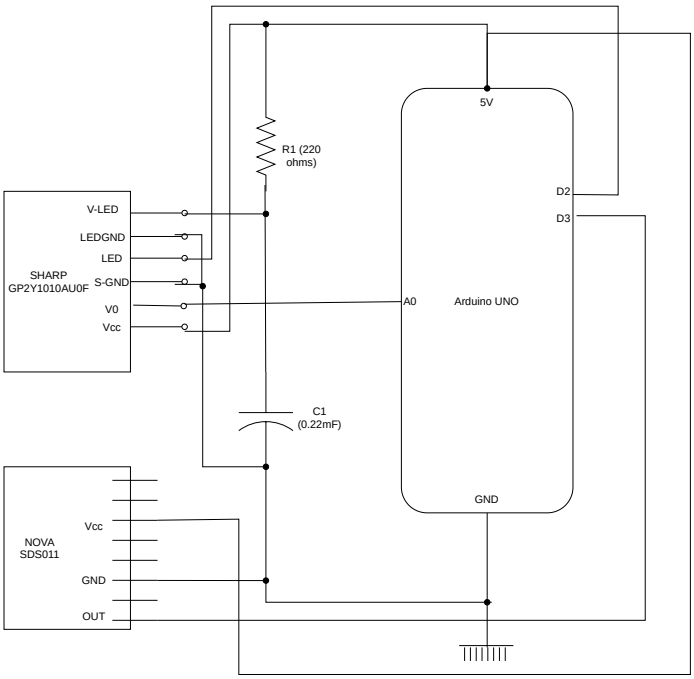


Fig. 4. Circuit Diagram to Connect the Sensors to an Arduino Uno.

of the overall data acquisition and persistence framework.

1) Message Queueing Telemetry Transport: MQTT is a publish-subscribe type messaging protocol which is widely used in IoT (Internet of Things) applications. One of the major advantages of MQTT is that the client applications are very light weight, making it ideal for deployment in small embedded systems such as low power micro-controllers. Additionally, the protocol is scalable, potentially capable of incorporating millions of devices.

Fig. 5 shows an example of an IoT system using MQTT. There are mainly three types of nodes in a MQTT network, viz., the publisher, the subscriber and the MQTT broker. A node can take the role of a subscriber or publisher or both simultaneously. The broker is part of the infrastructure and resides in a server machine. The publisher node publishes data using a specific *topic*, while subscriber nodes which have subscribed for the specific topic, will recieve the message with the data.

MQTT also provides an option for the broker to store messages when a subscriber is down, so that the messages are delivered when the client comes up. Another advantage of a publish/subscriber model over a traditional client servers is that the publisher does not need to be aware of the nature or number of subscribers. Although, the subscribers need to be aware of the message format and the topic of the publisher. MQTT and publish/subscribe model for communication is an efficient model for low power sensor networks such as the one used in the study. It also supports OAUTH based authentication, so that the data is protected from unauthorised access. It also provides for TLS/SSL data encryption.

In this study, each edge node containing an Arduino Uno is assigned a MD5 hash, which acts as an identifier string, and each node publishes the measured PM2.5 concentration, temperature and humidity using a topic which includes the

identifier. For example, 84474b87cafa5f22d8aa2a5b990bfe98 is the identifier for the node which was colocated with the Jayanagar 5th Block monitoring station, and the sensor measurements from the node is published using a topic strings 84474b87cafa5f22d8aa2a5b990bfe98_dht11, 84474b87cafa5f22d8aa2a5b990bfe98_gp2y, and, 84474b87cafa5f22d8aa2a5b990bfe98_sds011 for the temperature and humidity values from the DHT11 sensors, PM2.5 concentrations from the SHARP GP2Y1010AU0F and the Nova SDS011 sensors respectively in the form of a JSON object.

2) Server Infrastructure: The data published by the edge nodes need a MQTT broker to route the messages to the subscribers. Also, there is a need for a scalable database where the data can be persisted for future analysis. Additionally, a HTTP server interface is helpful for a web based dashboard with visualization and realtime data showcasing. In this study, eclipse ponte is used, to accomplish the above requirements. Ponte provides a seamless integration with MongoDB for persisting the data as well as sessions. Additionally, it provides HTTP and CoAP services with REST like APIs.

Fig. 6 shows the schematic architecture of Ponte, with interfaces for MQTT, HTTP and CoAP (Constrained Application Protocol) protocols. The figure also shows the persistence module which connects to MongoDB which not only stores the published data, but also the session information for facilitating seamless communication even in fickle networks.

On the server, different analysis and visualisation applications can provide good insights into the individuals exposure to pollutants. Additionally, a cumulative exposure index (C) can be calculated based on pollutant concentrations exposed to and the amount of time spent. It can be computed as an integral over time, or a simplified summation (C') for discrete readings as shown in equation 1 and 2.

$$C = \int_t d_t dt \tag{1}$$

$$C' = \sum_t d_t \tag{2}$$

where d_t is the pollutant concentration at time t. The measurements, d_t can be geographically distributed across the city.

III. Discussion

Low cost sensors and their efficacy in measuring pollutant concentrations is the subject of multiple studies, hence this study does not go into the details of these factors. However, there are only a few studies which elaborate on the framework and infrastructure involved in collecting the measurements in a database and the interfaces required for analysing the data. This study provides a comprehensive guide to the processes, system and framework necessary for measurement, persistence and analysis of pollution data.

Sensors used in the current study require periodic calibration. Chain calibration technique, where a correctly calibrated low cost sensor co-located with the deployed sensors can be used to avoid sensor drift and other errors in the measurements. Additionally, temperature and humidity play a major role in the accuracy of the measurements. Hojaiji et al. [23] provide a way to calibrate the SHARP sensor to compensate the effects of temperature and humidity. The SDS011 sensor is relatively robust to temperature and humidity variations. In the current study, we smooth the variations caused due to temperature and relative humidity for the SHARP sensor. The study also ignores readings from both sensors when the humidity level rises beyond 70%. Spatial and temporal regression techniques can be used to interpolate the measurements from the monitoring stations. In a future work, a combination of long short term memory (LSTM) networks and land use regression (LUR) in order to build a dense spatio-temporal pollution map is evaluated.

MQTT is a robust transmission protocol which supports the necessary data distribution, encryption and authentication features. Publish/subscribe model of communication is also a very efficient mode for medium scale networks which is necessary for pollutant modeling for a city. However, MQTT transmit cycles are not well equipped for large networks, CoAP could be a feasible alternative when designing a network larger than 250 measurement nodes. Ponte server, used in the study, supports publishing data through CoAP using a simple REST API. This would be the focus of a future study to augment the findings presented in this article.

The framework presented above, is capable of being extendible to different types of sensors, different and more efficient protocols, more stringent and effective security measures, and different persistence technologies. The framework presented above serves as an outline for a more complex and feature rich system implementation.

A. Conclusion and Future Scope

In this paper, a system and framework for measurement, persistence and analysis of pollution data using low cost sensors was proposed. A SHARP GP2Y1010AU0F and a SDS011 dust sensors were used to measure the PM2.5 concentration in ambient atmosphere and the data was sent to a server running ponte using MQTT. The data was persisted in MongoDB which is used for further analysis on the server. Additionally, in order to compensate for the effect of temperature and relative humidity, a DHT11 sensor provided the necessary readings.

The above work is the first step towards a comprehensive estimation and prediction of individual exposure to pollutants and it's deposition in the human lungs. In the future, as a progression of this work, the authors intend to estimate dense spatio-temporal pollution maps with the fusion of data measured using these low cost sensors, and from measurement stations and provide a predictive methodology to quantify lung deposition of these pollutants for an individual.

References

[1] Dorairaj Prabhakaran, Siddhartha Mandal, Bhargav Krishna, Melina Magsumbol, Kalpana Singh, Nikhil Tandon, KM Venkat Narayan, Roopa Shivashankar, Dimple Kondal, Mohammed K Ali, et al. Exposure to particulate matter is associated with elevated blood pressure and incident hypertension in urban india. *Hypertension*, 76(4):1289–1298, 2020.

[2] Marco Ciabattini, Emanuele Rizzello, Francesca Lucaroni, Leonardo Palombi, and Paolo Boffetta. Systematic review and meta-analysis of recent high-quality studies on exposure to particulate matter and risk of lung cancer. *Environmental Research*, page 110440, 2020.

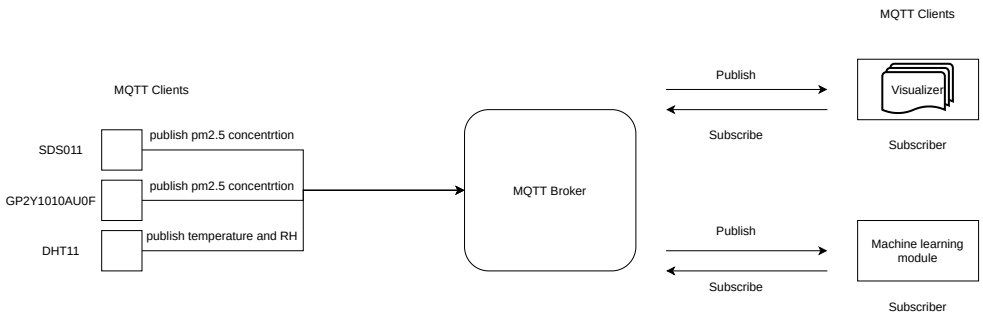


Fig. 5. MQTT Publish Subscribe Mechanism.

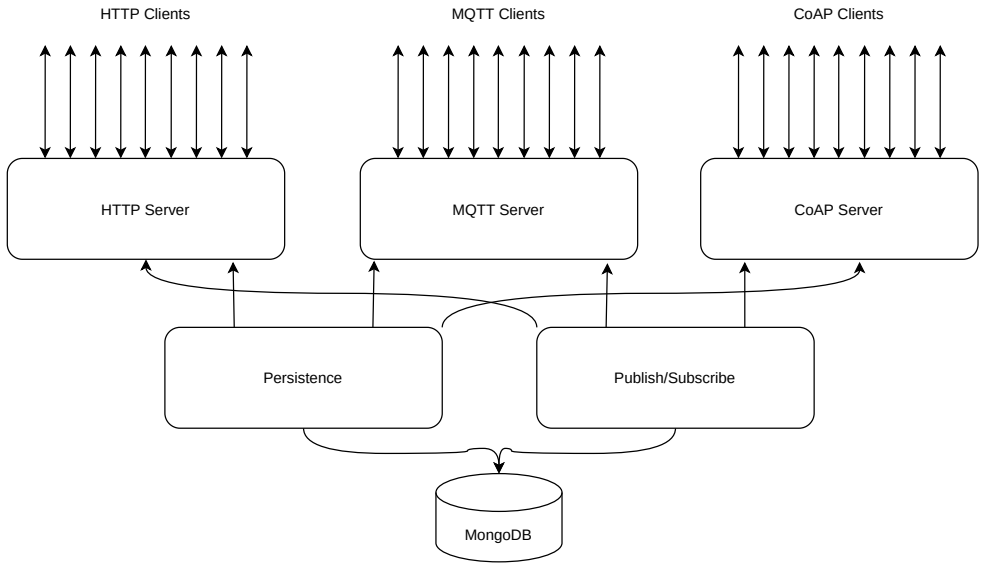


Fig. 6. Eclipse Ponte Architecture (reproduced from Eclipse [22]).

[3] Jalil Jaafari, Kazem Naddafi, Masud Yunesian, Ramin Nabizadeh, Mohammad Sadegh Hassanvand, Mansour Shamsipour, Mohammad Ghanbari Ghozikali, Hamid Reza Shamsollahi, Shahrokh Nazmara, and Kamyar Yaghmaeian. The acute effects of short term exposure to particulate matter from natural and anthropogenic sources on inflammation and coagulation markers in healthy young adults. *Science of The Total Environment*, 735:139417, 2020.

[4] VL Tapia, BV Vasquez, B Vu, Y Liu, K Steenland, and GF Gonzales. Association between maternal exposure to particulate matter (pm 2.5) and adverse pregnancy outcomes in lima, peru. *Journal of exposure science & environmental epidemiology*, 30(4):689–697, 2020.

[5] Wentao Zhu, Huiqiu Zheng, Jieyu Liu, Jiajie Cai, Gechao Wang, Yi Li, Haochong Shen, Jing Yang, Xuemei Wang, Jing Wu, et al. The correlation between chronic exposure to particulate matter and spontaneous abortion: A meta-analysis. *Chemosphere*, page 131802, 2021.

[6] Shuang Zhou, Lizi Lin, Zheng Bao, Tong Meng, Shanshan Wang, Gongbo Chen, Qin Li, Zheng Liu, Heling Bao, Na Han, et al. The association of prenatal exposure to particulate matter with infant growth: A birth cohort study in beijing, china. *Environmental Pollution*, 277: 116792, 2021.

[7] Nadya Y Rivera Rivera, Marcela Tamayo-Ortiz, Adriana Mercado García, Allan C Just, Itai Kloog, Martha Maria Téllez-Rojo, Robert O Wright, Rosalind J Wright, and Maria José Rosa. Prenatal and early life exposure to particulate matter, environmental tobacco smoke and respiratory symptoms in mexican children. *Environmental Research*, 192: 110365, 2021.

[8] Jiaonan Wang, Tiantian Li, Yuebin Lv, Virginia Byers Kraus, Yi Zhang, Chen Mao, Zhaoxue Yin, Wanying Shi, Jinhui Zhou, Tongzhang Zheng, et al. Fine particulate matter and poor cognitive function among chinese older adults: evidence from a community-based, 12-year prospective cohort study. *Environmental health perspectives*, 128(6):067013, 2020.

[9] Changwoo Han, Jongmin Oh, Youn-Hee Lim, Soontae Kim, and Yun-Chul Hong. Long-term exposure to fine particulate matter and development of chronic obstructive pulmonary disease in the elderly. *Environment International*, 143:105895, 2020.

[10] Federico Karagulian, Maurizio Barbiere, Alexander Kotsev, Laurent Spinelle, Michel Gerboles, Friedrich Lagler, Nathalie Redon, Sabine Crunaire, and Annette Borowiak. Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere*, 10(9):506, 2019.

[11] Priyanka DeSouza, Amin Anjomshoaa, Fabio Duarte, Ralph Kahn, Prashant Kumar, and Carlo Ratti. Air quality monitoring using mobile low-cost sensors mounted on trash-trucks: Methods development and lessons learned. *Sustainable Cities and Society*, 60:102239, 2020.

[12] Sachit Mahajan and Prashant Kumar. Evaluation of low-cost sensors for quantitative personal exposure monitoring. *Sustainable Cities and Society*, 57:102076, 2020.

[13] Naser Hossein Motlagh, Eemil Lagerspetz, Petteri Nurmi, Xin Li, Samu Varjonen, Julien Mineraud, Matti Siekkinen, Andrew Rebeiro-Hargrave, Tareq Hussein, Tuukka Petaja, et al. Toward massive scale air quality monitoring. *IEEE Communications Magazine*, 58(2):54–59, 2020.

[14] L-W Antony Chen, John O Olawepo, Felicia Bonanno, Aman Gebreselassie, and Mi Zhang. Schoolchildren's exposure to pm 2.5: a student club–based air quality monitoring campaign using low-cost sensors. *Air Quality, Atmosphere & Health*, 13(5):543–551, 2020.

[15] Agustín Candia, Soledad Natacha Represa, Daniela Giuliani, Miguel Ángel Luengo, Andrés Atilio Porta, and Luis Armando Marrone. Solutions for smartcities: proposal of a monitoring system of air quality based on a lorawan network with low-cost sensors. In *2018*

Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación (CACIDI), pages 1–6. IEEE, 2018.

[16] Philipp Schneider, Nuria Castell, Matthias Vogt, Franck R Dauge, William A Lahoz, and Alena Bartonova. Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment international*, 106:234–247, 2017.

[17] AQMesh. AQMesh. https://www.aqmesh.com/products/aqmesh/, 2011. [Online; accessed 24-Oct-2021].

[18] Chris C Lim, Ho Kim, MJ Ruzmyn Vilcassim, George D Thurston, Terry Gordon, Lung-Chi Chen, Kiyoung Lee, Michael Heimbinder, and Sun-Young Kim. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in seoul, south korea. *Environment international*, 131:105022, 2019.

[19] HabitatMap. AirCasting. https://www.habitatmap.org/aircasting, 2011. [Online; accessed 24-Oct-2021].

[20] Xiaoting Liu, Rohan Jayaratne, Phong Thai, Tara Kuhn, Isak Zing, Bryce Christensen, Riki Lamont, Matthew Dunbabin, Sicong Zhu, Jian Gao, et al. Low-cost sensors as an alternative for long-term air quality monitoring. *Environmental research*, 185:109438, 2020.

[21] Marina Zusman, Cooper S Schumacher, Amanda J Gassett, Elizabeth W Spalt, Elena Austin, Timothy V Larson, Graeme Carvlin, Edmund Seto, Joel D Kaufman, and Lianne Sheppard. Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environment international*, 134:105329, 2020.

[22] Eclipse. PONTE. https://github.com/eclipse/ponte, 2014. [Online; accessed 24-Oct-2021].

[23] Hannaneh Hojaiji, Haik Kalantarian, Alex AT Bui, Christine E King, and Majid Sarrafzadeh. Temperature and humidity calibration of a low-cost wireless dust sensor for real-time monitoring. In *2017 IEEE sensors applications symposium (SAS)*, pages 1–6. IEEE, 2017.

Sustainable Android Malware Detection Scheme using Deep Learning Algorithm

Abdulaziz Alzubaidi
Computer Science Department,
College Computing in Al-Qunfudhah
Umm Al-Qura University

***Abstract*—The immense popularity of smartphones has led to the constant use of these devices for productive and entertainment purposes in daily life. Among the different operating systems, the Android system plays a very important role in the development of mobile technology as it is the most popular operating system. This makes it a target for cyberattack, with severe negative effects in terms of monetary and privacy costs. Thus, this study implements a detection scheme using effective deep learning algorithms (LSTM and MLP). Also, it tests their ability to detect malware by employing private and public datasets, with accuracy of over than 99%.**

***Keywords*—Smartphone security; machine learning; mobile malware; classification; big data**

I. Introduction

The current century has witnessed various inventions such as smartphone devices. These devices are characterized by advanced features in terms of sophisticated operating systems and gigabytes of memory. They are equipped with advanced sensors such as accelerometer, magnetometer, global positioning system (GPS), and biometric sensors. Due to these developed features, the owner of smartphones can perform various activities, such as sending/ receiving electronic emails, performing financial transactions, contacting with others, taking photos and recording videos, etc. [1]. Statistics show the number of smartphone users surpassed 6.4 billion users globally in 2021, and it is expected to reach 7 billion in 2026 [2].

Mobile applications (apps) are implemented to perform one or more tasks. They are available in official markets and third parties. As of 2021, the popular mobile market apps are: Google Play (3,482,452 apps), Apple app store (2,226,823 apps), Windows store (669,000 apps), and Amazon App store (460619 apps) [2]. Smartphones and mobile market apps are targets for attackers in terms of privacy and security. According to [3], the instances of mobile cybercrime surpassed 14.4 million attacks in the second quarter of 2021, 95% of Android devices can be hacked using a simple text message, and 87% are exposed to serious vulnerability. In September of 2020, the Apple store has pulled 40 apps infected by the XcodeGhost botnet attack [4], which indicates that the malware apps might be found in official apps. Therefore, several approaches from academic and industrial fields have been proposed using machine learning algorithms.

Machine learning is an effective method for intelligent detection of malware on smartphones. Malware detection on smartphones is based on feature analysis by static, dynamic, and hybrid methods [2]. The detection and prediction effectiveness of any machine learning algorithm relies on selecting suitable data and understanding malware behavior.

Problems and Motivation

Previous studies have been proposed to detect known and unknown malware samples using public and private datasets. However, most of utilized public datasets are collected between 2010 and 2017, which raise an important question about how they can detect recent implemented malware while the behavior of mobile malware is changeable. Therefore, there is need to collect updated apps . Besides that , understanding the malware patterns and classifying them into families is an effective way to detect unknown malware.

This study proposed a sustainable and cost-effective malware detection scheme with respect to collecting an updated dataset, classifying malware families , and observing malware behavior .

Contributions

This work address the above-mentioned issues related to detecting Android malware. Its contributions are listed below.

1) A dataset is build with 30,000 samples at the present time, plan to expand to be larger and make it publicly available
2) An Android malware detection approach is proposed using machine and deep learning algorithms with respect to sustainability metrics.

The remaining part of this article is organized as follows: Section II summarizes recent studies in this field. Sections III and IV introduce the proposed methodology and describes the dataset used. Finally, the conclusion of this research work is presented in Section V.

II. Prior Research

Detecting Android malware has gained attention last two decades. There are several proposed approaches employed machine and deep learning. Alzubaidi [2] provides a comprehensive survey in terms of static, dynamic and hybrid feature analysis methods using machine learning algorithms, while Qiu et al. [5] review recent deep learning approaches and addressed challenges like the architecture of deep learning. This sections discusses sustainable Android malware detection approaches then summarizes common public datasets.

A. Sustainable Detection Malware Approaches

Onwuzurike et al. [6] introduced a static approach to detect Android malware, which is known as MaMaDroid. This approach is comprised of three phases. The authors first acquired a dataset with size of 43,940 apps (35,493 malware apps from Drebin and 8,447 normal apps from the Google Play store). Second, they extracted the API calls for each feature, then used principal component analysis (PCA) [7] to rank them. Third, they employed random forest (rf) [8], 1-Nearest Neighbor (1-NN), 3-Nearest Neighbor (3-NN) [9], and support vector machine (SVM) [10] to construct their approach. The authors performed two experiments on detecting unknown malware samples and examined the sustainability of their approach. For the first experiment, they achieved accuracy of 0.99. For the second experiment, they examined the sustainability of the samples in terms of one-year and two-year periods and obtained accuracies of 0.87 and 0.75, respectively.

Zhang et al. [11] developed a method to detect Android malware employing sustainability analysis, which they called APIGRAPH. To construct APIGRAPH, a private dataset was collected consisting of 322,594 samples (290,505 normal apps and 32,089 malware apps). The authors extracted API calls, exceptions, and permissions features and employed RF [8], Model Pool, SVM [10], and deep learning neural networks (DNNs) [12] classifiers. They evaluated their developed method using MamaDroid [13], DroidEvolver [14], Drebin-DL [15] and Drebin [16], based on sustainability, and found that the average enhancement for [13], [14], [15] and [16] was 19.2%, 19.6%, 15.6% and 8%, respectively.

Cai and Jenkins [17] investigated how Android app behavior might change over time. For this purpose, the authors used 155 predefined metrics from [18], which are based on general, ICC and security perspectives. They added the extent, frequency, and distribution for the source and sink invocations of sensitive API calls. A dataset was built including 6432 apps (3431 normal apps and 3001 malware apps). In order to evaluate their approach, they constructed two groups of datasets, based on the year, then performed a comparison using the predefined metrics to rank the most informative metrics and found 52 features to be most informative, which were used for further evaluations. They employed RF and obtained an accuracy of 93% while achieving an accuracy of 82% for their sustainability metric.

Cai et al. [19] implemented a scheme called Droidcat to detect Android malware in terms of systemic app-level profiling. The authors created a private dataset consisting of 34343 apps (17,365 normal apps, 16,978 malware apps). Then, they reduced the samples to 271 apps (136 normal apps, 135 malware apps) meeting their requirements. A total of 122 metrics were defined based on structure (method calls, declaring classes, callback), ICC (Intent, carrying data through URI only), and security (distributions of sources, sinks). The authors utilized RF to detect unknown malware samples, and obtained Precision, Recall and F1 of 97.96%, 97.91%, and 97.84%, respectively. Another experiment was carried out to evaluate the sustainability and obtained results with small standard deviations of 1.34-2.38% in terms of F1.

An approach was introduced by Cai [20] called DroidSpan based on behavioral profiling features. The author collected a total of 26382 samples (13,627 normal apps and 12,755 malware apps), then extracted 52 features based on the extent of sensitive access, categorization of sensitive data and operations accessed, and sensitive method-level control flows. Then, the approach employed RF [8], k-NN [9], SVM with both linear and radial basis function kernels [21] , decision trees [22], naïve Bayes [23] with three models (Gaussian, Multinomial, and Bernoulli), AdaBoost [24], Gradient Tree Boosting [25], Extra Trees, and the Bagging classifier, and evaluated DroidSpan in terms of F1-measure, recall and precision. Among all classifiers, RF obtained the best results. Then, another experiment was performed to examine the sustainability based on same-period detection and obtained 92.88%, 92.68% and 92.61% for precision, recall and F1-score.

B. Common Public Datasets

Current approaches rely on two types of data: private and public datasets. This section summarizes common public datasets.

1) Drebin: Arp et al. [16] built a dataset called Drebin between 2010 and 2012 and comprised of 123, 453 samples for normal applications and 5560 abnormal samples from 179 different families. It is available on https://www.sec.cs.tu-bs.de/~danarp/drebin/.

2) AndroZoo: Allix et al. [27] began building a public dataset in the latter part of 2011, called AndroZoo. The authors implemented a crawling tool to examine the application if it had not been downloaded previously. Then, they installed the application, calculated the s SHA256 checksum, and stored the sample. These samples were submitted to VirusTotal, a portal that allows users to analyze potential malware using various antivirus scanners, including several commercial products, such as McAfee, Symantec, and Avast. The total instances of AndroZoo is 10,774,100 samples and available on https://androzoo.uni.lu/.

3) Malgenome: Malgenome was introduced by Yajin and Xuxian [28] with total of 1260 malware samples covering 49 Android malware families between 2010 and 2011. Malgenome is available on http://malgenomeproject.org/.

4) Contagio Mobile Mini-dump: Contagio Mobile Mini-dump was developed by Mila [29]. Data collection involved a blog published in 2008 that allowed researchers to upload and download malware. As of April 2020, 370 malware samples had been collected. We tested this dataset and found some samples cannot be installed. Contagio can be found on http://contagiodump.blogspot.com/?m=1.

5) PRAGuard: PRAGuard is a publicly available dataset introduced by Maiorca et al. [30]. The dataset contains 10479 malware samples from on 50 malware families. PRAGuard is available on http://pralab.diee.unica.it/en/AndroidPRAGuardDataset.

6) Android Malware Dataset (AMD): The AMD was compiled by Wei et al. [31], and consisted of 24,650 samples collected between 2010 and 2016. The dataset includes various types of malware, such as Trojan, backdoor, and ransomware.

TABLE I. SUMMARY OF UTILIZED PUBLIC DATASET

Dataset	# of samples	Years of collection
Drebin [16]	5560	2011 - 2012
CICAndMal2017 [26]	10854	2015 - 20017
AndroZoo [27]	16,487,972	Late 2011- 2016

The authors employed antivirus scan results as well as automation methods to classify these samples based on behavioral semantics (135 varieties), and 71 malware families. It is available on http://amd.arguslab.org/ .

III. PROPOSED SCHEME

The proposed scheme categorizes into two parts: first part relies on examining how feature engineering might affect on obtained results. The machine learning classifier extracts static and dynamic features, finds most informative features and employs state of the art machine classifiers, while the deep learning scheme studies how deep learning might enhance the gained performance. We performed several scenarios to present the effectiveness of developed scheme.

A. Machine Learning Classifiers

The Machine Learning classifiers scheme consisted of four parts. First, we build a dataset comprised of public and private datasets. Then, two types of features are extracted: permission-based and network traffic features. Once the features are extracted, the most informative features are ranked for further analysis in the third phase. Finally, the developed scheme is evaluated among four scenarios: detecting malware in terms of binary classification, classifying the samples based on their packages and families, and finally considering sustainability metrics in our scheme. Fig. 1 depicts the structure of the proposed scheme.

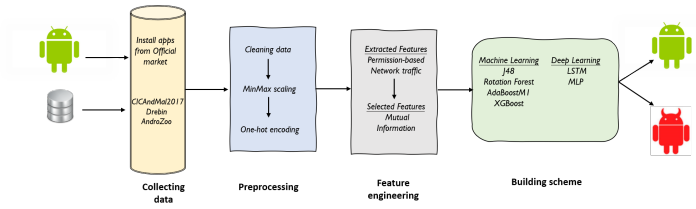


Fig. 1. Architecture of the Implemented Scheme.

1) Acquiring Data: Since the commonly used public datasets were collected between 2008 and 2017 [2], we started to construct a public dataset to be available in the near future for academic research purposes. For this purpose, we picked three public datasets: Drebin [16], CICAndMal2017 [26] and AndroZoo [27], which are summarized in Table I. We, then, targeted the official market for Android devices: the Google Play store to download up to 500 popular apps for 30 different categories using 15 different smartphone devices. Finally, we analyzed and stored downloaded apps in a local database. The process of collecting apps started November 2020 and we are continuing to build our dataset.

2) Pre-processing the Data: Once the data was collected, we performed a pre-processing step for the data which aims to observe any possible noise, remove duplicate apps and keep the updated version of the app, as well as to find missing, redundant and invalid data. For normalization, we used Min-Max scaling [32], and One Hot Encoding [33] to transform non-numeric data into numeric values. After performing this process, we had a total dataset of 15,000 malware apps, and 16500 normal apps. For balance, we constructed an updated dataset with a total of 30,000 samples (15,000 normal apps and 15,000 malware apps) and continued working to make the dataset larger.

3) Extracting Features: To extract the features from our dataset, we developed a tool called AndroAPKF Analyzer, installed it on our devices, ran each app, then extracted permission-based and network traffic features. We extracted two types of features: permission-based (280 extracted features) and network traffic (80 extracted features).

Once the features are extracted, they are stored temporarily in the phone, then sent to the local database for further process. The final stored data had a CSV file extension.

4) Selecting Features: Finding the most informative features is a subsequent phase during construction of the scheme. Using selected features will help to reduce evaluation time and over-fitting and improve the obtained results. There are several approaches that can be used to find most informative features such as mutual information (MI), which is a method to measure the mutual independence of the amount of information that a variable contains about the occurrence of another variable [34].

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} P(x,y) log_p \frac{x,y}{p(x)p(y)} \quad (1)$$

where X, Y are discrete or continuous random variables, $p(x)p(y)$ is the product of marginal distributions. For our approach, we adopted the implemented ranking scheme proposed by Alzubaidi et. al [35] to find the most informative features. We ranked permission features and network traffic individually in terms of the top 10, top 50, and all features. Tables II and III present highest informative extracted permission and network traffic features, respectively.

TABLE II. TOP 10 PERMISSION-BASED FEATURE RANKING USING [35]

Feature set	Rank
Android.permission.INTERNET	0.937
Android.permission.READ_PHONE_STATE	0.925
Android.permission.ACCESS_NETWORK_STATE	0.909
Android.permission.SEND_SMS	0.886
Android.permission.WRITE_EXTERNAL_STORAGE	0.881
Android.permission.RECEIVE_BOOT_COMPLETED	0.867
Android.permission.ACCESS_WIFI_STATE	0.852
com.Android.launcher.permission.INSTALL_SHORTCUT	0.834
Android.permission.INSTALL_PACKAGES	0.785
com.android.alarm.permission.SET_ALARM	0.748

B. Experimental Setup

We ran experiments using Microsoft Windows 10 on a 2.11 GHz Intel Core i7 178 processor with 16 GB RAM dGPU

TABLE III. TOP 10 NETWORK TRAFFIC FEATURE RANKING USING [35]

Feature set	Rank
Source IP address	0.913
Source Port	0.906
Destination IP address	0.883
Destination Port	0.876
Flow Duration	0.833
Total Foward Packets	0.795
Total Length of Forward Packets	0.792
Total Backward Packets	0.748
Total Length of Bakward Packets	0.741
Maximum Forward Packet	0.705

device. We also used a virtual machine p2.xlarge EC2 instances to perform further experiments and a local database for storing the data.

C. Performance Evaluation Metrices

Common evaluation metrics that can be used to examine the performance of the implemented methods for detecting malware such as true positive, false positive, accuracy, recall, precision and sustainability are summarized in Table IV. In our implemented scheme, the data are evaluated using accuracy, precision, recall and f1-score.

TABLE IV. COMMON EVALUATION METRICS

Metric	Definition
True positive	A sample is a true positive if the sample is labelled positive as well as prediction being positive
False positive	A sample is a false positive if the sample is labelled negative while it is predicted as a positive
False negative	A sample is called false negative when the sample is labelled positive and is predicted as a negative class
Precision	It is the proportion of correctly classified instances to the instances predicted as positive
Recall	It is the proportion of instances predicted to be positive to the total positive instances
Sustainability	Indicates how the model is adopted for new samples with retraining (re-usability) and without retraining (stability) over times

D. Experimental Evaluation

We initially utilized four machine learning classifiers to examine the effectiveness of the ranking features: J48 [36], rotation forest [37], AdaboostM1 [38], and XGBoost [39]. Table V outlines these classifiers with respect to their definition and platform.

We evaluated the machine learning classifiers using ranking features with the top 10, 50 and all features for the detection of unknown malware (binary classification) using permission and network traffic features.

Using Permission based Feature: We compare the utilized classifiers using top 10,50 and all features in terms of accuracy, precision, recall and f1-score. Among all three feature sets, rotation forest achieved better results with accuracies of 94.4%,98.6% and 96.5%. The rest results are illustrated in Tables VI, VII, and VIII.

TABLE V. CHARACTERISTICS OF THE CLASSIFIERS EMPLOYED

Classifier	Definition	Availability
J48 [36]	Implemented classifier of C4.5 decision tree	Weka 3.8.5
Rotation Forest [37]	Ensemble classifier, categorizes features into subsets, then run Principal Component Analysis on each subset	Weka 3.8.5
AdaBoostM1 [38]	An ensemble classifier used for boosting a nominal-class	Weka 3.8.5
XGBoost [39]	A boosted decision tree, which associates objective functions based on the gradient of the loss optimized function	R 4.0.0

TABLE VI. MALWARE DETECTION BASED ON PERMISSION FEATURE SET (TOP10 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	91.8	91.8	91.7	91.7
Rotation Forest	**94.4**	**94.5**	**94.3**	**94.5**
AdaBoostM1	89.6	88.6	89.7	89.5
XGBoost	93.7	93.6	93.5	93.7

1) Using Network Traffic Feature: We also evaluated the proposed approach using network traffic with respect to top 10,50 and all features. Among employed aforementioned classifiers,XGBoost is able to gain the best results with accuracies of 89.8%, 94.2% and 91.6%. The rest results are illustrated in Tables IX, X, and XI.

IV. DEEP LEARNING SCHEME

Deep learning algorithms are a subset of machine learning algorithms based on artificial neural networks, which can be applied to various fields such as computer vision, speech recognition, natural language processing, and lately malware detection. We employed two deep learning classifiers: long short-term memory (LSTM) [40] and multilayer perceptions (MLPs) [41] to enhance the obtained results.

A. Long-Short Term Memory

A brief definition Long short-term memory (LSTM) considers recurrent structures having the ability to learn a sequence of data. Generally, the sequence of features for the Android apps are $x = (x_1, x_2, x_3, ...x_t)$, which represent the input values to the LSTM. Then, the calculated output will be denoted as $ot = (o_{t1}, o_{t2}, o_{t3}, ...o_{tt})$ with respect to continuous calculation of the forget state vector, input/update gate activation vector, output vector, and cell state vector for time $T = 1, 2, 3, ...t$. The calculation of the hidden layer can be calculated as listed below.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} b_f)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$

TABLE VII. MALWARE DETECTION BASED ON PERMISSION FEATURE SET (TOP50 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	97.3	97.2	97.3	97.3
Rotation Forest	**98.6**	**98.5**	**98.4**	**98.5**
AdaBoostM1	95.4	95.5	95.4	95.4
XGBoost	98.5	98.5	98.6	98.5

TABLE VIII. MALWARE DETECTION BASED ON PERMISSION FEATURE SET (ALL 280 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	95.4	95.3	95.2	95.4
Rotation Forest	**96.5**	**96.6**	**96.4**	**96.5**
AdaBoostM1	91.2	91.1	91.2	91.3
XGBoost	96.4	96.3	96.3	96.4

$$\tilde{c_t} = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t o c_{t-1} + i_t o \tilde{c_t}$$

$$h_t = o_t o \sigma_h(c_t)$$

where ft, it, ot,ct, ht, and xt denote the forget state vector, input/update gate activation vector, output vector, cell state vector, hidden state vector, and input vector, respectively.

B. Experimental Result

We evaluated the deep learning scheme in three scenarios, as listed below.

1) Detecting unknown Malware using Binary Class Classification: We initially performed a comparison among: LSTM, MLP and XGBoost in terms of detecting unknown malware using accuracy, precision, recall and f1-score using permission-based features. The obtained results show the superiority of LSTM compared to MLP and XGBoost, which achieved accuracy, precision, recall and f1-score higher than 99%. Fig. 2 depicts the results.

2) Detecting Categories of Malware : Another comparison was performed to examine how these classifiers are able to detect malware based on malware categories. Our malware samples belong to four categories: Adware, Ransomware, Scareware and SMS Malware. We performed four multi-class classifications to examine to which malware family each malware belong to. Comparing LSTM, MLP and XGBoost, LSTM is achieved the better results with accuracy of 97.5%. Fig. 3 outlines the comparison among these classifiers in terms of accuracy, precision, recall and F1-score.

3) Detecting Malware Families: A third comparison was performed to detect which package the malware belongs to. In our evaluation, we defined 45 malware families; therefore, we performed 45 class- classification procedures to examine the samples. We compared LSTM, MLP, and XGBoost and found LSTM achieved best accuracy of 99.5%. Fig. 4 illustrates the obtained results.

TABLE IX. MALWARE DETECTION BASED ON NETWORK TRAFFIC FEATURE SET (TOP10 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	87.7	87.6	87.6	87.7
Rotation Forest	88.6	88.5	88.6	88.5
AdaBoostM1	83.3	83.2	83.2	83.3
XGBoost	**89.8**	**89.6**	**89.7**	**89.6**

TABLE X. MALWARE DETECTION BASED ON NETWORK TRAFFIC FEATURE SET (TOP50 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	90.4	90.4	90.3	90.3
Rotation Forest	92.6	92.5	92.4	92.5
AdaBoostM1	86.5	86.5	86.4	86.5
XGBoost	**94.2**	**94.1**	**94.2**	**94.1**

4) Examining the Sustainability Performance: Third evaluated of our scheme is examined how our scheme is sustained. We divided the datasets into normal and malware apps based on implemented year with a span of nine years (2010 to 2019) . Then, we trained and tested apps that developed in the same year. The average obtained accuracy is 92.21%. Table XII presents the obtained results in terms of accuracy, precision and recall metrics.

C. Discussion

Although previous studies have obtained promising results in terms of sustainable detecting malware, there are some limitations, which are:

1) Choosing an updated dataset and selecting informative features are vital methods to detect unknown malware and improve obtained results. For example, Onwuzurike et al. [6] utilized a dataset collected between 2010 and 2017, neglected feature selection during building their scheme, and obtained accuracy between 75- 87% .
2) Studying mobile malware in terms of detecting unknown malware families, and categories using multi-class classifications has not covered from Onwuzurike et al. [6], Zhang et al. [11], Cai and Jenkins [17], Cai et al. [19], and Cai [20]

Table XIII presents a comparison between the proposed approach and prior studies. The introduced approach achieved accuracies of 99.2%,99.5%, 97.5%, and 92.2% for detecting unknown malware, detecting malware families, detecting malware categories, and sustainable malware detection, respectively.

V. CONCLUSION

Since the amount of mobile malware increases annually, it is important to implement malware detection that is able to detect possible threats to smartphone devices. This study leveraged machine and deep learning to detect unknown malware apps with accuracy over 99%. We also employed multi-class classification to detect the packages and families of mobile malware, and finally we examined the sustainability of our implemented scheme.

TABLE XI. MALWARE DETECTION BASED ON NETWORK TRAFFIC FEATURE SET (ALL 80 FEATURES)

Classifier	Accuracy	Precision	Recall	F1-score
J48	88.7	88.6	88.5	88.5
Rotation Forest	90.5	90.4	90.5	90.3
AdaBoostM1	82.3	82.2	82.3	82.4
XGBoost	**91.6**	**91.4**	**91.5**	**91.5**

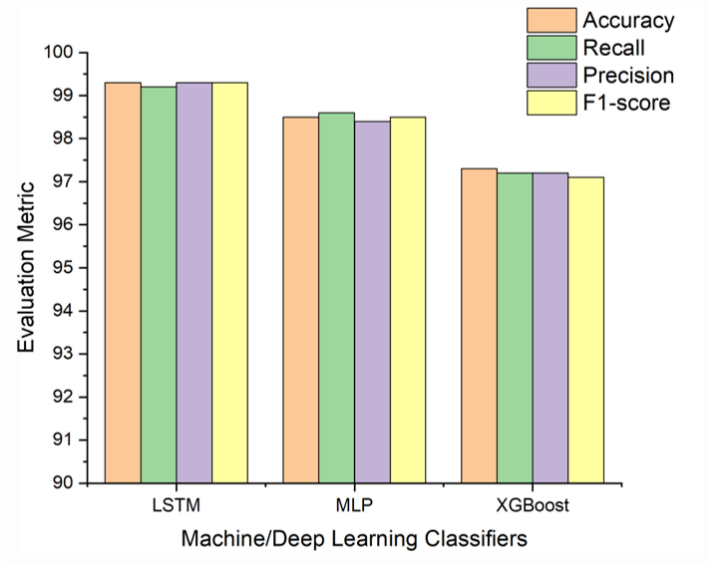


Fig. 2. Performance Evaluation Metrics based on Detecting unknown Malware.

TABLE XII. DATASET EVALUATION BASED ON SAME PERIOD OF DEVELOPED YEAR

Dataset	Accuracy	Recall	Precision
DS1	92.2	92.1	92.2
DS2	93.5	92.4	92.5
DS3	91.6	91.6	91.6
DS4	89.7	89.6	89.5
DS5	92.3	92.4	92.2
DS6	90.3	90.2	90.4
DS7	94.8	94.7	94.8
DS8	91.6	91.6	91.6
DS9	93.9	93.8	93.9
Average	**92.21**	**92.04**	**92.08**

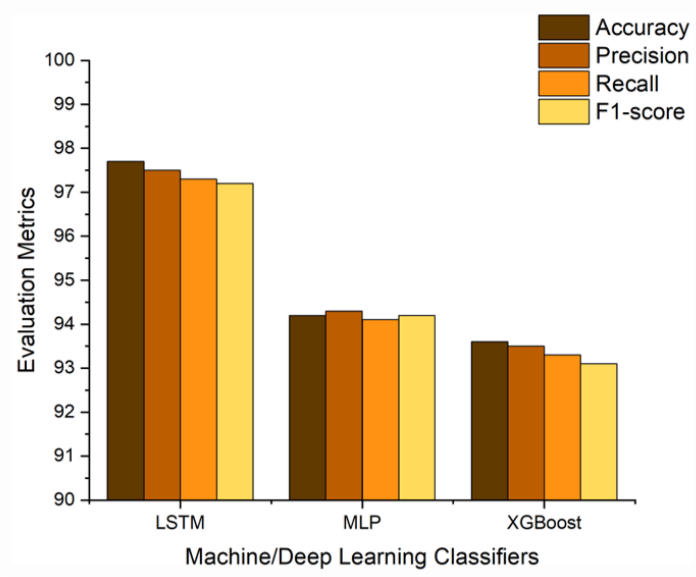


Fig. 3. Performance Evaluation Metrics based on Classifying Malware in Terms of Categories.

For future work, we aim to extend our dataset to download more apps, then make it available for academic use. We also plan to use various deep learning classifiers with different dynamic and static features to detect unknown malware, as well as assess datasets over a period of years for sustainability purposes.

ACKNOWLEDGMENT

The author would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: 18-COM-1-01-0007.

REFERENCES

[1] Alzubaidi, A., & Kalita, J. (2016). Authentication of smartphone users using behavioral biometrics. IEEE Communications Surveys & Tutorials, 18(3), 1998-2026.

[2] Alzubaidi, A., 2021. Recent Advances in Android Mobile Malware Detection: A Systematic Literature Review. IEEE Access.

[3] IT threat evolution in Q2 2021. Mobile statistics . Available online: https://securelist.com/it-threat-evolution-q2-2021-mobilestatistics/103636/ (accessed on: 12 - 10 - 2021)

[4] Mobile Malware . Available online: https://usa.kaspersky.com/resource-center/threats/mobile-malware (accessed on: 12 - 10 -2021)

[5] Qiu, J., Zhang, J., Luo, W., Pan, L., Nepal, S., & Xiang, Y. (2020). A survey of Android malware detection with deep neural models. ACM Computing Surveys (CSUR), 53(6), 1-36.

[6] Onwuzurike, L., Almeida, M., Mariconti, E., Blackburn, J., Stringhini, G., & De Cristofaro, E. (2018). A family of droids: Analyzing behavioral model based Android malware detection via static and dynamic analysis. arXiv preprint arXiv:1803.03448.

[7] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

[8] Pal, M., 2005. Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), pp.217-222.

[9] Laaksonen, J., & Oja, E. (1996, June). Classification with learning k-nearest neighbors. In Proceedings of International Conference on Neural Networks (ICNN'96) (Vol. 3, pp. 1480-1483). IEEE.

[10] Suykens, J.A. and Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural processing letters, 9(3), pp.293-300.

[11] Zhang, X., Zhang, Y., Zhong, M., Ding, D., Cao, Y., Zhang, Y., Zhang, M. and Yang, M., 2020, October. Enhancing state-of-the-art classifiers with API semantics to detect evolved android malware. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (pp. 757-770).

[12] Hinton, G.E. and Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. science, 313(5786), pp.504-507.

[13] Mariconti, E., Onwuzurike, L., Andriotis, P., De Cristofaro, E., Ross, G. and Stringhini, G., 2016. Mamadroid: Detecting android malware by building markov chains of behavioral models. arXiv preprint

TABLE XIII. COMPARISON OF OUR PROPOSED WITH OTHER SUSTAINABLE DETECTION SCHEME. NOTE: NP REFERS TO NOT PROVIDED

Approach	Extracted features	# of samples	Detecting unknown malware	Detecting malware family	Detecting malware category	Sustainable malware detection
Onwuzurike et al. [6]	API calls	43,940	99%	NP	NP	87%
Cai and Jenkins [17]	extent, frequency, distribution for the source, API calls	6432	NP	NP	NP	82%-93%
Cai [20]	behavioral profiling features	26382	NP	NP	NP	92.88
Our Scheme	permission based, network traffic	30,000	99.2%	99.5%	97.5%	92.21%

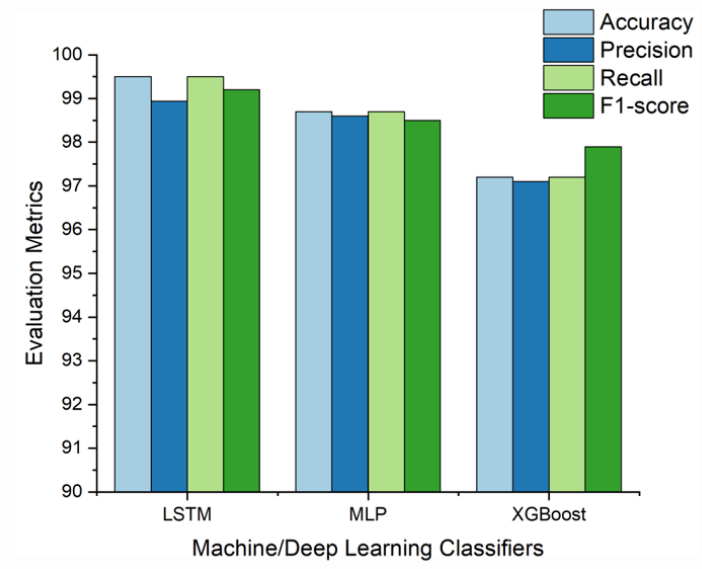


Fig. 4. Performance Evaluation Metrics based on Classifying Malware in Terms of Family.

arXiv:1612.04433.

[14] Xu, K., Li, Y., Deng, R., Chen, K. and Xu, J., 2019, June. DroidEvolver: Self-evolving Android malware detection system. In 2019 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 47-62). IEEE.

[15] Grosse, K., Papernot, N., Manoharan, P., Backes, M. and McDaniel, P., 2017, September. Adversarial examples for malware detection. In European symposium on research in computer security (pp. 62-79). Springer, Cham.

[16] Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K. and Siemens, C.E.R.T., 2014, February. Drebin: Effective and explainable detection of android malware in your pocket. In Ndss (Vol. 14, pp. 23-26).

[17] Cai, H. and Jenkins, J., 2018, May. Towards sustainable android malware detection. In Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings (pp. 350-351).

[18] Cai, H. and Ryder, B.G., 2017, September. Understanding Android application programming and security: A dynamic study. In 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME) (pp. 364-375). IEEE.

[19] Cai, H., Meng, N., Ryder, B. and Yao, D., 2018. Droidcat: Effective android malware detection and categorization via app-level profiling. IEEE Transactions on Information Forensics and Security, 14(6), pp.1455-1470.

[20] Cai, H., 2020. Assessing and improving malware detection sustainability through app evolution studies. ACM Transactions on Software Engineering and Methodology (TOSEM), 29(2), pp.1-28.

[21] Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), pp.1-27.

[22] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[23] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[24] Schapire, R.E., 2003. The boosting approach to machine learning: An overview. Nonlinear estimation and classification, pp.149-171.

[25] Schapire, R.E., 2003. The boosting approach to machine learning: An overview. Nonlinear estimation and classification, pp.149-171.

[26] Lashkari, A.H., Kadir, A.F.A., Taheri, L. and Ghorbani, A.A., 2018, October. Toward developing a systematic approach to generate benchmark android malware datasets and classification. In 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1-7). IEEE.

[27] Allix, K., Bissyandé, T.F., Klein, J. and Le Traon, Y., 2016, May. Androzoo: Collecting millions of android apps for the research community. In 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR) (pp. 468-471). IEEE.

[28] Zhou, Y., & Jiang, X. (2012, May). Dissecting android malware: Characterization and evolution. In 2012 IEEE symposium on security and privacy (pp. 95-109). IEEE.

[29] Parkour, M, Mobile Signatures, Available online: http://contagiodump.blogspot.com, 2008, (Accessed on: 11/11/2021)

[30] Maiorca, D., Ariu, D., Corona, I., Aresu, M., & Giacinto, G. (2015). Stealth attacks: An extended insight into the obfuscation effects on android malware. Computers & Security, 51, 16-31.

[31] Wei, F., Li, Y., Roy, S., Ou, X., & Zhou, W. (2017, July). Deep ground truth analysis of current android malware. In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 252-276). Springer, Cham.

[32] Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

[33] Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing, 75, 21-31.

[34] Anthony, M., Bartlett, P.L. and Bartlett, P.L., 1999. Neural network learning: Theoretical foundations (Vol. 9). Cambridge: cambridge university press.

[35] Alzubaidi, A., Roy, S. and Kalita, J., 2017, May. Ranking most informative apps for effective identification of legitimate smartphone owners. In 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 790-795). IEEE.

[36] Mathuria, M., 2013. Decision tree analysis on j48 algorithm for data mining. Intrenational Journal of Advanced Research in Computer Science and Software Engineering, 3(6).

[37] Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J., 2006. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence, 28(10), pp.1619-1630.

[38] Cortes, E. A., Martinez, M. G., & Rubio, N. G. (2007). Multiclass corporate failure prediction by Adaboost. M1. International Advances in Economic Research, 13(3), 301-312.

[39] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree

boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[40] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[41] Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. Computer Standards & Interfaces, 16(3), 265-278.

Solving the Steel Continuous Casting Problem using an Artificial Intelligence Model

Achraf BERRAJAA
INSA Euro-Mediterranean, Euromed Research Center,
Euromed University of Fes, Fes, Morocco

***Abstract*—Over the past decade, the steel continuous casting problem has revolutionized in important and remarkable ways. In this paper, we consider a multiple parallel device for the steel continuous casting problem (SCC) known as one of the hardest scheduling problem. The SCC problem is an important NP-hard combinatorial optimization problem and can be seen as three stages hybrid flowshop problem. We have proposed to solve it a recurrent neural network (RNN) with LSTM cells that we will executed in the cloud. For our problem, we consider several machines at each stage that are the converter stage, the refining stage and the continuous casting stage. We formulate the mathematical model and implemented a RNN with LSTM cells to approximately solve the problem. The proposed neural network has been trained on a big dataSet, which contains 10 000 real use cases and others generated randomly. The performances of the proposed model are very interesting such that the success rate is 93% and able to resolve large instances while the traditional approaches are limited and fail to resolve very large instances. We analyzed the results taking into account the quality of the solution and the prediction time to highlight the performance of the approach.**

***Keywords*—Artificial intelligence; SCC Program; RNN; LSTM; big data**

I. Introduction

Iron and steel industry is the cornerstone of an industrialized economy. Since it is capital and energy intensive, companies have constantly laid great emphasis on technological advances to be employed in the production process in order both to increase productivity and to save energy. Due to the complex production process and potential constraints, the latter faces difficult planning and scheduling problems. For example, in the scheduling problem, we usually define a set of n resources, a set of m tasks and a specific optimization goal, called makespan Sol_{max}. The classic flowshop (FS) considers scheduling a set of tasks on one machine at each stage, while the hybrid flowshop (HFS) aims to schedule a flow shop with multiple parallel machines at each stage [1]. In the steelmaking industry, we have three main stages principals for the production that are the converters (CV), the refining stands (RS) and the continuous castings (CC) stages. Each one can include one or more devices, and each product is processed on only one device in each stage. Fig. 1 summarized this configuration, the SCC problem can be seen as a hybrid flow shop. More generally, the goal of SCC is to determine the sequence, timing, and system of equipment involved in the entire production process. This problem is a NP-hard combinatorial optimization and it is considered to be one of the more difficult scheduling problems in the literature [30].

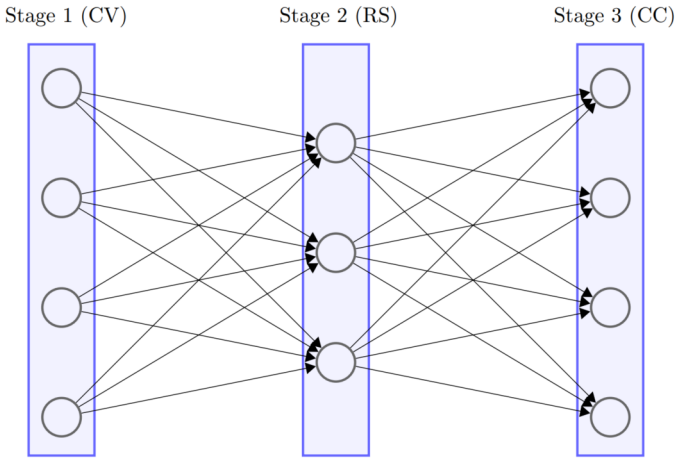


Fig. 1. The Principle Stage of the SCC.

The principal stages and constraints are as follows :

1) The length of stay is minimized,
2) The deadline must be met,
3) The continuity constraint must be satisfied.

We propose in this research, a system with three stage M_i machines at each stage $i(i = 1, \ldots, 3)$, such that M_1 identical parallel CV machines are available in the first stage (denoted as CV_{M_1}), M_2 unify the parallel RS machines at stage two (denoted as RS_{M_2}) and M_3 unrelated parallel CC machines at stage three (denoted as CC_{M_3}). In addition, we consider the inter-sequence correlation setting time between every two consecutive sequences λ to be processed on the same CC machine with non-preemptive scheduling. Similarly to the notation by [8], we write our system as a hybrid flowshop (HFS) : $CV_{M_1}, RS_{M_2}, CC_{M_3} \mid \lambda \parallel Sol_{max}$. The first modeling of the problem was considered as a special case of m-stages HFS, which has irrelevant machines and related setup time, which proved to be NP-Hard [14]. For example, [22] has proven the NP hardness of a single machine with a set time.

Some recent research work have been developed to study the steel making continuous casting problem. Most of the developed models deal with high complexity, so solving them as optimal is not always effective, especially for the large instances problems. The most of the methods proposed for the HFS problems are approximation methods, most of which are meta-heuristics. The SCC problem is a special set time scheduling problem, which is difficult to solve due to the high

computational complexity [22]. The SCC problem is based on a chronological step chain (the three stages: CV, RS, and CC), where CV, RS, and CC devices are considered to be single, and only one sequence of loads is processed on the machine M_i according to the rules. Among the key rules of the SCC problem, the charging will not start processing on the device until the previous charge is completed. This architecture can be represented by a neural network (forward propagation principle [11]) which allows the automation of the problem SCC. To the best of our knowledge, this is the first work that proposes a model of artificial intelligence to solve the problem of SCC, in particular a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells.

In the remainder, the article is organized as follows: Section 2, presents a literature review on the works related to the steelmaking industry in general and the continuous casting in particular. Sections 3 and 4 detail the structure, the sequences and charges, the constraints and formulate the objective function and the mathematical model for the SCC. Section 5 we present the construction of learning big data and we describe our recurrent neural network with LSTM cells for SCC. Section 6 is devoted to the numerical experiments. We also discuss and comment the obtained results and show the efficiency of the method. Finally, Section 7, presents a conclusion that summarizes the study and gives some potential perspectives for the developed approach to improve the current results for the SCC.

II. Related Work and Problem Statement

Continuous casting is one of the most commonly used processes in steel production and has received special attention in the past ten years [31]. Several research works have been developed, both exact and heuristic algorithms have been elaborated to solve several SCC problem variants. The authors in [28] reviewed several SCC models used in steelmaking production, and other works considered mathematical and non-mathematical techniques [3] to solve these variants. Optimization methods such as genetic algorithm (GA) [33], taboo search (TS) [14], mathematical model [32] or swarm intelligence optimization [10] have been developed to optimize production. The hybrid method [27] is also tailored for a continuous steel plant.

Also exact MILP methods are proposed to unravel the SCC. Usually, these methods can only solve small problem instances with commercial solvers, but the real challenge is to use smart and optimized methods to solve large instances. The author in [29] proposed a MILP model for the production order scheduling problem. This model describes a system with more than two machines in each stage, and also more than two sequences must be processed in the final stage, in addition there is an inter-sequence setting time. This model is solved by commercial solvers for relatively small instances. In [9] a mathematical model of the process and a state-of-the-art industrial control system are presented, and they mentioned the importance of using a real-time computational model. In [6], a method has been developed for a steelmaking plant with 2 converters, 2 refining stands, and 2 continuous casters. The authors in [12] develop a discrete-time mixed-integer linear programming (MILP) formulation for a new SCC scheduling problem where different processing routes are used to produce diversified and personalized slab products. The authors in [23] developed a method to minimize the total delay and waste of scheduling problems for different product types by minimizing batches.

Scheduling problems are mainly NP-Hard optimization problems ([5]). In order to overcome their complexity, heuristics and approximation methods may proposed a solution for this types of problems. In the literature approximation methods were developed for SCC. In [18], heuristic-based combinatorial auction technology is also used to solve the SCC containing 4 production orders and a 10-hour planning range. The authors in [16] have developed approximate techniques to solve some planning and scheduling problems of downstream production lines. In [13], the authors have developed a soft-decision based two-layered approach for strong uncertain scheduling. The paper [26] presents an Improved Artificial Bee Colony (IABC) algorithm for the SCC scheduling. Other uncertainty optimizations have been developed for SCC, such as a novel efficient solution algorithm using Augmented Lagrange multiplier method (e-ALM) through relaxation of the coupling constraints and incorporation of penalty components [12].

Several evolutionary algorithms have also been developed for SCC. In [20], we find the description several approaches for computerized scheduling solutions. They include application of techniques in operations research, artificial intelligence, and a hybrid of these two. Nature-inspired optimization methods have also been used in SCC. For example, in [7] the steel industry has been modeled through a combination of a steady-state heat transfer approach and a pareto-converging genetic algorithm (PCGA). In [2], the authors developed a algorithm, that is based on a combination of ant colony optimization and non-linear optimization methods. The authors in [17] have implemented a multi-objective GA to minimize penalties for completion and lateness. In [24], the authors developed a swarm heuristic method for SCC. In the paper of [25], the authors develop a discrete-time mixed-integer linear programming (MILP) formulation for a new SCC scheduling problem where different processing routes are used to produce diversified and personalized slab products. A multi-objective hybrid genetic algorithm combined with local search was presented in [32], in which the enhanced evolutionary mechanisms combined with the improved genetic operators and the local search were also designed. Regarding this work, our approach is based on an artificial neural network, where artificial neural network, is a system whose design is originally schematically inspired by the functioning of biological neurons. The idea has no relation to the way a human being reacts, but to the way of designing the data. It adapts and gives good results when there is a lot of sample information (Features). To the best of our knowledge, this is the first work that proposes a model of artificial intelligence to solve the problem of SCC, in particular a recurrent neural network with long short-term memory cells.

A. Problem Statement

The system processes the costs of different stages under the continuity constraint $(i = 1, \ldots, 3)$. These sequences are pre-ordered costs on the continuous castings machine, and their costs are allocated to any converter machine. Since handling and transportation operations occur in the entire process, the transfer time between stages $\tau_{i,i+1}$ (i = 1, 2) is considered.

We also believe that all converters devices have the earliest and latest start time for the first charge of a sequence with a usable date. The working principle of the system is as follows:

- A sequence represent a set of charges dedicated to one continuous casting machine, and has priority constraints on fees;
- The processing time is limited for any converter, who can be used in the first stage;
- There is a limited residence time (transmission) between the termination of charging in converter and the start of continuous casting to comply with the required temperature;
- Due to continuous constraints, there is no idle time for the charges on the continuous casting;
- The processing time of the continuous casting stage is a decision variable belonging to a given interval;
- The sequence start time at continuous casting is bounded, and the delay between two consecutive sequences is the setup time we define.

The sorted total batch represents the industrial order book so that the sequence can be assigned to the continuous casting.

1) Concept of Sequence and Continuity Constraints: The sequence Seq_{l_3} is defined as a set of $n_{Seq_{l_3}}$ charges to be used in the M_3 casters CC_{l_3} $(l_3 = 1, \ldots, M_3)$. The relevant constraints are summarized as follows:

1) The converter can be used in the first stage, there is no specification, and it has a constant processing time and availability date;
2) There is a transition time $\tau_{i,i+1}$ between all stages i (i = 1, 2);
3) The stay (transit) time of the charge j (the time between the termination of the first stage $i = 1$ and the start of the final stage) must not exceed a certain value $T_{l_3,j}$;
4) No idle time is allowed between two consecutive charges in the same sequence in the third stage (continuous constraint);
5) The setting time depends on the sequence time;

III. Continuous Casting of the Ordered Sequences

This section describes all the data (sets, indexes, etc.), parameters, and decision variables that describe the problem being studied. For a given CC_{l_3}, a pre-ordered dedicated sequence Seq_{l_3} $(Seq_{l_3} = 1, \ldots, S_{l_3})$ is a list of $n_{Seq_{l_3}}$ charges to continuously process on CC_{l_3} with setup times.

Collections, constants, and indexes :
i : the index of the stage, $i = 1, \ldots, 3$;
M_i : the number of the machines l_i at stage i $(l_i = 1, \ldots, M_i)$;
S_{l_3} : the number of the sequences Seq_{l_3} to be processed on the machine CC_{l_3};
Seq_{l_3} : a sequence to process on CC_{l_3} $(Seq_{l_3} = 1, \ldots, S_{l_3})$;
$n_{Seq_{l_3}} = |Seq_{l_3}|$: the number of charges (jobs) j of the sequence Seq_{l_3};
$n_{l_3} = \sum_{Seq_{l_3}=1}^{S_{l_3}} n_{Seq_{l_3}}$: the total number of the charges to process on CC_{l_3} $(l_3 = 1, \ldots, M_3)$;
$n = \sum_{l_3=1}^{M_3} n_{l_3}$: the total number of the charges;
j : the charge index dedicated to the sequence Seq_{l_3} of CC_{l_3}, $j = 1, \ldots, n_{l_3}$;
l_1 : the index of the first stage machine CV_{l_1}, with $(l_1 = 1, \ldots, M_1)$;

The consider assumptions :
ω_{l_1} : position of a charge on CV_{l_1}, $\omega_{l_1} = 1, \ldots, \Pi_{l_1}$;
$l_1 = 1, \ldots, M_1 - 1$
$\Pi_{l_1} = \frac{\sum_{l_3=1}^{M_3} n_{l_3}+1}{M_1}$;
ω_{M_1} : position of a charge on the last machine CV_{M_1},
$\omega_{M_1} = 1, \ldots, \Pi_{M_1}$;
$\Pi_{M_1} = \frac{\sum_{l_3=1}^{M_3} n_{l_3}}{M_1}$;

Settings and parameters
pro^1 : processing time (constant) for a charge j on any one of the CV_{l_1} at the 1^{st} stage;
$pro^2_{l_2}$: processing time for a charge j on the RS_{l_2};
$[P^{min}_{l_3,j}, P^{max}_{l_3,j}]$: the interval of $pro^3_{l_3,j}$ is dedicated to the processing time of the cost j of CC_{l_3};
$[\lambda^{min}_{Seq_{l_3}}, \lambda^{max}_{Seq_{l_3}}]$: the interval between sequences depends on the setup time $\lambda_{Seq_{l_3}}$.
$date_{l_1}$: available date of CV_{l_1};
$T_{l_3,j}$: maximum allowed sojourn time for a charge j between the termination of any processing in CV_{l_3} and the start of processing in CC_{l_3} $(j = 1, \ldots, n_{l_3}, l_i = 1, \ldots, M_i, i = 1, 3)$;
τ_{12} (resp. τ_{23}) : transfer time required between CV_{l_1} (stage 1) and RS_{l_2} (stage 2) (resp. RS_{l_2} (stage 2) and CC_{l_2} (stage 3)).

Decision variables
The model considers continuous and binary decision variables. For any $l_3 = 1, \ldots, M_3$:
$x^i_{l_3,j}$: start time of the charge j dedicated to CC_{l_3} at stage i $(i = 1, 2, 3)$ for $j = 1, \ldots, n_{l_3}$,
$pro^3_{l_3,j}$: processing time dedicated to the charge j of CC_{l_3} at the 3^{rd} stage for $j = 1, \ldots, n_{l_3}$,
$\lambda_{Seq_{l_3}}$: setup time between two consecutive sequences, $(Seq-1)_{l_3}$ and Seq_{l_3} to process at CC_{l_3}, that occurs between the charge $n_{(Seq-1)_{l_3}}$ and the charge $l_{Seq_{l_3}}$.

$$y^{l_3}_{j,\omega_{l_1}} = \begin{cases} 1 & \text{if charge j dedicated to } CC_{l_3} \\ & \text{is assigned to a position } \omega_{l_1} \text{ in } CV_{l_1} \\ 0 & \text{otherwise} \end{cases}$$

$j = 1, \ldots, n_{l_3}, \omega_{l_1} = 1, \ldots, \Pi_{l_1}, l_1 = 1, \ldots, M_1.$

IV. The Mathematical Model for the SCC

In the research problem, we intend to maximize productivity, that is, minimize the manufacturing span (Sol_{max}) and the sequence-dependent setup time, which represents the required duration between two consecutive sequences.

Our modeling method uses the position of the sequence charge and the priority defined by the processing start time of stage 1 (CV). We are also considering pre-orders for charging CC machines in the third stage.

A. The Constraints

We may remember that due to the high complexity of its structure and function, this problem is subject to several constraints that we detailed below:

$$\sum_{l_1=1}^{M_1}\sum_{\omega_{l_1}=1}^{\Pi_{l_1}} y_{j,\omega_{l_1}}^{l_3}=1;\quad j=1,\ldots,n_{l_3},\quad l_3=1,\ldots,M_3 \tag{1}$$

$$\sum_{l_3=1}^{M_3}\sum_{j=1}^{n_{l_3}} y_{j,\omega_{l_1}}^{l_3}=1;\quad \omega_{l_1}=1,\ldots,\Pi_{l_1},\quad l_3=1,\ldots,M_3 \tag{2}$$
$$\text{and}\ \sum_{l_3=1}^{M_3} y_{1,1}^{1}=1$$

$$y_{j+1,t_{(l_1=1)}+1}^{l_3}\le\sum_{l_1=1}^{M_1-1}\sum_{\omega_{l_1}=1}^{t_{l_1}} y_{j,\omega_{l_1}}^{l_3}; j=1,\ldots,n_{l_3},\quad t_{l_1}=1,\ldots,\Pi_{l_1}-1 \tag{3}$$

$$y_{j+1,t_{l_{M_1}}}^{l_3}\le\sum_{l_1=1}^{M_1}\sum_{\omega_{l_1}=1}^{t_{l_1}} y_{j,\omega_{l_1}}^{l_3}-y_{j,t_{l_{M_1}}}^{l_3};\ j=1,\ldots,n_{l_3}-1,\quad t_{l_1}=1,\ldots,\Pi_{l_1}; \tag{4}$$

$$z_{l_3,j}^{1}=\sum_{l_1=1}^{M_1}\sum_{\omega_{l_1}=1}^{\Pi_{l_1}}(date_{l_1}+pro^{1}(\omega_{l_1}-1))y_{j,\omega_{l_1}}^{l_3};\ j=1,\ldots,n_{l_3}; \tag{5}$$

$$z_{l_3,j}^{2}\ge z_{l_3,j}^{1}+pro^{1}+\tau_{12}; j=1,\ldots,n_{l_3}; \tag{6}$$

$$z_{l_3,j}^{3}\ge z_{l_3,j}^{2}+pro_{l_2}^{2}+\tau_{23}; j=1,\ldots,n_{l_3}; l_2=1,\ldots,M_2; \tag{7}$$

$$z_{l_3,j+1}^{2}\ge z_{l_3,j}^{2}+pro_{l_2}^{2}; j=1,\ldots,n_{l_3}-1; l_2=1,\ldots,M_2; \tag{8}$$

$$z_{l_3,j+1}^{3}=z_{l_3,j}^{3}+pro_{l_3,j}^{3};\quad \forall j\notin\{n_1,n_1+n_2,\ldots,\sum_{Seq_{l_3}=1}^{S_{l_3}} n_{Seq_{l_3}}\},\forall\ l_3=1,\ldots,M_3; \tag{9}$$

$$z_{l_3,j+1}^{3}\ge z_{l_3,j}^{3}+pro_{l_3,j}^{3}+\lambda_r;\quad \forall j=\sum_{Seq_{l_3}=1}^{r} n_{Seq_{l_3}}, r=1,\ldots,S_{l_3}-1; \tag{10}$$

$$P_{l_3,j}^{min}\le pro_{l_3,j}^{3}\le P_{l_3,j}^{max}; j=1,\ldots,n_{l_3} \tag{11}$$

$$z_{l_3,j}^{3}-(z_{l_3,j}^{1}+pro^{1})\le T_{l_3,j}; j=1,\ldots,n_{l_1} \tag{12}$$

$$\lambda_{Seq_{l_3}}^{min}\le\lambda_{Seq_{l_3}}\le\lambda_{Seq_{l_3}}^{max}; Seq_{l_3}=1,\ldots,S_{l_3}-1 \tag{13}$$

The constraint (1) represents that the charge j is allocated to only one converter machine and is located in only one position ω_{l_1}. The constraint (2) represents that the converter CV_{l_1} must process the charge j once and only once at a specific position ω_{l_1} ($l_1=1,..,M_1$). In addition, must assign the charge $j=1$ to the first position of the converter CV_1 ($l_1=1$). The constraint (3) demand that the charge (j+1) must be affected on the CV_1 converter where ($l_1=1$) at the position ($t_1=1$) only if the charge j is to process at any of the positions in $1,\ldots,t_{l_1}$, on one of the M_1-1 first converters CV_{l_1}. The same way, for two consecutive charges j and $j+1$, for the last converter CV_{M_1}, constraint (4) has the same meaning. The constraint (5) is set for the start time to process the charges in the converter stage (CV_{l_1}; $l_1=1,\ldots,M_1$). Constraints (6)-(7) represent the sequencing in the same charge, exactly for two consecutive operations. The last operation can be started after the first operation reaches its end time and the charge has been brought to the next stage. In addition, constraint (6) (resp. (7)) represents the priority rules between CV_{l_1} and RS_{l_2} (between RS_{l_2} and CC_{l_3} (respectively)). The constraint (8) means that for two consecutive charges to process on the same refining stand (RS_{l_2}), the second charge can only be processed when the first one has reached its end time. The constraint (9) represents the continuity constraints for all the sequences. The constraint (10) defines the inter-sequence correlation setting time between two consecutive sequences to be processed on the same continuous casting machine (CC_{l_3}). Constraint (11) defines the limit of the third stage charging processing time on the machine CC_{l_3}. The constraint (12) means that the stay time (transport) of the charge is finite, and (13) defines the boundary of the set time between the sequence of the third stage. They set the necessary preparation time before the first charging of the sequence after the last charging end time of the previous sequence on the same machine in the third stage.

B. The Objective Function

We define makespan (Sol_{max}) as a function representing the time required to completely process all sequence sets (from the start time of the first charge in the first stage to the termination time of the last charge at the third stage with their inter-sequence dependent setup times.

The objective function we consider for the problem has the following mathematical form:

$$Sol_{max}=\max_{1\le l_3\le M_3}\{\lambda_{S_{l_3}}+z_{l_3,S_{l_3}}^{3}+pro_{l_3,S_{l_3}}^{3}\}$$

In order to maximize productivity and the goal is to minimize completion time, we have defined the following goals:

$$Minimize\ Sol_{max}=min\max_{1\le l_3\le M_3}\{\lambda_{S_{l_3}}+z_{l_3,S_{l_3}}^{3}+pro_{l_3,S_{l_3}}^{3}\}$$

In this form, the problem is non-linear. Therefore, in order to avoid this situation, we define a new non-negative decision variable z so that we can obtain a new objective (linear) equal to minimizing z and add new constraints:

$$\lambda_{S_{l_3}}+z_{l_3,S_{l_3}}^{3}+pro_{l_3,S_{l_3}}^{3}\le z\quad\forall l_3=1,\ldots,M_3$$

V. The Proposed Contribution to Solve the SCC

Several evolutionary algorithms are developed to solve complex optimization problems, among them we cite [21]. But the real challenge for the SCC problem is to solve instances with an approach intelligent and optimal. Here we develop an adapted neural networks to the SCC with inter-sequence setup times at the last stage using LSTM cells. The adapted neural network trained on a large database of 10 000 use cases solved with CPLEX based on the mathematical model that we presented in the first part of this article.

In the following we detail the evolutionary strategy that we have adopted for this problem but before that we will explain what deep learning is and why exactly the use of the recurrent neural network (RNN) in particular LSTM.

A. Deep Learning

Machine learning is a field of study of artificial intelligence that gives computers the ability to learn from data, that is, to improve their performance at solving tasks without being explicitly programmed. In several areas of machine learning research, it is about creating neural networks.

Deep learning can be defined as special kind of neural networks composed of multiple layers. These networks are better than traditional neural network in persisting the information from previous event. Recurrent neural network is one such machine that has a combination of networks in loop. The networks in loop allow the information to persist. Each network in the loop takes input and information from previous network performs the specified operation and produces output along with passing the information to next network. Some applications require only recent information while others may ask for more from past, exactly the case of the SCC problem. The common recurrent neural networks lag in learning as the gap between required previous information and the point of requirement increases to a large extent. But fortunately Long Short Term Memory Networks [15], a special form of RNN are capable in learning such scenarios.

B. LSTM Neural Networks

Long-term memory (LSTM) is an alternative solution proposed in [15]: the traditional architecture of a Recurrent Neural Network (RNN) that is based on a simple activation function is modified in such a way that the vanishing gradient problem is explicitly avoided, while the learning method remains unchangeable. For more information on this architecture ([4]). But what are the strengths of LSTM ? why the LSTM will be effective to solve the steel continuous casting problem ?

A LSTM neuron network is made up of several cell that have not just one activation function but rather three that are represented as an input gate, a forget gate and an output gate. Each cell remembers the state of the problem treat in several time intervals, and the three gates regulate the flow of information in and out of the cell. The LSTM network is very suitable for classification, processing and prediction based on time series data, because there may be lags of unknown duration between important events in the time series. This is what is needed to schedule tasks between the three stages (CV, RS and CC) of the SCC problem. Also as we explained, LSTM was developed to deal with the explosion and disappearance of gradients that may be encountered when training traditional RNNs. Fig. 2 shows the internal architecture of a LSTM cell.

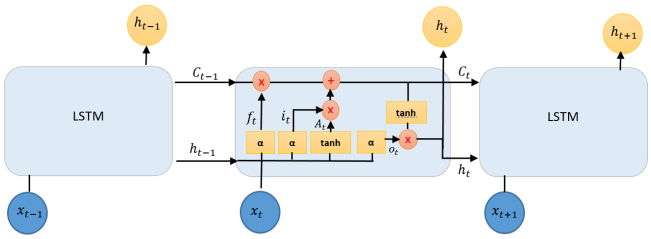


Fig. 2. The Internal Architecture of a LSTM Cell [4].

- **Forget gate:** It is the power of forgetting information. Unlike the classic neural network where it must to memorize all the information in a long sequence, a LSTM has the power to forget unnecessary information that will probably not be used in the prediction. In the LSTM the power to forget non-useful information is represented by the function $f_t = \alpha(W_f * [h_{t-1}, x_t] + b_f)$, where α is a sigmoid function, W_f is the weights and $[h_{t-1}, x_t]$ is the concatenation of the two vectors h_{t-1} and x_t.

- **Input gate:** is responsible for adding relevant information and providing new information. The function that allows this is $C_t = C_{t-1} * f_t + i_t * A_t$ where $i_t = \alpha(W_i * [h_{t-1}, x_t] + b_i$ and $A_t = tanh(W_c * [h_{t-1}, x_t] + b_c)$.

- **Output gate:** This last operation allows to define the current state of the unit. So far, we have forgotten informations, and we have added new informations to the memory. We still need to define the state of the current cell, which represents the output of this cell and which will be the input of the next cell. This is summarized by the following functions: $h_t = \alpha(W_o * [h_{t-1}, x_t] + b_o) * tanh(C_t)$.

C. The Evolutionary Strategy: Training Data, Network Architecture and Choice of Parameters

The proposed neural network will be used as a heuristic to solve the SCC with smart solutions. For this a LSTM is applied to approach the objective function. The proposal to solve the SCC problem is explained in Fig. 3. The system generates a data set in the domain of variables to train a neural network. The objective function of the optimization problem is redefined with the multilayer neural network that transforms the function, allowing to generate a polynomial equation to solve the optimization problem. To define a neural network, it is necessary to establish parameters, such as the training data, the type of neural network, connections, number of layers, activation functions, propagation rules, etc.

Various methods exist to train these networks to produce a specific output for a specific input. Among the current training methods, we have error propagation, which consists of adjusting the network by adapting the weights of each neuron. The use of the partial derivative makes it possible to know in which direction we must modify the weights of our neural network to have the requested output. Also a genetic algorithms are used to train a neural network [19]. By training

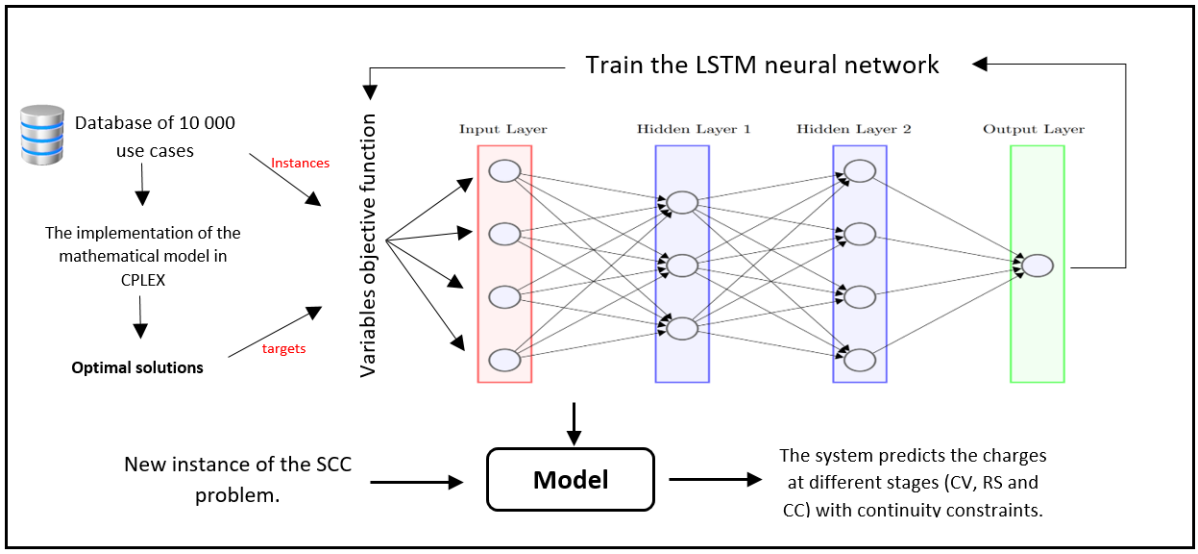


Fig. 3. The Evolutionary Strategy for the SCC.

these networks on a set of data for which the correct output is known, the network will return the appropriate results for similar data. We trained our LSTM on data from 10 000 use cases (instances), where the correct output was generated by the CPLEX based on the mathematical model that we proposed above.

Regarding the architecture, there are several architectures in the literature, but just some differences between them in the neural network language model. The architecture that we propose to solve the steel continuous casting problem was obtained after several tests and it is the one that which gave best results. It can be summarized as follows:

Input sequence represents the input of our LSTM, in our implementation it is encoded by an encoding from 1 to Sl_3 where Sl_3 is number of the sequences $Seql_3$ to process on the machine CC_{l_3}. Note that the sequenes are generated in 128 batches with 5% diversity, where the purpose of adding diversity in the learning stage is to give flexibility to the model and to avoid overfitting. The model predicts the assignment of sequence loads from different stages (CV, RS and CC) with continuity constraints.

The network topology, an architecture with three hidden layers has been applied, each with 256 LSTM units. Each cell LSTM use the "relu" activation function (to have faster convergence compared to other activation functions as tanh or sigmoid). However the output layer, the "softmax" activation function is used to generate the correct normalized probability value in order to select sequential charges in different stages (CV, RS and CC).

As a learning criterion, Adam optimizers is used with a precision of 0,001.

The following algorithm summarizes the main steps :

Algorithm 1 : The proposed LSTM model.

- According to the unified law on [0;1], the weights are initialized by random drawing.
- Codage : list all sequence charges, represent each one by a number from 1 to Sl_3.
- Creating a LSTM network : create four layers, where the three hidden layers have 256 LSTM units for each one. The output layer uses the "softmax" activation function to predict the sequence charges of different stages (CV, RS, and CC) with continuity constraints.

while counter $\leq$ iterMax **do**
1. Generate 128 batches with 5% noise.
2. Train the model.
3. The learning rate of the Adam optimizer is 0.001.
4. Update each weight

end while

VI. Experimental Tests

All experiments were performed on the Google colab under GPU, namely CPLEX 12.6 was used to determine the optimal solutions for the SCC instances in order to prepare the training data and also to compare with the results of the LSTM.

A. The Learning and Test Rate:

The learning rate gives an idea of the quality of the model. In Fig. 4, we present a graph that represents the learning rate (93%) and the test rate (91%) per epoch.

As can be seen from the convergence of the cost function (Fig. 5), our model performed well and the fact that there is not a large discrepancy between the loss of the training and the loss of the test allows us to conclude that we do not have an overfitting. We have also to mention that the learning rate is 93% and the test rate is 91%.

B. Test Instances

For the instances taken form the literature, we compare our LSTM with the CPLEX solutions, the proposed RNN with LSTM cells has improved the total makespan Sol_{max} and was

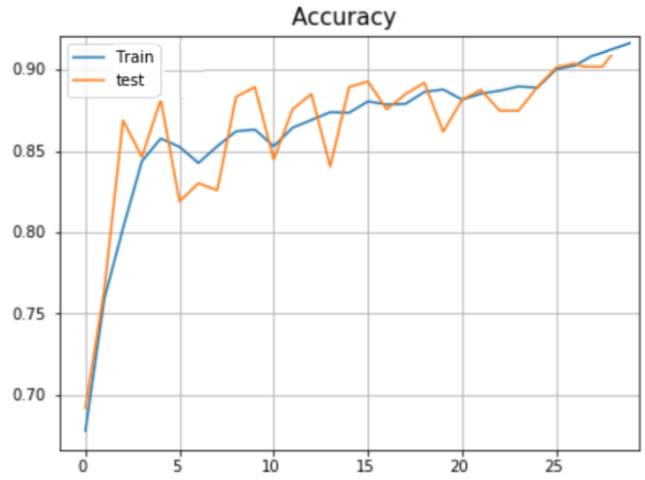


Fig. 4. The Learning and Test Rate per Epoch.

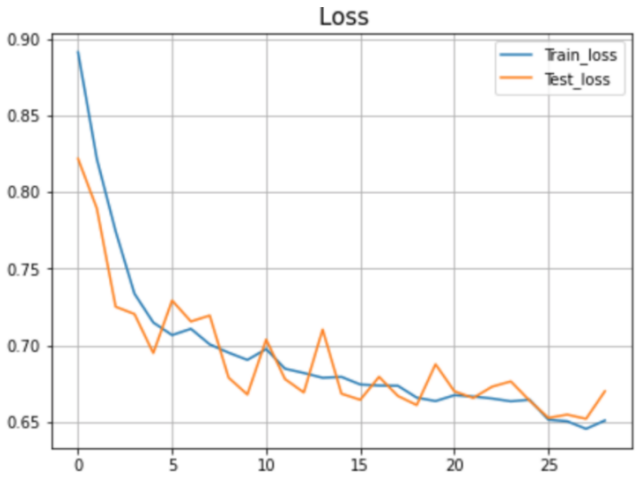


Fig. 5. Convergence of the Cost Function per Epoch.

TABLE I. RESULTS FOR DIFFERENT SEQUENCES SIZES WITH 3 CC AND FOR SEVERAL INTERVALS $[\lambda^{min}_{Seql_3}, \lambda^{max}_{Seql_3}]$

Order: [.][.][.][.]	Sol_{max} CPLEX	Time CPU (s)	Sol_{max} LSTM	Time LSTM(s)
[15, 10, 12][5, 5, 5]	1559.00	5.75	1133.23	1.0
[15, 14, 30][5, 5, 5]	1807.01	16.55	1721.10	1.3
[10, 13, 17, 8][5, 5, 5, 5, 5] [18, 11, 23, 15, 27]	1756.72	540.99	1440.67	1.5
[5, 5, 5, 5, 5, 5, 5, 5, 5, 5]	2506,04	2105.31	2420	1.7
[5][5][5]	427.58	1.09	281.39	0.17
[10, 10][10, 10][10, 10] [5, 15, 5, 10][10, 10, 10, 5]	706.07	21.45	651.26	1.1
[5, 5, 5, 5]	1239.34	790.23	1075.37	1.68
[5, 10, 5, 10, 5][10, 10, 10, 5, 5] [5, 5, 5, 5, 5]	1678.23	3367.86	1210.16	1.69
[10, 20, 5, 30][20, 5, 10, 5] [15, 5, 10, 10]	-	-	2910.54	1.57
[20, 10, 5, 40][30, 20, 15][20, 15]	-	-	3372.58	1.63
[35, 15, 25][10, 19, 15, 4, 5] [8, 35, 17]	-	-	3375.44	1.93
[15, 15, 25][20, 14, 25, 11, 5] [15, 35, 17, 10, 30]	-	-	4841.10	2.32

TABLE II. RESULTS ON SEQUENCES SIZES UNTIL 10 CC

Number of CC	Total number of charges	Sol_{max} (LSTM)	Time (s)
4	280	4320.28	1.21
4	340	4550.91	2.30
4	470	6330,13	2.42
4	520	6955.39	3.26
4	600	7615.39	3.62
5	235	3872.15	1.12
5	315	4226.28	3.15
5	400	4712.15	3.91
5	625	6980.50	5.76
5	865	10101.45	6.74
6	366	4821.67	3.12
6	515	6316.53	4.75
6	700	7005.22	5.36
8	450	5924.67	3.48
8	635	6473.31	6.85
8	965	8567.92	7.83
10	368	2365.86	2.10
10	823	8391.50	6.34
10	1256	10341.19	8.90

up to 1000 times faster than the CPLEX on CPU as shown in Table I.

Table II shows our tests on randomly generated large instances for which no solution is known. These instances span a large number of charges until 10 CC machines at the last stage. As an example, for the following batches size of a 4 CC system [30,25,20,25][25,30,15,25,15] [30,35,30,20,30 [40,20,55] (number of charges per each sequence) with different $[\lambda^{min}_{Seql_3}, \lambda^{max}_{Seql_3}]$ ranges of setup times, the CPLEX fails to solve it. However, the LSTM runs it on 1.42 seconds and gave a solution with Sol_{max} = 6330,13. This allows us to say that the obtained numerical results show the efficiency of the proposed recurrent neural network with LSTM cells with a total success of solving all the instances.

VII. CONCLUSION

In this paper, we have implemented a recurrent neural network with LSTM cells in order to solve the SCC with intersequence dependent setup times and dedicated machines at the last stage known as one of the harder problem in scheduling. Especially, we have shown that with RNN with LSTM cells, one can tackle very large instances arising in complex industrial systems where the number of sequences, of charges or of any devices type (CV, RS or CC) is bigger than 10. Better solutions are obtained with better quality and execution time. The performances of the proposed model are very interesting such that the success rate is 93% and able to resolve large instances while the traditional approaches are limited and fail to resolve very large instances.

One of the future works that we intend to develop is to generalize the approach on a cluster of GPUs in order to deal with more complex and robust cases and to enable solving very large size instances in order to improve the quality of the up to day known solutions. Also, we intend to generalize our approach to similar SCC problems in particular or to solve very complex hybrid flowshops in general. Another feature that could also be envisaged is the lagrangean relaxation for typical hard constraints that we could relax.

ACKNOWLEDGMENT

The author would like to thank the Editor-in-chief and anonymous reviewers for their comments and suggestions to improve the quality of the paper

REFERENCES

[1] Aqil S. and Allali K. Two efficient nature inspired meta-heuristics solving blocking hybrid flow shop manufacturing problem. *Engineering Applications of Artificial Intelligence*, *100*, 104196, (2021).

[2] Atighehchian A., Bijari M. and Tarkesh H. A novel hybrid algorithm for scheduling steel-making continuous casting production. *Computers and Operations Research*, *36(8)*: 2450-2461, (2009).

[3] Basu, S. and Dutta, G. A Survey of the Non-Optimization techniques used in an integrated steel plant. *Management Dynamics*, *6*: 33–68, (2006).

[4] Berrajaa, A. and Ettifouri, E. H. The Recurrent Neural Network for Program Synthesis. *In International Conference on Digital Technologies and Applications*, 77-86, (2021).

[5] Blazewicz J., Ecker K. H., Pesch E., Schmidt G. and Weglarz J. Scheduling computer and manufacturing processes. *springer science and Business media*, (2013).

[6] Cappel J., Kaiser H. P. and Schlüter J. Time management at the HKM-Huckingen BOF-Shop. *In Proceedings 4 th European Oxygen Steelmaking Conference (EOSC)*, *Vol.12*: 15.5, (2003).

[7] Chakraborti N., Kumar R. and Jain D. A study of the continuous casting mold using a pareto-converging genetic algorithm. *Applied Mathematical Modelling*, *25(4)*: 287-297, (2001).

[8] Chang, P. C., and Chen, S. H. Integrating dominance properties with genetic algorithms for parallel machine scheduling problems with setup times. *Applied Soft Computing*, *11(1)*: 1263-1274, (2011).

[9] Chen Z. Control of constrained moving-boundary process with application to steel continuous casting. *(Doctoral dissertation, University of Illinois at Urbana-Champaign)*, *36(8)*:2450-2461, (2020).

[10] Ferretti I., Zanoni S. and Zavanella L. Production–inventory scheduling using Ant System metaheuristic. *International Journal of Production Economics*, *104(2)*: 317-326, (2008).

[11] Foumani S. N. A., Guo C. and Luk W. An Analysis of Alternating Direction Method of Multipliers for Feed-forward Neural Networks. *arXiv preprint arXiv:2009 :02825*, (2020).

[12] Han D., Tang Q., Zhang Z., Yuan L., Rakovitis N., Li D. and Li J. An Efficient Augmented Lagrange Multiplier Method for Steelmaking and Continuous Casting Production Scheduling. *Chemical Engineering Research and Design*, *168*: 169-192, (2021).

[13] Hao J., Liu M., Jiang S. and Wu C. A soft-decision based two-layered scheduling approach for uncertain steelmaking-continuous casting process. *European Journal of Operational Research*, *244(3)*: 966-979, (2015).

[14] Helal, M., Rabadi, G. and Al-Salem, A. A tabu search algorithm to minimize the makespan for the unrelated parallel machines scheduling problem with setup times. *International Journal of Operations Research*, *3*: 182–192, (2006).

[15] Hochreiter S. Long Short-Term Memory. *Neural Computation*, *9(8)*, (1997).

[16] Hohn W., Konig F. G., Mohring R. H. and Lubbecke M. E. Integrated sequencing and scheduling in coil coating. *Management Science*, *57(4)*: 647-666, (2011).

[17] Ko C. H. and Wang S. F. Precast production scheduling using multi-objective genetic algorithms. *Expert Systems with Applications*, *38(7)*: 8293-8302, (2011).

[18] Kumar V., Kumar S., Tiwari M. K. and Chan F. T. S. Auction-based approach to resolve the scheduling problem in the steel making process. *International journal of production research*, *44(8)*: 1503-1522, (2006).

[19] Lamos-Sweeney, J. D. Deep learning using genetic algorithms. *Rochester Institute of Technology*, (2012).

[20] Lee H. S., Murthy S. S., Haider S. W. and Morse D. V. Primary production scheduling at steelmaking industries. *IBM Journal of Research and Development*, *40(2)*: 231-252, (1996).

[21] Lee T. and Loong Y. A review of scheduling problem and resolution methods in flexible flow shop. *International Journal of Industrial Engineering Computations*, *10(1)*: 67-88, (2019).

[22] Michael L. P. Scheduling: theory, algorithms, and systems. *Springer*, (2018).

[23] Naphade K. S., Wu S. D., Storer R. H. and Doshi B. J. Melt scheduling to trade off material waste and shipping performance. *Operations Research*, *49(5)*: 629-645, (2001).

[24] Pan Q. K., Wang L., Mao K., Zhao J. H. and Zhang M. An effective artificial bee colony algorithm for a real-world hybrid flowshop problem in steelmaking process. *IEEE Transactions on Automation Science and Engineering*, *10(2)*: 307-322, (2012).

[25] Pan Q. K. An effective co-evolutionary artificial bee colony algorithm for steelmaking-continuous casting scheduling. *European Journal of Operational Research*, *250(3)*: 702-714, (2016).

[26] Peng K., Pa, Q. and Zhang, B. An improved artificial bee colony algorithm for steelmaking–refining–continuous casting scheduling problem. *IEEE Chinese journal of chemical engineering*, *26(8)*: 1727-1735, (2018).

[27] Tan Y., Zhou M., Zhang Y., Guo X., Qi L. and Wang Y. Hybrid scatter search algorithm for optimal and energy-efficient steelmaking-continuous casting. *IEEE Transactions on Automation Science and Engineering*, *17(4)*: 1814-1828, (2020).

[28] Tang L., Liu J., Rong A. and Yang Z. A review of planning and scheduling systems and methods for integrated steel production. *European Journal of Operational Research*, *133(1)*: 1-20, (2001).

[29] Tang L. and Liu G. A mathematical programming model and solution for scheduling production orders in Shanghai Baoshan Iron and Steel Complex. *European Journal of Operational Research*, *182(3)*: 1453-1468, (2007).

[30] Tang L., and Wang G. Decision support system for the batching problems of steelmaking and continuous-casting production. *Omega*, *36(6)*: 976-991, (2008).

[31] Wu X., Jin H., Ye X., Wang J., Lei Z., Liu Y., ... and Guo Y. Multiscale Convolutional and Recurrent Neural Network for Quality Prediction of Continuous Casting Slabs. *Processes*, *9(1)*: 33, (2021).

[32] Xu Z., Zheng Z., and Gao X. Energy-efficient steelmaking-continuous casting scheduling problem with temperature constraints and its solution using a multi-objective hybrid genetic algorithm with local search. *Applied Soft Computing*, *95*: 106554, (2020).

[33] Yang J. M., Che H. J., Dou F. P. and Zhou T. Genetic algorithm-based optimization used in rolling schedule. *Journal of Iron and Steel Research International*, *15(2)*: 18-22, (2008).

Predicting Stock Closing Prices in Emerging Markets with Transformer Neural Networks: The Saudi Stock Exchange Case

Nadeem Malibari[1], Iyad Katib[2], Rashid Mehmood[3]
Department of Computer Science, Faculty of Computing and Information Technology[1,2]
High Performance Computing Center[3]
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—Deep learning has transformed many fields including computer vision, self-driving cars, product recommendations, behaviour analysis, natural language processing (NLP), and medicine, to name a few. The financial sector is no surprise where the use of deep learning has produced one of the most lucrative applications. This research proposes a novel fintech machine learning method that uses Transformer neural networks for stock price predictions. Transformers are relatively new and while have been applied for NLP and computer vision, they have not been explored much with time-series data. In our method, self-attention mechanisms are utilized to learn nonlinear patterns and dynamics from time-series data with high volatility and nonlinearity. The model makes predictions about closing prices for the next trading day by taking into account various stock price inputs. We used pricing data from the Saudi Stock Exchange (Tadawul) to develop this model. We validated our model using four error evaluation metrics. The applicability and usefulness of our model to fintech are demonstrated by its ability to predict closing prices with a probability above 90%. To the best of our knowledge, this is the first work where transformer networks are used for stock price prediction. Our work is expected to make significant advancements in fintech and other fields depending on time-series forecasting.**

***Keywords*—Stock price prediction; time-series forecasting; transformer deep neural networks; Saudi Stock Exchange (Tadawul); financial markets**

I. Introduction

We have come a long way in developing our societies, improving and optimising every task and thing we do, and artificial intelligence (AI) is at the heart of these endeavours [1], [2]. Machine and deep learning-based AI has revolutionised many aspects of our daily activities, be it healthcare [3], [4], transportations [5], [6], big data [7], distance learning [8], disaster management [9], risk prediction in aviation systems [10], DNA profiling [11], smart cities [12], [13], and more. The use of machine and deep learning in the financial sector is one of the most lucrative tasks. Forecasting time-series data is an important topic that plays a key role in analysis, decision-making, and resource management in many industrial sectors. For example, in the financial sector, forecasting based on historical data can be helpful for investors in maximizing return and reducing risk on investments [14], [15]. Many works have been reported on the use of AI for the financial sector, such as the use of multilayer perceptrons (MLP) for NSADA stock index [16], the use of stacked autoencoders for US stock forecasting [17], and the use of Long short-term memory network (LSTM) to predict the closing prices of iShares MSCI United Kingdom index [18] (for further motivation on the subject, see Section II).

A time-series forecast is a way of determining future values based on historical experience. Correlational data is used for this process, either time-based correlations (years, months, weeks, etc.) or sequential correlations, for gaining insights that inform decisions. A range of methods has been developed to predict, ranging from traditional to machine-learning approaches. Despite their wide usage, traditional time-series prediction methods such as auto-regression (AR), Seasonal Naïve, ETS, and integrated moving average ARIMA are designed to fit each time-series separately [19]. Moreover, practitioners should learn how to select specific trends, seasonal components, and other data components manually, especially for financial data series with highly nonlinear and fluctuating data. These drawbacks have limited their applications in advanced large-scale time-series prediction tasks.

The challenges mentioned above can be overcome by algorithms that can capture the patterns in the data and the dynamics underlying them. In deep neural networks, continuous developments have led to breakthroughs that are proposed as another alternative. An array of deep neural network architectures has been applied to time-series models to understand trends and patterns by learning from ground truth data. However, many challenges remain. For example, while a Recurrent Neural Network (RNN) can model and process sequential and time-series data, its gradient vanishing and exploding properties prevent them from detecting long-term dependencies (relationship between entities that are several steps apart). In real-world forecasting, there are long-term and short-term repeating patterns [20], which means that complex RNN models are required to analyze long-term time series and study long-term effects. Therefore, long short-term memory (LSTM) models have been proposed to improve the standard RNN model for time series analysis. Theoretically, they are explicitly geared towards minimizing long-term dependency problems. However, according to [21], LSTM has an adequate context size of 200 tokens on average, but they are only able to distinguish 50 tokens within a context, suggesting that even it is incapable of capturing long-term trends. Furthermore, RNNs and all their variants use mostly sequential operations, thus cannot benefit from the performance advantages offered by modern GPUs.

Rather than RNNs, the next big step was a completely new architecture – Transformer [22] utilizes attention mechanisms that leverage self-attention mechanisms to process the entire sequence of data. The transformer architecture is the most prevalent model for natural language modelling and has proven quite successful in several other applications. The complexity of the space corresponding to self-attention can grow quadratically as sequence length increases; for this reason, self-attention cannot be extended to extremely long sequences [20]. The quadratic complexity of computing poses a significant challenge when forecasting time series with long-term solid dependence and fine granularity. Researchers had the same challenges adapting transformers from language to computer vision applications due to pictures containing more significant amounts of information than sentences. However, they are able to replace this quadratic computational complexity with a linear computational complexity to image size.

In this work, we specifically delve into adapting the computer vision transformer model [23] to time series forecasting. We propose a novel fintech machine learning method that uses Transformer neural networks for stock price predictions. In our method, self-attention mechanisms are utilized to learn nonlinear patterns and dynamics from time-series data with high volatility and nonlinearity. Our Contributions follow.

- We propose a novel predictive Transformer based model with divided time series data into patches for predicating future value. Regardless of how complex a situation is, our proposed method can discover the broad conditional probability distribution of the future values.
- The model makes predictions about closing prices for the next trading day by taking into account various inputs, Open, High, Low, Volume, and Closing Prices. We used pricing data from the Saudi Stock Exchange (Tadawul) to develop this model. We validated our model using a range of metrics; Mean Absolute (MAE), Square (MSE), Root MSE (RMSE), and Percentage (MAPE) Error.
- The applicability and usefulness of our model to fintech are demonstrated by its ability to predict closing prices with a probability above 90%.

Novelty: As mentioned earlier, transformers are relatively new and while these have been applied for NLP and computer vision, they have not been explored much with time-series data. Our work is expected to make significant advancements in fintech and other fields depending on time-series forecasting. To the best of our knowledge, this is the first work where transformer networks are used for stock price prediction.

The structure of this paper is as follows. We discuss related research in Section II as well as past innovations using deep learning for stock forecasts. The methodology for this study, the dataset, and the transformer model with divided space are described in Section III. This section also provides details of data prepossessing, and hyperparameters selection. Section IV provides the results and analysis. Section V concludes and provides directions for the future work.

II. Related Work

In the not-so-distant past, Neural Networks (NN) were criticized by many forecasting practitioners as not suitable and not being competitive in forecasting fields [24]. Consequently, practitioners have usually selected statistical methods that were considered more straight forward to apply [19]. However, with the ever-increasing availability of data, neural networks (NNs) and deep learning have revolutionized and achieved remarkable success in many research fields and practical scenarios, including medical predictions, NLP, image recognition, etc. Because of their capabilities to identify complex nonlinear patterns and explore unstructured relationships without hypothesizing them a priori. These technological breakthroughs have attracted significant attention from the enthusiasts' researcher community presenting many complex novel NN architectures on time series forecasting. Over recent decades, plenty of works and research exist where deep learning is used for forecasting. There is a possibility to predict stock price changes and foreign exchange rates according to [14]. As a result, AI applications are becoming increasingly popular among investors to increase returns and reduce the risk [15].

Selvin et al. [25] illustrated how deep neural network architectures can capture hidden dynamics and can be used to forecast. Guresen, Kayakutlu, and Daim [16] predicts the NASDA stock index by using multilayer perceptrons (MLP), dynamic, and hybrid artificial neural networks. Using a stacked autoencoder and deep neural network, Takeuchi and Lee [17] obtains an accuracy of 53.36 % when predicting the US stock direction.

Since their inception in 2014 by Hochreiter and Schmidhuber [26], the Long short-term memory network (LSTM) introduced is a variation of the Recurrent neural network model (RNN), which is the most commonly used architecture for sequence prediction problems [8]. In contrast to RNN, LSTM networks are capable of detecting long-term dependency and can prevent gradient vanishing. It utilizes historical information via the input, forget and output gates. In their study, Nikou, Mansourfar, and Bagherzadeh [18] predict the closing prices of iShares MSCI United Kingdom index using an LSTM model. The model performed significantly better than the ANN, Support Vector Regression (SVR), and RF models. LSTMs are utilized in another study by [27] in order to forecast future stock returns. Also, an Autoregressive Integrated Moving Average (ARIMA) and an LSTM model were utilized to improve forecast accuracy [28]. According to Nelson, Pereira, and De Oliveira [29], the average accuracy for predicting the direction of some stocks traded on the Brazilian stock exchange could reach up to 55.9% with the LSTM model.

Because of its powerful pattern recognition ability, the convolutional neural network (CNN) is a variation of the multilayer perceptron (MLP). Its use has extended increasingly for time-series forecasting. The work by [30], [31], and [32] used CNN to predict stock trends. Ugur Gudelek, Arda Boluk, and Murat Ozbayoglu [32] have also experimented with 2D CNN for trend detection. The model performance evaluation has 72% accuracy values and looks promising.

A comparison study of differences between Multi-layer Perceptron (MLP), Convolutional Neural Network, and Long Short-Term Memory (LSTM) was performed by [33]. They

TABLE I. SUMMARY OF RELATED WORKS

Research	Model Architecture								Dataset	Accuracy
	ANN	MLP	RF	SVR	RNN	CNN	LSTM	Transformer		
Selvin et al. [25]					✓	✓	✓		1721 NSE	2.36
Guresen et al. [16]	✓	✓							NASDAQ	MSE: 1472
Nikou et al. [18]	✓		✓	✓			✓		iShares MSCI	MSE: 0.09396
Naik et al. [27]					✓		✓		Indian NSE	RMSE: 23.78
Nelson et al. [29]							✓		BMF Bovespa	55.9%
Gudelek et al. [32]						✓			ETF	72%
Persio et al. [33]		✓				✓	✓		SP500	MSE: 0.2491
Our Study								✓	Tadawul	90%

adopted a sliding window approach, using 30 previous days to predict the value of the 31st day using historical data of the S&P500 index from 1950 to 2016. CNN's results are exceptional even without the use of additional features such as technical analysis.

Recently, the well-known self-attention-based Transformer [22] was proposed for sequence modeling is the most prevalent model for natural language modeling and has proven quite successful in several other applications such as as translation, speech, image generation, and music [22], [34], [35]. The extension of self-attention to extremely long sequences would, however, be computationally prohibitive since space complexity increases quadratically with the sequence length [20]. However, Vision Transformer (ViT) [23] and TimeSformer (Time-Space Transformer) [36] offer entirely new architectures for image classification, and video understanding based solely on Transformers eliminating the problems associated with long sequences. In particular, ViT divides an image into patches (also called tokens) with fixed length; then following the practice of using transformers to model language, ViT then uses transformer layers to model the relationship among tokens for classification. The TimeSformer, on the other hand, translates the input video into a sequence of image patches derived from the individual frames. The model then captures the semantic information about each patch through comparison with those of the other patches. This allows TimeSformer to capture the space-time dependency based on the whole video. Transformers' recent success in natural language processing (NLP) has motivated researchers to implement this model in computer vision applications and tasks.

Table I summarises the related works discussed in this section. It lists the various ML models that the researchers have used for stock price prediction along with the respective datasets used and the model accuracies reported in the respective works. The most commonly used architecture for problems involving stock price prediction is the LSTM. It can detect long-term dependency and prevent gradient vanishing to some extent. However, LSTM accuracy is much less than the convolutional neural network (CNN) because of CNN's powerful pattern recognition ability. The accuracy metrics are reported in the table if these are provided by the researchers, otherwise, we reported the numeric value from the article without the accuracy metric name. As shown in the table, the best result achieved is 72% accuracy. Our transformer model with its attention features has provided 90% or higher accuracy. We have kept the content in the table to the minimum due to the space issue, please refer to the listed works for details.

III. METHODOLOGY, DATASETS AND MODEL DESIGN

The objective of our study is to predict the subsequent and future closing of the trades in the Saudi Stock Exchange (Tadawul). We use a transformer-based temporal model architecture. In this section, we describe our methodology, Transformer neural network model design, datasets, preprocessing, and validation metrics.

We first present an overview of our methodology in Section III-A. The transformer-based temporal model architecture is described in Section III-B. The the Saudi Stock Exchange (Tadawul) datasets are explored in Section III-C. The data modelling methodology using transformer neural networks is summarised in Section III-D. In Section III-E, we describe the preparation of the dataset, including data splitting, normalization, and feature selection. We discuss the hyperparameter configuration for the model. Section III-F describes the concept of sliding window for framing the dataset. Section III discusses hyperparameter configuration of our model.

A. Methodology Overview

The overall methodology we have adopted is depicted in Fig. 1. It consists of seven main phases as highlighted in the figure. The first process involved extracting Saudi Stock Exchange (Tadawul) data, followed by data cleaning and normalization. As a result of this procedure, we only get data that is appropriate for machine learning algorithms. We then select the four features (open, low, high, previous closing) that the model will use. Thereafter, the data are sorted into non-overlapping batches, which are then fed into the model until performance measures are optimized. Ultimately, the optimized model is used to forecast future closing prices for unseen stock data.

B. Transformer Neural Network Architecture

A significant influence on our architecture is a vision transformer (ViT) [36] using divided space. The vision transformer (ViT) is among the first attempts to apply the outstanding performance of Transformers [22] to image classification tasks rather than natural language processing. The vision transformer (ViT) model, which comprises three main elements: a linear layer for patch embedding, a stack of transformer blocks with multi-head self-attention and feed-forward layers, and a linear layer classification score prediction.

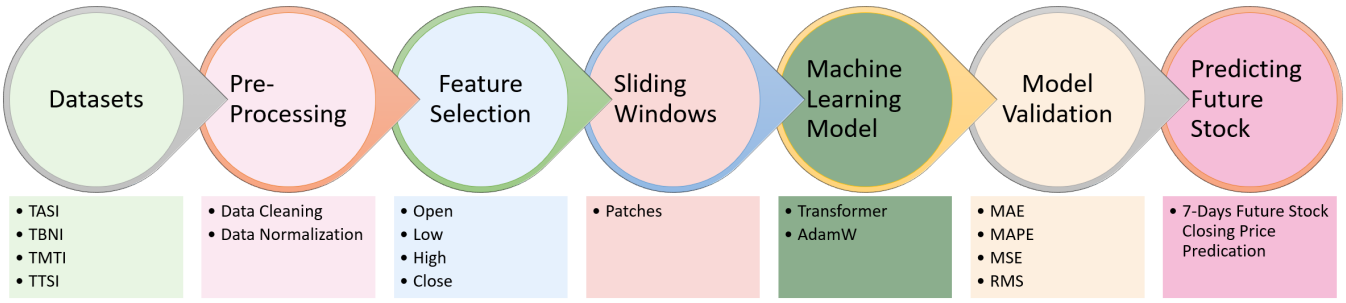


Fig. 1. An Overview of Our Methodology.

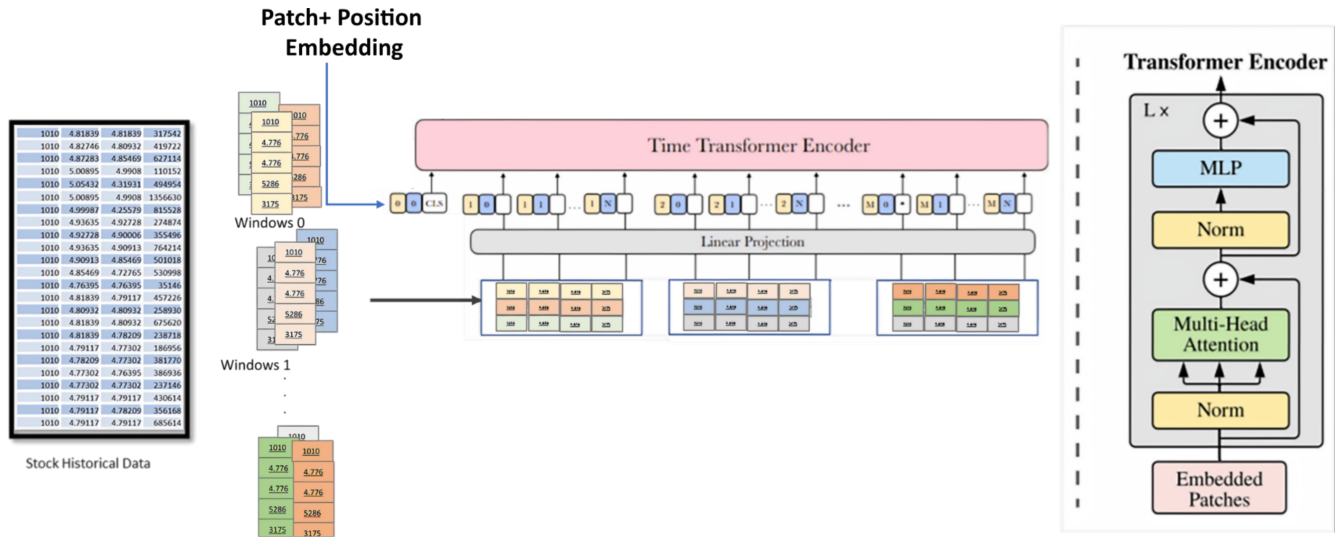


Fig. 2. Proposed Transformer Encoder Model Overview (Illustration Inspired by [22]).

An overview of our suggested model is depicted in Fig. 2. The Vision Transformer (ViT) model serves as the basis for our predictive model. Our suggested model added one more component to ViT architecture. Its primary purpose is to create sliding windows from historical data. Since daily trading volumes on the stock market are substantial, historical data on the market can be challenging to manipulate, and manipulating it can cause a computational burden. Furthermore, the effect of more recent data on a training model is greater than that of older data [37]. Braverman et al. [38] developed a sliding-window method that utilizes recent data while disregarding older observations to solve this problem.

The range of data of interest is selected using a window. The sliding window represents a period that stretches backward in time from the present to the past. The sliding window is held steady (the number of data stays constant), and only the window is moved. Resulting, the training data volume is reduced while maintaining the model's efficiency and general usability [37].

In summary, Fig. 2 depicts our proposed model as follows. The historical data is split into windows and then those windows are divided into fixed-size patches. Linear embeddings are then applied to the patches, followed by position embeddings. Then we feed the resulting sequence of vectors to the Transformer encoder. As a standard approach, we add an extra token to the sequence of learnable tokens to perform prediction. The Transformer encoder diagram in Fig. 2 was inspired by [22].

C. Datasets

The Saudi Stock Exchange (Tadawul) database contains stock trading information for more than 200 Saudi Arabian listed companies. The companies are grouped into sectors with different indices for each industry. The data we downloaded spans the period from 1993-01-02 through 2021-06-17 and consists of 772,189 trading days. Listed companies' and indices trading information includes their Open, High, Low, Volume, and Closing Price for each trading day. From the dataset, we extracted four indices to illustrate model capabilities and performance. These are Tadawul All Share Index (TASI), the Banks Index (TBNI), Materials Index (TMTI), and Telecommunication Services Index (TTSI).

Table II lists a small selection of the dataset. Specifically, it shows the trading information in the dataset for the TASI index for the period 1994-01-26 to 2021-07-01, which corresponds to 7311 trading days. The rows correspond to one trading day and contain the following features: the index column, the transaction date, the ticker code, High, Low, Volume, and Closing Price. Fig. 3 depicts the histograms of the closing price feature of the four datasets. There are four panels in the figure,

TABLE II. TASI SAMPLE DATA

	date	ticker	open	low	high	vol	close
0	1994-01-26	TASI	1751.71	1751.71	1751.71	312907	1751.71
1	1994-01-29	TASI	1751.71	1750.91	1751.71	204831	1750.91
...	...	...	...	...	...	...	...
7310	2021-06-30	TASI	11002.74	10940.44	11009.70	374658538	10984.15
7311	2021-07-01	TASI	10987.13	10968.11	11006.66	352200486	10979.05

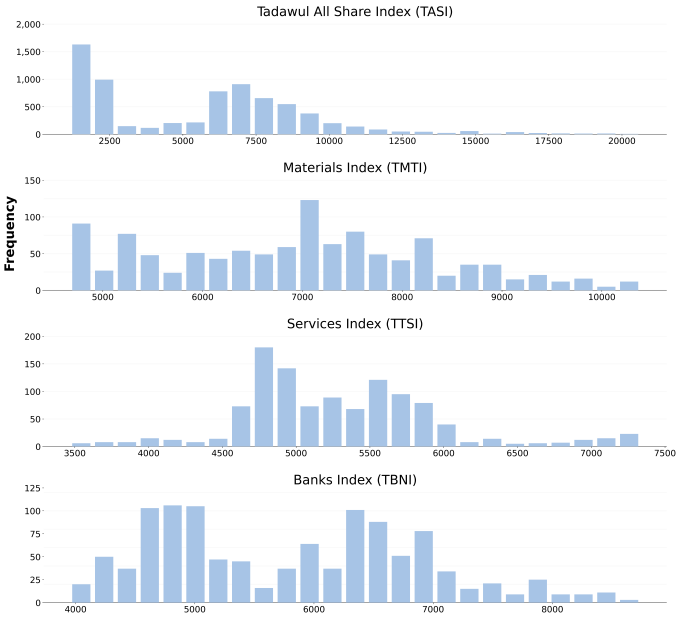


Fig. 3. The Histograms of the Closing Price for the Four Datasets.

each showing the histogram of its respective index. A display of the closing price is shown on the x-axis for each panel, grouped into 25 bins of equal width. Each bin is plotted as a bar whose height (the y-axis) indicates the number of closing prices (frequencies) occurrences in that bin.

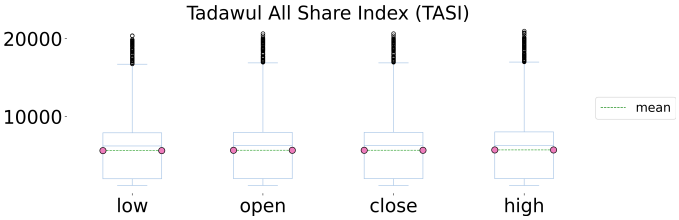


Fig. 4. TASI Boxplot.

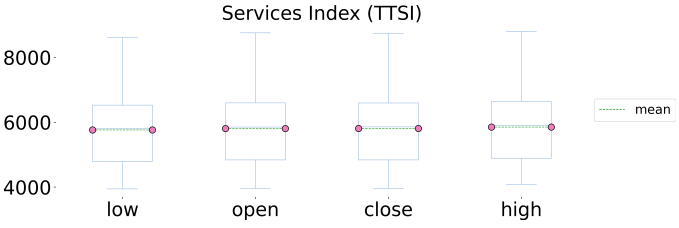


Fig. 5. TTSI Boxplot.

Fig. 4 to 7 depict the four features (open, low, high, close) of the four indices in our dataset as a boxplot. A boxplot is an in-depth statistical data analysis tool for gaining a broad perspective on the center and spread of the data distribution, which can assist with checking for errors and protecting other analyses. The median, interquartile range box, and whiskers are the primary elements of the boxplot to help understand the center and spread of the sample data. You'll see the green line representing the median in each box, which is the center of each feature. The interquartile range (the range between the third quartile and the first quartile) box, on the other hand, represents the middle 50% of the data and reflects how the data is distributed. The whiskers extend from both sides of the box (the bottom line is called lower whiskers, whereas the upper one is called higher whiskers). The whiskers denote the ranges for the bottom 25% and the top 25% of the data values, excluding outliers. Graphs that are skewed have the majority of data on the high or low side. Skewed graphs indicate that the data isn't normally distributed.

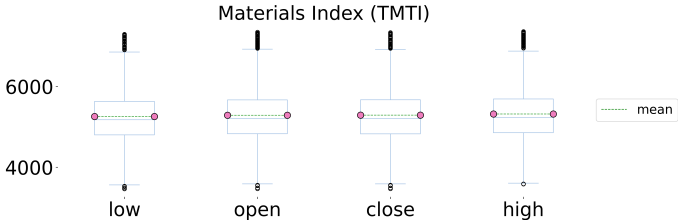


Fig. 6. TMTI Boxplot.

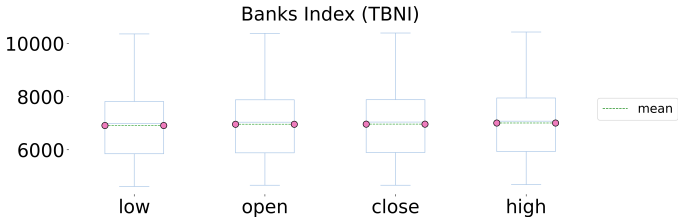


Fig. 7. TBNI Boxplot.

The data distribution for the TTSI, TMTI, and TBNI in the figures (Fig. 5 to 7) is almost normally distributed while it is positively skewed for the TASI index (Fig. 4). Moreover, any value greater than higher whiskers and less than lower whiskers values is an outlier and is represented in the figure as circles beyond the minimum and maximum values. Fig. 4, shows reasonable outliers points for TASI, which is expected as the closing of TASI is directly impacted by each and every listed company.

Fig. 8 highlights the correlation between the features (High, Low, Volume, and Closing Price), which is considered an essential step in the feature selection phase of data pre-

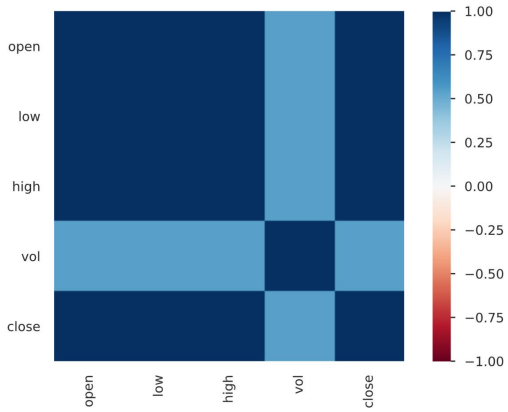


Fig. 8. TASI Correlation Matrix.

processing, especially if the data types of the features are continuous. As you can see in the figure, there is a high correlation between volume and the other features.

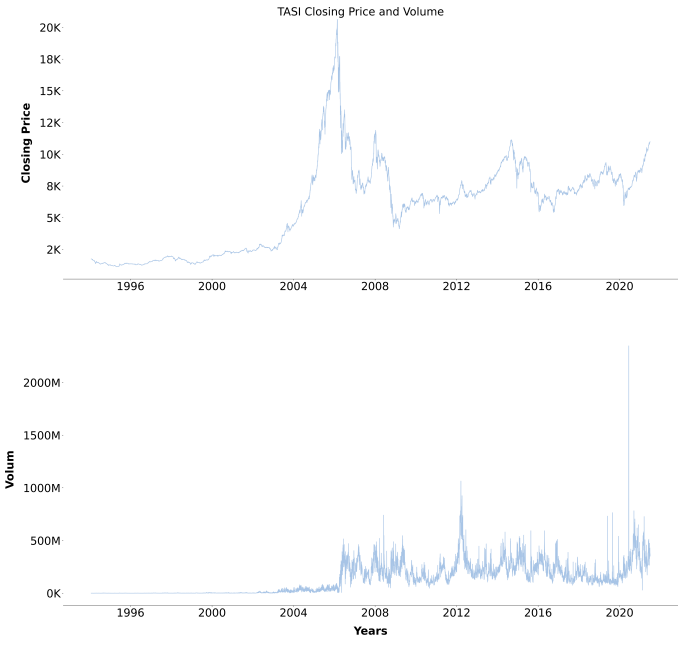


Fig. 9. TASI Closing Price and Volume.

Tadawul All Share Index (TASI) volume and closing figures are shown in Fig. 9.

D. Data Modelling Methodology

At first, the historical stock data of earlier days $X \in \mathbb{R}^{M \times F}$ consisting of M periods with F features (previous closing, opening, high ,low and volume) is split into a sequence of flattened 2D Windows $\mathbf{x}_w \in \mathbb{R}^{L \times F}$ of size $M-L$, where L is look back time interverls. Then the input window is divided into non-overlapping temporal patches of size $\mathbf{x}_p \in W \times (F \times 2)$.

Finally, following the protocol in ViT ,the patches $\mathbf{x}_p \in \mathbb{R}^{W \times (F \times 2)}$ are flattened forming a sequence of embeddings. Using learnable 1D position embeddings, we embed positional information into the patch embeddings so that all patches within a given window w are given the same temporal position. This allows the model to determine the temporal positions of patches.

E. Data Prepossessing

It is imperative to preprocess data in order to achieve good predictions. The indexes data were checked to determine whether the Tadawul Dataset contained inconsistencies. All the numerical data were normalized, and the missing values were removed. The open, high, low, volume and close prices were used to calculate the features, but information such as the stock code and stock name was omitted since they do not make sense. The following sections describe how the various preprocessing steps are implemented.

1) Splitting the Dataset: The training and test datasets are separated, similar to the ideas presented by [39]. We reserve apart from the end the training for validation from each time series. This approach is illustrated in the Fig. 10.

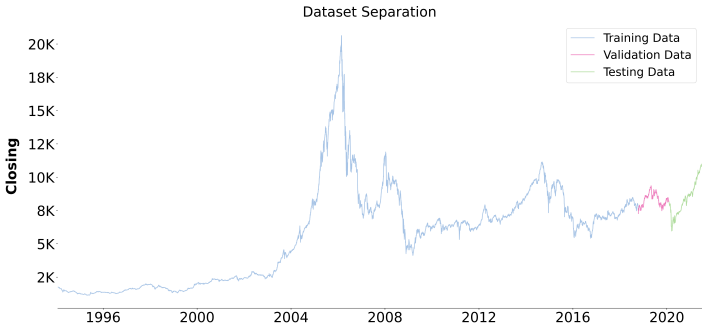


Fig. 10. Training, Validation and Testing Dataset Allocation.

2) Data Normalization: Normalization refers to the process of changing the range of values in a set of data. As we use prices and volume data, all the stock data must be within a typical value range. In general, machine learning algorithms converge faster or perform better when they are close to normally distributed and/or on a similar scale. Also, in a machine learning algorithm, the activation function, such as a sigmoid function, has a saturation point after which the outputs are constant [40]. As a result, when using model cells, the inputs should be normalized before being used. This process was done using MinMaxScaler methods of the scikit-learn library. When MinMaxScaler is applied to a feature, it subtracts the minimum value from each value in the feature and divides the range by the result. Thus, the range of a feature is the difference between the maximum and minimum values. In this way, MinMaxScaler preserves the shape of the original distribution. MinMaxScaler normalizes input values to be between [0,1].

3) Feature Selection: The downloaded data contains several features, including stock code, stock name, opening price, high price, low price, volume and closing price. Aside from some features that may not make any sense, these initial data have a lot of noise. For this reason, the data should be neglected when it is being trained. Based on [41] using open price, high price, low price, volume and close price, the input features will

yield a satisfactory result. Therefore, we have selected the first five features as our input and have neglected irrelevant data like stock names and stock codes.

F. Divided Space

At this stage, we apply the concept of the sliding window for framing the dataset. With a window size of 2, we use the data before two days to predict the subsequent day closing. The process is repeated until all data are segmented. Then, the framing dataset is further split into patches.

G. Hyperparameter Selection

A number of parameters, called hyperparameters, are usually included in all deep learning models (apart from Naïve Bayes) that need to be adjusted to optimize results [42]. The various hyperparameters used during training are summarized in Table III. The AdamW optimizer is used during training with a learning rate of 0.001 and a weight decay of 0.0001. We train the model for 500 epochs with early stopping and dropout to prevent overfitting using TensorFlow [43] library.

TABLE III. VARIOUS HYPERPARAMETERS USED IN THIS MODEL WITH THEIR VALUES

Hyperparameter	Value
Learning Rate	0.001
Optimizer	AdamW
Batch size	256
Epochs	500
Early stopping	Patience = 70 epochs Monitoring parameter = validation loss
Loss Function	MSE

H. Evaluation Metrics

Deep learning evaluation is categorized into accuracy index, financial index, and error-index [44]. Accuracy and financial index are widely used for prediction by classifying data (e.g., price direction prediction) and stock trading and portfolio management. On the other hand, error terms are frequently used in the evaluation for predicting numeric dependent variables (for instance, exchange rates or stock market predictions). The error terms evaluation rules compare the Real Data Y_{t+1} and the prediction data F_{t+1} using performance metrics: MAE, MSE, MAPE, and RMSE. Detailed information about the measures is provided below:

MSE is used to assess model performance based on the average error of forecasting. The formula of the MSE is given below:

$$\sum_{i=1}^{m} \frac{(Y_{t+i} - F_{t+i})^2}{m} \tag{1}$$

RMSE is one of the most commonly used error metrics in regression. It is equal to the square root of the MSE. RMSE is a measure of how spread out the residuals are. Based on the RMSE formula, it is possible to determine how well the data was focused around the optimal line. The optimal RMSE value is close to zero.

$$\sqrt{\sum_{i=1}^{m} \frac{(Y_{t+i} - F_{t+i})^2}{m}} \tag{2}$$

MAPE shows how much error was in the forecast. It measures how accurate the forecast is. A value of accuracy is calculated by subtracting the actual values from the average values of the previous period. The concept of MAPE is separated from the measurement level by data conversion. MAPE has minimal deviation in practice and cannot tell which direction the error is coming from. Ideally, MAPE should be close to zero. MAPE can be calculated using the following equation:

$$\frac{100}{m} \sum_{i=1}^{m} \frac{|Y_{t+i} - F_{t+i}|}{Y_{t+i}} \tag{3}$$

IV. RESULT AND DISCUSSION

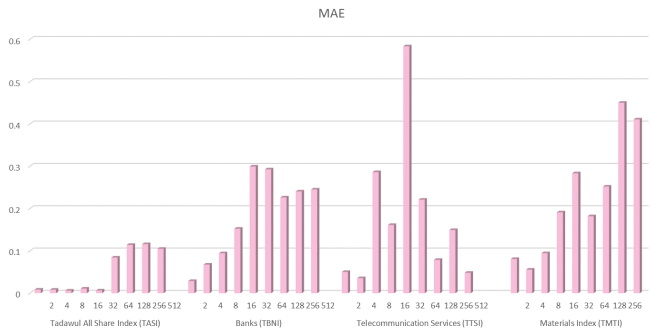
We now discuss the results beginning with results on model optimisation in Section IV-A, model validation in Section IV-B, and future stock closing price prediction in Section IV-C.

A. Model Optimisation

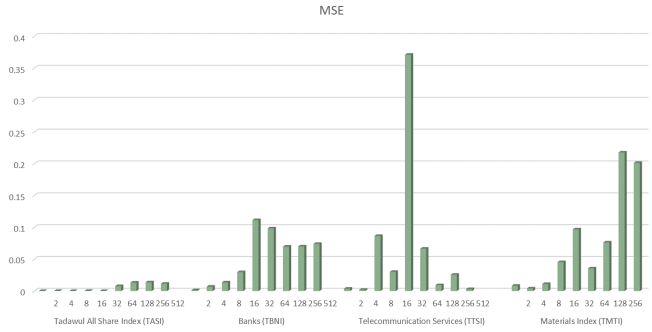
For this study, we experimented with different batch sizes and kept all the other hyper-parameters unchanged. Our predictive transformer model is implemented using TensorFlow written in Python. The study found that training smaller batches yielded better estimates but had a long training process. Our findings indicate that models perform better for all the four indices until the batch sizes reach around 4, with other batches not delivering significant performance improvements worth the time and effort devoted to estimating them. Fig. 11 depicts the four prediction performance measures MAE, MAPE, MSE, and RMSE results, respectively, for the Tadawul All Share Index (TASI), the Banks Index (TBNI), the Materials Index (TMTI), and the Telecommunication Services Index (TTSI).

Fig. 11a Mean Absolute Error (MAE) measure. It shows the batch size of the Tadawul All Share Index(TASI) is increased, and we find that forecast measures enhance until it reaches its best results at a batch size of 8 at a value of 0.0001, while it gets fluctuated for the other indices. Taking the MSE for Banks (TBNI) Index as an example, the optimal value for the TBNI index at batch size 2 is 0.0013, and it increases to 0.1114 at batch size 32. Thereafter, the index decreases with batch size. Similarly, when batch size over batch size exceeds eight, the Materials Index (TMTI) and Telecommunication Services Index (TTSI) also apply. Fig. 11b illustrates the mean square error (MSE) measure. It shows the batch size of the Tadawul All Share Index(TASI) is increased, and we find that forecast measures enhance until it reaches its best results at a batch size of 8 at a value of 0.0001, while it gets fluctuated for the other indices. Taking the MSE for Banks (TBNI) Index as an example, the optimal value for the TBNI index at batch size 2 is 0.0013, and it increases to 0.1114 at batch size 32. Thereafter, the index decreases with batch size. Similarly, when batch size over batch size exceeds eight, the Materials Index (TMTI) and Telecommunication Services Index (TTSI) also apply.

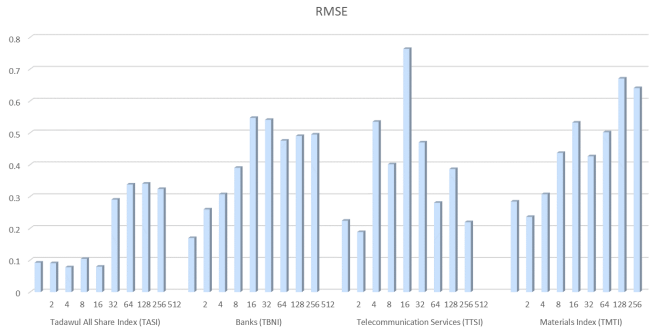
Fig. 11c, however, presents the root means square error (RMSE) of each batch size determined by the indices. For each batch size, each experiment was repeated 500 times(number of Epochs). As indicated in the figure, the RMSE is substantially higher for the Banks Index (TBNI), the Materials Index



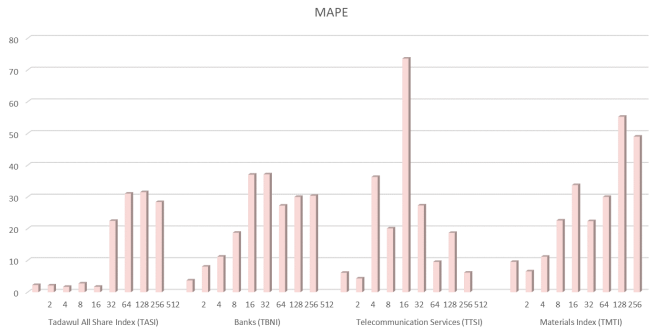
(a) MAE



(b) MSE



(c) RMSE



(d) MAPE

Fig. 11. Model Performance Versus Varying Batch Sizes.

(TMTI), and the Telecommunication Services Index (TTSI) mainly because of the number of trading days for these indices compared to the Tadawul All Share Index(TASI). Results show an evident dependence on the number of trading days. The RMSE ranges from 0.1697 to .5409 obtained for the Banks Index (TBNI), from 0.1885 to 0.7638 for Telecommunication Services (TTSI), from 0.2361 to 0.5323 for Materials Index (TMTI). Fig. 11d shows the Mean Absolute Percentage Error (MAPE) for each batch size for the four indices. Despite outperforming practically all other indices with a MAPE value of 1.681 for batch size 8, there is another batch size where TASI does just as well on this accuracy measure. Considering the effects of sampling (trading days) on results, it makes sense that the result would differ.

B. Model Validation

Fig. 12 to 15 depict the predicted versus actual closing price for the four datasets: Tadawul All Share Index (TASI), Banks (TBNI), Telecommunication Services (TTSI), and Materials Index (TMTI) indices. These graphs show the best results based on a comparison between the actual and forecasted stock prices (close prices). On each chart, orange and blue lines depict the actual values and predicted values, respectively. The plots provide the timelines of the whole dataset. Fig. 12 plots the closing prices for the TASI dataset for the period from early 1990s to 2021. Note in the figure that there is a relatively bigger difference between the actual and predicted values of the stock closing prices in the earlier period of the data. However, the differences get smaller for the later time periods. Overall, all the four figures show a reasonably small differences between the actual and predicted values, indicating a good model performance. The results of our study indicate that the proposed model is very effective in analyzing and capturing trends, as well as forecasting them accurately.

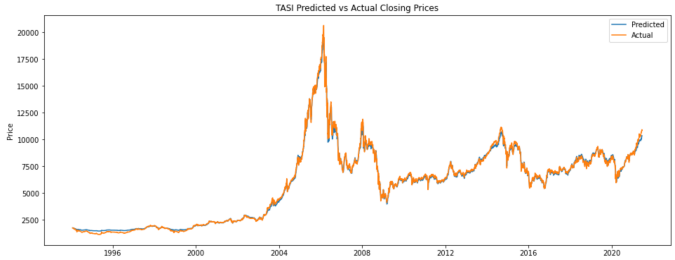


Fig. 12. TASI-Predicted vs Actual Closing Price for the whole dataset.

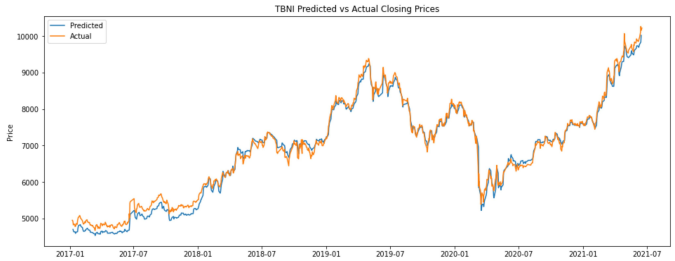


Fig. 13. TBNI-Predicted vs Actual Closing Price.

C. Predicting Future Stock Closing Prices

The next day's closing price of the selected stock is derived from the model prediction. Fig. 16 depicts the predicted and

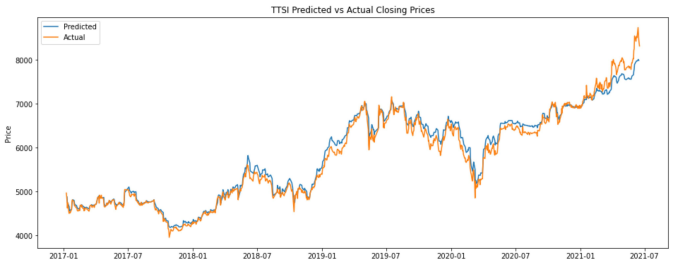


Fig. 14. TTSI-Predicted vs Actual Closing Price.

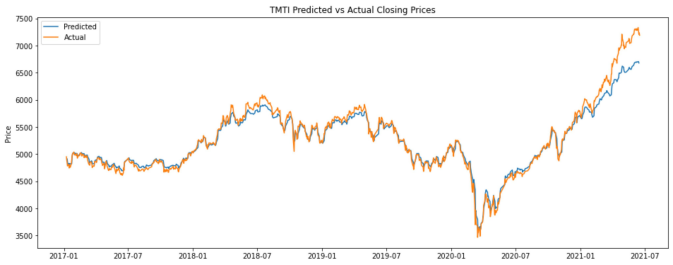


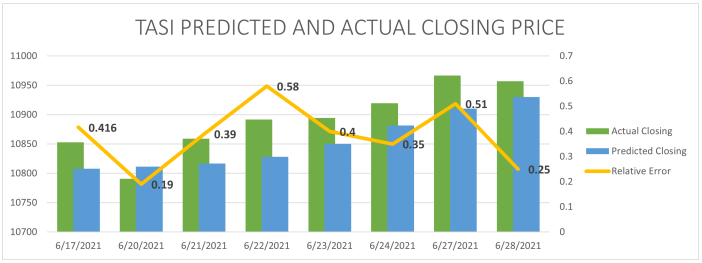
Fig. 15. TMTI-Predicted vs Actual Closing Price.

actual closing for eight trading days starting from 6/17/2021 till 6/28/2021 by using the model for the four indices TASI, TBNI, TTSI, and TMTI. The figure illustrates that the range of relative error fluctuation within the eight working days is between 0.19 and 0.58 for Tadawul All Share Index (TASI) and between 4.43 and 6.15 for the Banks index. As a result, the model accurately predicted the closing price of TASI and Bank with more than 99 and 94 percent, respectively. According to the model, TASI's closing price, for example, on 2021/06/17, will be 10807.94, while it was actually 10853.12 at the time. 45.18 points is a relatively small difference. In contrast, the relative error of Telecommunication Services (TTSI) fluctuated between 0.2, and 2.12, while Materials Index (TMTI) fluctuated between 5.25, and 7.33. Consequently, TMTI and TTSI closing prices were correctly predicted with more than 92, and 97 percent, respectively. The proposed model predicts the market closing price with a better than 90% accuracy, making it an exceptionally effective and practical model.

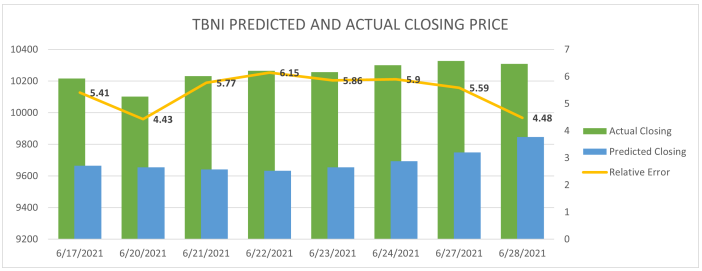
V. Conclusion and Future Work

We propose a transformer-based formalization model for stock price prediction. A significant influence on our architecture is a vision transformer (ViT) [36] using divided space. The vision transformer (ViT) is among the first attempts to apply the outstanding performance of Transformers. Using transformer network architectures with split time series into patches shows that hidden dynamics can be captured and predictions made reasonably. The model was trained using data from the Saudi Stock Exchange (Tadawul). As a result, we were able to predict the stock price of the TadawulAll Share Index (TASI), Telecommunication services Index (TTSI), Banks Index (TBNI), and Materials Index (TMTI) with accuracy that exceeds 90%.

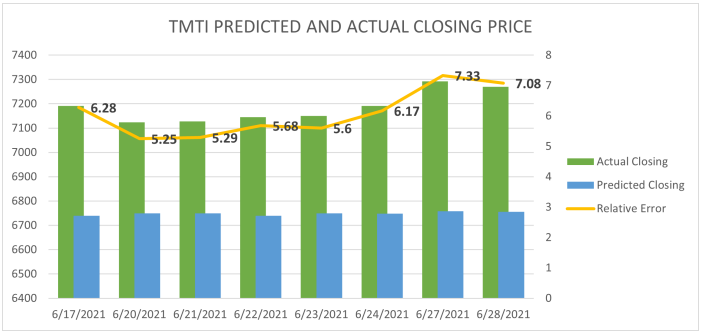
We evaluated the proposed transformer model using four accuracy metrics, MAE, MSE, MAPE, and RMSE. We described the experimental results related to model optimisation



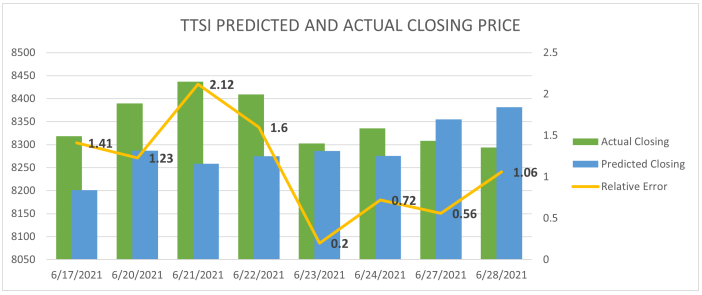
(a) TASI



(b) TBNI



(c) TMTI



(d) TTSI

Fig. 16. Prediction of Unseen Future Stock Closing Price.

and model validation for all the four datasets. Subsequently, we presented results for the prediction of future stock closing prices. We were able to achieve over 90% accuracy compared to the best 72% reported in the literature (see Table I). Furthermore, the experiments showed that the proposed model architectures that split time series into patches were able to identify the dynamics and complex patterns from irregularities in financial time series. Transformer architecture has also been shown to identify sudden changes in stock markets, as reflected in the results. However, the changes occurring may not always

appear regularly or follow the same cycles each time.

ACKNOWLEDGMENT

The experiments reported in this paper were performed on the Aziz supercomputer at King Abdulaziz University.

REFERENCES

[1] T. Yigitcanlar, L. Butler, E. Windle, K. C. Desouza, R. Mehmood, and J. M. Corchado, "Can Building "Artificially Intelligent Cities" Safeguard Humanity from Natural Disasters, Pandemics, and Other Catastrophes? An Urban Scholar's Perspective," *Sensors*, vol. 20, no. 10, p. 2988, may 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/10/2988

[2] T. Yigitcanlar, N. Kankanamge, M. Regona, A. Maldonado, B. Rowan, A. Ryu, K. C. Desouza, J. M. Corchado, R. Mehmood, and R. Y. M. Li, "Artificial Intelligence Technologies and Related Urban Planning and Development Concepts: How Are They Perceived and Utilized in Australia?" *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 4, p. 187, dec 2020. [Online]. Available: https://www.mdpi.com/2199-8531/6/4/187

[3] E. Alomari, I. Katib, A. Albeshri, and R. Mehmood, "COVID-19: Detecting Government Pandemic Measures and Public Concerns from Twitter Arabic Data Using Distributed Machine Learning," *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, p. 282, jan 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/1/282

[4] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning," *Applied Sciences*, vol. 10, no. 4, p. 1398, feb 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/4/1398

[5] E. Alomari, I. Katib, A. Albeshri, T. Yigitcanlar, R. Mehmood, and A. A. Sa, "Iktishaf+: A Big Data Tool with Automatic Labeling for Road Traffic Social Sensing and Event Detection Using Distributed Machine Learning," *Sensors*, vol. 21, no. 9, p. 2993, apr 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/9/2993

[6] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, and A. Albeshri, "A Deep Learning Model to Predict Vehicles Occupancy on Freeways for Traffic Management," *IJCSNS - International Journal of Computer Science and Network Security*, vol. 18, no. 12, pp. 246–254, 2018.

[7] S. Usman, R. Mehmood, and I. Katib, "Big data and hpc convergence for smart infrastructures: A review and proposed architecture," in *Smart Infrastructure and Applications Foundations for Smarter Cities and Societies*. Springer Cham, 2020, pp. 561–586.

[8] R. Mehmood, F. Alam, N. N. Albogami, I. Katib, A. Albeshri, and S. M. Altowaijri, "UTiLearn: A Personalised Ubiquitous Teaching and Learning System for Smart Societies," *IEEE Access*, vol. 5, pp. 2615–2635, 2017.

[9] M. Aqib, R. Mehmood, A. Alzahrani, and I. Katib, *A smart disaster management system for future cities using deep learning, gpus, and in-memory computing*, 2020.

[10] A. Omar Alkhamisi and R. Mehmood, "An Ensemble Machine and Deep Learning Model for Risk Prediction in Aviation Systems," in *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*. Riyadh, Saudi Arabia: Institute of Electrical and Electronics Engineers (IEEE), mar 2020, pp. 54–59. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9044233

[11] H. Alotaibi, F. Alsolami, and R. Mehmood, "DNA Profiling: An Investigation of Six Machine Learning Algorithms for Estimating the Number of Contributors in DNA Mixtures," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, pp. 130–137, 2021.

[12] R. Mehmood, S. See, I. Katib, and I. Chlamtac, *Smart Infrastructure and Applications: foundations for smarter cities and societies*, R. Mehmood, S. See, I. Katib, and I. Chlamtac, Eds. Springer International Publishing, Springer Nature Switzerland AG, 2020.

[13] S. Alotaibi, R. Mehmood, and I. Katib, "Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect," in *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*. IEEE, 2019, pp. 330–335.

[14] Z. Hu, Y. Zhao, and M. Khushi, "A Survey of Forex and Stock Price Prediction Using Deep Learning," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 9, feb 2021. [Online]. Available: https://www.mdpi.com/2571-5577/4/1/9

[15] J. Sirignano and R. Cont, "Universal features of price formation in financial markets: perspectives from deep learning," *Quant. Financ.*, vol. 19, no. 9, pp. 1449–1459, 2019. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/14697688.2019.1622295

[16] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10 389–10 397, 2011. [Online]. Available: https://www.researchgate.net/publication/220219343

[17] L. Takeuchi and Y. Lee, "Applying Deep Learning to Enhance Momentum Trading Strategies in Stocks," Tech. Rep. December 1989, 2013. [Online]. Available: http://cs229.stanford.edu/proj2013/TakeuchiLee-ApplyingDeepLearningToEnhanceMomentumTradingStrategiesInStocks.pdf

[18] M. Nikou, G. Mansourfar, and J. Bagherzadeh, "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms," *Intell. Syst. Accounting, Financ. Manag.*, vol. 26, no. 4, pp. 164–174, 2019. [Online]. Available: https://www.researchgate.net/publication/337735594

[19] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *Int. J. Forecast.*, vol. 37, no. 1, pp. 388–427, 2021.

[20] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 32, jun 2019. [Online]. Available: http://arxiv.org/abs/1907.00235

[21] U. Khandelwal, H. He, P. Qi, and D. Jurafsky, "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, may 2018, pp. 284–294. [Online]. Available: http://arxiv.org/abs/1805.04623 http://aclweb.org/anthology/P18-1027

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem. Neural information processing systems foundation, jun 2017, pp. 5999–6009. [Online]. Available: https://arxiv.org/abs/1706.03762v5

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," oct 2020. [Online]. Available: http://arxiv.org/abs/2010.11929

[24] R. J. Hyndman, "A brief history of forecasting competitions," Tech. Rep. 1, 2020. [Online]. Available: http://monash.edu/business/ebs/research/publications

[25] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1643–1647, 2017.

[26] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, nov 1997. [Online]. Available: http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf

[27] N. Naik and B. R. Mohan, "Study of stock return predictions using recurrent neural networks with LSTM," in *Commun. Comput. Inf. Sci.*, vol. 1000. Springer Verlag, may 2019, pp. 453–459. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-20257-6_39

[28] T. Skehin, M. Crane, and M. Bezbradica, "Day ahead forecasting of FAANG stocks using ARIMA, LSTM networks and wavelets," in *CEUR Workshop Proc.*, vol. 2259, 2018, pp. 186–197.

[29] D. M. Nelson, A. C. Pereira, and R. A. De Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, 2017, pp. 1419–1426. [Online]. Available: https://www.researchgate.net/publication/318329563

[30] S. Y. Shih, F. K. Sun, and H. yi Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, no. 8-9, pp. 1421–1441, sep 2019. [Online]. Available: https://doi.org/10.1007/s10994-019-05815-0

[31] G. Lai, W. C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," Tech. Rep., 2018. [Online]. Available: https://doi.org/10.475/123_4

[32] M. U. Gudelek, S. A. Boluk, and A. M. Ozbayoglu, "A deep learning based stock trading model with 2-D CNN trend detection," in *2017 IEEE Symp. Ser. Comput. Intell.* IEEE, nov 2017, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/8285188/

[33] L. Di Persio and O. Honchar, "Artificial neural networks architectures for stock price prediction: Comparisons and applications," Tech. Rep., 2016.

[34] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for ASR," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., sep 2018, pp. 5874–5878.

[35] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," Tech. Rep., jul 2018. [Online]. Available: http://proceedings.mlr.press/v80/parmar18a.html

[36] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" feb 2021. [Online]. Available: http://arxiv.org/abs/2102.05095

[37] J.-S. Chou, D.-N. Truong, and T.-L. Le, "Interval Forecasting of Financial Time Series by Accelerated Particle Swarm-Optimized Multi-Output Machine Learning System," *IEEE Access*, vol. 8, pp. 14 798–14 808, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8955860/

[38] V. Braverman, R. Ostrovsky, and C. Zaniolo, "Optimal sampling from sliding windows," *J. Comput. Syst. Sci.*, vol. 78, no. 1, pp. 260–272, jan 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2011.04.004 https://linkinghub.elsevier.com/retrieve/pii/S0022000011000493

[39] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert Syst. Appl.*, vol. 140, p. 112896, feb 2020.

[40] S. Smyl and K. Kuber, "Data Preprocessing and Augmentation for Multiple Short Time Series Forecasting with Recurrent Neural Networks," Tech. Rep., 2016. [Online]. Available: https://www.researchgate.net/publication/309385800

[41] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," in *2015 IEEE Int. Conf. Big Data (Big Data)*. IEEE, oct 2015, pp. 2823–2824. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7364089/ http://ieeexplore.ieee.org/document/7364089/

[42] M. Nabipour, P. Nayyeri, H. Jabani, S. S., and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," *IEEE Access*, vol. 8, pp. 150 199–150 212, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9165760/

[43] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," mar 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[44] J. Huang, J. Chai, and S. Cho, "Deep learning in finance and banking: A literature review and classification," *Front. Bus. Res. China*, vol. 14, no. 1, p. 13, dec 2020. [Online]. Available: https://fbr.springeropen.com/articles/10.1186/s11782-020-00082-6

Real Time Multi-Object Tracking based on Faster RCNN and Improved Deep Appearance Metric

Mohan Gowda V[1]
Dept. of Computer Science and Engineering
GITAM School of Technology
GITAM Deemed to be University
Bengaluru, India

Megha P Arakeri[2]
Dept. of Information Science and Engineering
Center of Imaging Technologies
Ramaiah Institute of Technology
Bengaluru, India

***Abstract*—Computer Vision has set a new trend in image resolution, object detection, object tracking, and more by incorporating advanced techniques from Artificial Intelligence (AI). Object detection and tracking have many use cases such as driverless cars, security systems, patient monitoring, and so on. Various methods have been proposed to overcome the challenges such as long-term occlusion, identity switching, and fragmentation in real-time multi-object detection and tracking. However, reducing the number of identity switches and fragmentation remains unclear in multi-object detection and tracking. Hence, in this paper, we proposed a multi-object detection and tracking technique that involves two stages. The first stage helps to detect the multiple objects with high uniqueness using Faster RCNN and the second stage, Improved Sqrt cosine similarity, helps to track the multiple objects by using appearance and motion features. Finally, we evaluated our proposed technique using the Multi-Object Tracking (MOT) benchmark dataset with current state-of-the-art methods. The proposed technique resulted in enhanced accuracy and reduces identity switching and fragmentation.**

***Keywords*—Multi-object detection; tracking; faster RCNN; convolution neural network; data association**

I. Introduction

MOT tracks moving objects with the regular time interval via camera as the input device. In 1998, Zenon Pylyshyn [1] was first developed multi-object tracking. Each detected object is assigned a unique identification number. This identity number retains its association with the object when changing the object's appearance or object moving and draw the motion trajectories of the object based on the unique identities. Fig. 1 shows the basic steps of object detection and tracking. Multi-object detection finds the objects under a unique frame and MOT is integrated with the detected objects in the sequence of frames.

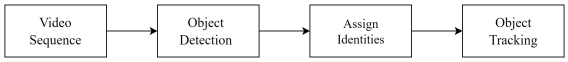


Fig. 1. Basic Steps of Object Detection and Tracking.

A wide range of real-time applications is implemented using MOT, with which it is having extraordinary significance nowadays [2]. Some real-time applications are Human tracking, monitoring Alzheimer's activities, monitoring security, autonomous driving, robotic vision, traffic control, medical images, and others. Gulraiz *et al.* [3] say that tracking is useful because of a specific reason as follows. In the video frame, multiple objects are detected, then establish the identity of the targeted objects in the frame while tracking the objects. Suppose object detection fails, it may be possible to track the object using appearance features and stored location of the previous frame. Local search is initiated instead of global search during tracking. Whenever the movements of the targeted objects are high, the tracking algorithm loses track of the target objects. Hence, the proposed system integrates the detection and tracking methods. The proposed hybrid system has two stages. In the first stage, multi-object detection performs every n^{th} frame of the input. The second stage, multi-object tracking, will take the target objects of n^{th} to $(n+1)^{th}$ frame based on the appearance features and object position in the different frames. Some applications of the proposed work are Alzheimer's patient tracking, monitoring the activities like cooking, hand washing, dressing and others, as shown in Fig. 2.

Fig. 2. Applications of MOT.

In recent years MOT has had many scopes. By the incorporation of AI techniques, the present object detection and tracking techniques are giving better performance than the traditional methods. Most of the traditional methods track the objects from frame to frame. It gives good tracking performance. However, traditional methods are facing computational problems in complex scenarios. They also face difficulty in handling noise and occlusion problems.

Most of the traditional methods are not suitable to adopt in real-time applications. Batch-based movement tracking [4] and probability-based systems [5] must complete batch video processing to track the targeted object. These methods take more time to convert a batch to a performance tracking process. Hence, it is not suitable for real-time scenarios.

Currently, researchers are concentrating on tracking performance by reducing the missing detection rate with the combination of detection and tracking methods. However, we are facing some challenges in tracking moving objects. The

challenges are occlusion, identity switching, time efficiency, motion blurring, viewpoint variations, Background changing, low resolutions, etc. In this work we concentrate mainly occlusion in the long term, time efficiency, identity switching and fragmentation. The objective of this research work is given below.

1) Multiple object detection.
2) Develop a technique to reduce identity switching and fragmentation.
3) Compare the performance of the proposed technique with existing techniques.

The batch and probability-based tracking algorithms have some limitations in time efficiency to apply the real-time application. Much work has been done to overcome the above challenge.

Identity switch problem means detected object changing the identity number in the frame to frame. Identity switching problems mainly occur in two situations: failure of object detection and long-term occlusion occurs. To overcome the failure of object detection, we used the deep learning-based multi-object detection method to detect the person, Bread, Jam, Coffee Maker, Coffee cup and Spoon. To avoid the occlusion problem, we mainly focused on the localization features of the object's appearance. Every object has a different appearance and different locations of the frame. Hence we are implementing the tracking based on the appearance features.

Fragmentation of trajectories may fail when the identity switches have not occurred or object detection fails. With the help of motion and appearance features, a complete trajectory path of one fragmented frame to another fragmented frame can be obtained.

To overcome the above challenges, we proposed the novel multi-object detection and tracking technique. This work has two stages: the first stage detects multiple objects using Faster RCNN, and the second stage is tracking the multiple objects using unique appearance and motion features. The proposed method improves the tracking performance by making robustness in the occlusion.

Paper is organization is as followed: Related methodologies for object detection and object tracking are discussed under Section 2. Section 3 provides the architecture of the proposed system. Section 4 describes the evaluations and results. Finally, Section 5 gives the conclusions of the proposed work.

II. Literature Survey

Recent researchers have concentrated on tracking an individual object in various contexts with multi-object detection and tracking progress. The proposed multiple object tracking method is mainly focused on the association problem. The association problem is used to associate the detected object of one frame to another frame. Hence object detection is carried out before object tracking. The following section primarily focuses on the existing methods and methodologies for object detection and object tracking.

A. Object Detection Algorithms

In the 1990's Anil *et al.* [6] proposed object detection by object matching using deformable templates. These templates have prior knowledge of the object shape like edges and set of edge information. In the late 1990s, object detection was based on the associated geometric appearance feature [7]. The geometric appearance methods involve some geometric properties such as height, width, angle and so on.

Object recognition was moved to the low-level characteristics of the image in the 2000s, based on statistical classifiers. Ojala *et al.* [8] developed rotation invariant classification for grayscale images using the binary patterns locally. Dalal *et al.* [9] proposed Histogram oriented gradients method of object detection in static images. Lowe *et al.* [10] present a method to extract the invariant features of the images. The features are the uniform image scale and rotation, 3D viewpoint and addition of noise. Tuzel *et al.* [11] describe the covariance-based computation method on the internal images. Compared to other statistical methods, a covariance matrix is better to handle large rotations and illumination changes.

Handcrafted conventional features were adopted for object detection in the computer vision sector for many years. In 2012 the deep learning method gave terrific results for the image classification challenge [12]. After successful classification, the researcher concentrated on object detection using deep learning. A Convolution Neural Network (CNN) acts as a backbone network for object detection in deep learning. The CNN is used to extract the local and global feature maps of the input image. Researchers use different backbone networks such as VGG16, AlexaNet, MobileNet, ResNet, and others to achieve the best accuracy.

Now-a-days, research community has moved to region-based networks for object detection. Gulrciz *et al.* [13] proposed a video-based variety of object interaction and spatial relations of the objects. However, to solve a more complex object detection problem, we need to find the object's coordinates in the input image. Girshick *et al.* [14] developed a region Convolution neural network (RCNN) for object detection to overcome the above problem. Here instead of running the classification of many regions. Firstly, they use selective search to extract the region from the image. Then classification will run on the extracted regions. The RCNN has four steps as follows. The selective search algorithm passes on images to generate a region proposal network. Once the region of interest for each image is determined, then resize all proposed regions to match the predefined size of the classes. SVM classifier is used to classify the object and background of the image. Finally, train a linear regression model to generate the bounding boxes of the detected objects. The RCNN has some drawbacks that are as follows. RCNN consumes more time in the training process because the selective search generates the Region of interest. For classification purposes, they used a separate SVM classifier. It is expensive to extract the convolution feature maps for individual regions.

Spatial Pyramid Pooling (SPP-net) method [15] takes any size of the input image. In the SPP-net method, there is no need to compute the convolution feature maps of every region repeatedly. SPP-net generates the entire image features map at a time.

Girshick *et al.* [16] proposed the Fast RCNN to address the above problem. Fast RCNN is similar to RCNN, and Fast RCNN feeds full input images to CNN to generate the feature

map. Then identify the different region proposals from the convolution feature map. The region proposals are different in size. Hence they add these region proposals to the Region of Interest (ROI). Pooling Layer to generate the fixed-size feature maps of the individual regions. The ROI feature vector is further split into two divisions that are classification and regression. Using softmax layer the image is classified into a predicted object and background object and regression is used to generate the bounding boxes. Compared to RCNN, the Fast RCNN has better performance.

Ren *et al.* [17] proposed the Faster RCNN object detection algorithm similar to Fast RCNN. To predict the region proposal, they used a separate network instead of a selective search algorithm. Redom *et al.* [18] proposed the YOLO object detection method. YOLO does not use the region proposal step. It simultaneously detects all bounding boxes of all classes in the input image. Hence YOLO can be an optimized, end-to-end training model. YOLO describes the image into S*S grids. Each grid has probabilities of B (Bounding boxes), C (classes). Yolo can predict the object at 45fps while running the real-time images. However, missed detection raises to identity switch and fragmentation issues.

B. Object Tracking

Some Researchers have investigated spatial features for multiple object tracking [19] and appearance-based approach to capture the association between previous and currently detected frames [20]. Motion-based multiple object target tracking with similar appearance features was proposed in [19]. This method recovers missing data efficiently during the long-term occlusion and also reduces misidentification. However, this method fails to maintain the association between different frames. JuHog *et al.* [21] constructed a Relative Motion Network (RMN) method to track a relative movement between the camera and objects resulting in better data association between different frames and accurate tracking during the camera movement.

Donald *et al.* [22] developed an algorithm for multiple target tracking and Fortmann*et al.* [23] introduced the Joint Probabilistic Data Association (JPDA) algorithm. Rezatofighi *et al.* [4] revisited the JPDA method that has better performance. They utilized some current methods to discover the m-best solution for linear programming. However, there is a delay in decision making. Hence these methods are not suitable for real-time scenarios.

Gabin *et al.* [24] proposed a method to track the multiple football player trajectories from the multiple cameras using a distributed scene algorithm. Improve the performance of online tracking of multiple objects using existing trajectories. Some researchers are using the correlated association for online detection purposes. Yang *et al.* [25] prepared the multi-person online tracking of dynamic appearance features. The temporal dynamic approach incorporates the spatial structure appearance features. This method gives an accurate approach using the data association technique. It also improves the affinity management between the detection and trajectories of the frames. However, it faces the difficulty of online tracking for complex scenes. Xiang *et al.* [26] used the Markow Decision Process (MDP) method to track multiple online objects.

Recently researchers worked on deep learning-based online multi-object tracking of live videos. Alex *et al.* [27] proposed the Simple Online and Real-time Tracking (SORT) method. SORT mainly focused on the association of the objects from one frame to another in online tracking. Firstly, identify the quality of the object detection. Then the Kalman filter is used to estimate the different positions like center, height, aspect ratio and linear velocity of one frame to another. The Kalman filter eliminates the duplicate track of the frame using the Hungarian algorithm. Finally, detect and create the track identities of the object. But, SORT is unable to handle the object re-identification and long-term occlusion problem.

To avoid problems with the SORT method, Nicolai *et al.* [28] integrate the SORT with the in-depth appearance information. In a deep sort, they change the association matrix with the integration of motion and appearance information. Then apply CNN for the target appearance to a large-scale object re-identification dataset. It reduces the loss track of the long-term occlusion. However, this system misses the track during the poster changes in real-time and detects the large frames.

Gulraiz *et al.* [3] proposed multi-person detection and tracking using state-of-art methods. Faster RCNN is used for accurate detection and the deep SORT method is used for tracking. It gives better accuracy in real-time human detection. In tracking, it misses re-identification of objects from one frame to another.

In this paper we propose a system to reduce the number of objects being missed from detection in multiple objects using Faster RCNN. Improved sqrt Cosine similarity method is applied to overcome the object re-identification problem. It uses to find the association matric between two consecutive frames.

III. Proposed Method

The Deep learning methods are the better choice for multiple object tracking. Object detection is the foundation of multiple object tracking. To solve object detection problems with greater accuracy in real-time, deep learning algorithms are preferable for detection. We have solved the multiple object tracking problems using state-of-the-art techniques. The projected method gives critical components of multiple object position prediction in the future frames, and tracking the association of the different frames and managing the lifespan of the tracked multiple objects.

The CNN methods are commonly used in object detection, such as Region-Based Convolutional Neural networks (R-CNN), Fast CNN, Faster R-CNN, etc. The CNN method divides the image into various regions and then classifies every region into different objects, but it needs many regions to predict the object. R-CNN method selectively searches together with the region, but it requires high computation time and prediction is carried using three different models. Fast R-CNN involves a single model that takes features from regions, classifies them, and delivers the border boxes for each class simultaneously. However, it is a slow and time-consuming process to find the Region of Interest. So we conducted a survey to choose the best detection algorithm that is Faster R-CNN. We have divided the proposed work into three models

are shown in Fig. 3. The first one is object detection, the second one is tracking handling, and the third is an association.

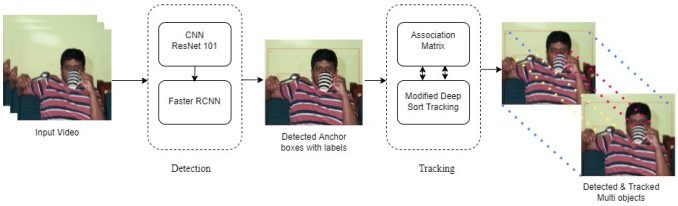


Fig. 3. Block Diagram of Proposed System.

A. Multi-Object Detection using Faster RCNN

The Faster R-CNN involves two stages: Region Proposed Network (RPN) and object detection network. The region proposed network is further divided into three steps. Initially, to extract features map by using a convolutional neural network. After it generates anchor Boxes for using the sliding window approach. Finally, generated anchor boxes are reanalyzed using a tiny network that computes the loss function to select the containing object. The CNN is the backbone of the RPN and object detection network. It requires a step for extracting the convolution feature map.

1) Residual Network-101: The Object detection problem mainly depends upon the feature extraction process. So, we used the ResNet-101 Network model in our method to produce a feature map. ResNet-101 is a convolutional neural network that contains 101 layers between the residential connections. The main advantage of the ResNet-101 is to train the module efficiently without increasing the training error. It also helps to solve the vanishing gradient problem by adding a shortcut connection technique. The shortcut connection is skipped one or more layers to perform the identity mapping. The shortcut connection takes the input x to the output after a few weighted layers. This x is added to the output of the sketch layer. The pictorial representation of the residential network shows in Fig. 4. In different scenarios, the two types of shortcut

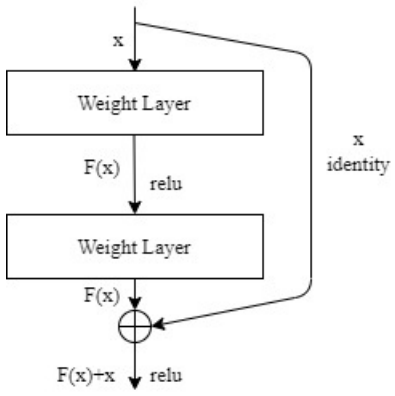


Fig. 4. Building Block of Residual Learning.

connections are used in ResNet. When the input and output are the same dimensions then shortcut(x) is used directly.

$$z = F(x, w_i) + x \tag{1}$$

When the input and output are different dimensions. Then identity mapping is preferred by padding extra 0 to make dimensions suitable.

$$z = F(x, w_i) + w_j x \tag{2}$$

In equations 1 and 2, where x is the feature map value of the previous layer, F is the convolution function and w is the weighted matrix.

2) Anchor Box Generation: The RPN takes the convolution feature map that the ResNet-101 generates as an input, and the output is anchor boxes generated by RPN using the Sliding window approach. This Sliding window approach adopts a 3 * 3 window size upon the future map. Sliding window traverse across the future map to generate the anchors. The sliding window generates a set of 9 anchors for each pixel, each pixel center point is (x,y). All 9 anchors are three different vertical scales such as 128 * 128, 256 * 256 and 512 * 512 and three different aspect ratios of 1:1, 1:2 and 2:1 as shown in Fig. 5. Then, determine the number of anchor boxes

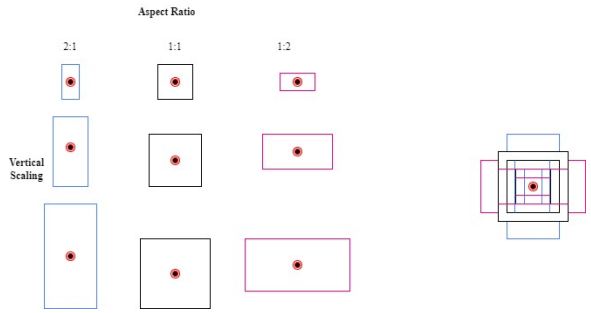


Fig. 5. Nine Anchor are Generted in Each Pixel.

that is overloaded with Background Classes (Bc) using the intersection of union approaches. We set the threshold value to the intersection of the union approach. If the threshold value is greater than 0.8 then consider the object is present in the region. Suppose the threshold value is less than 0.2 then no object is present in the region.

$$IOU = \frac{anchor \cap Bc > 0.8 = ObjectPresent}{anchor \cap Bc > 0.2 = NoObjectPresent} \tag{3}$$

3) Loss Function: The loss function is used to fine-tuning the selected anchor boxes. The loss function mainly contains two tasks: regression and classification. We use binary classification to predict whether the concerned anchor boxes contain the object or background. The regression determines the position of the predicted anchor box. The loss function calculates for both the classification and regression to fine-tuning the anchor box. The loss function is shown in equation 4.

$$L(p_i, t_i) = \frac{1}{N_{cls}(\sum_i L_{cls}(p_i, p_i^*))} + \frac{\lambda}{N_{reg}(\sum_i p_i^* * L_{reg}(t_i, t_i^*))} \tag{4}$$

In equation 4, Where p_i means the predicted probability of anchors containing objects, p_i^* means the ground-truth value of anchors contains an object, t_i coordinates of predicted anchors, t_i^* is ground truth coordinates associated with anchors,

L_{cls} classification loss, N_{reg} is normalization parameter of regression, L_{reg} regression loss and λ is a constant value.

4) Region of Interest (ROI): The output generated from the RPN is the input of the ROI Pool layer. The output of RPN anchor boxes are different sizes, so the task of the ROI is to reduce the different size anchor boxes to the same or fixed size anchor boxes. For this purpose, we use classification and regression methods. The classification method identifies the object or background class of the image. Then regression gives the bounding box values (dx,dy,dh,dw) to cover the complete object. Where (dx,dy) is the center point of the bounding box, dh is the height of the bounding box and dw is the bounding box width. The performance and accuracy of the Faster R-CNN is good enough than all of the available traditional object detection algorithms. The framework of the Faster R-CNN is shown in Fig. 6. We have trained Faster R-CNN and

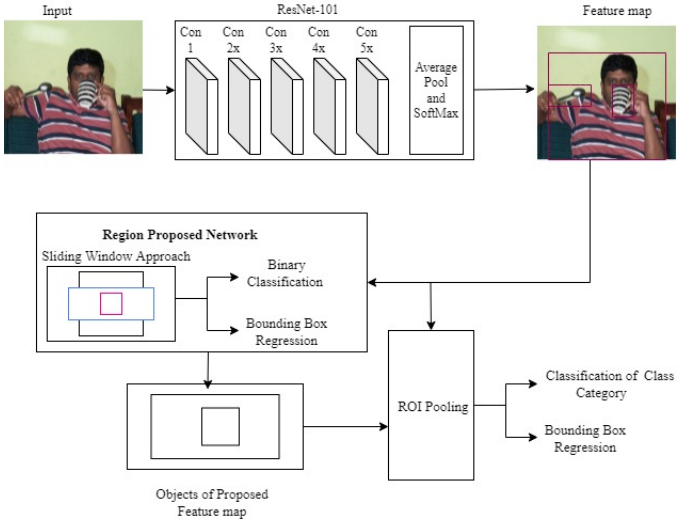


Fig. 6. Faster RCNN Complete Framework.

5000 annotated images like a person, coffee cup, coffee maker, bread, jam and spoon. The Faster R-CNN gives the improved detection accuracy effectively.

B. Multiple Object Track Handling of Feature Frames

We proposed a multi-object tracking method using the modified deep sort technique. The modified method takes the input as detected bounding boxes from Faster R-CNN. We are using the Kalman filter to extract the spatial and tracking information of the bounding boxes. The deep CNN model helps to extract the appearance feature of the frame. To find a track association between current and next frame of appearance feature using Mahabolish distance and improved sqrt-cosine similarity measures. If the track association's threshold value is equal to 1, the track is confirmed and updated; otherwise, delete the frame immediately. The modified deep sort is shown in Fig. 7. The frame-by-frame data association method and the Kalman filter are the essential components of the modified deep sort. The trackers scenario is based on the multidimensional state space(u,v,Y,h,a,b,c,d) that contains the center of bounding box is (u,v), expectation ratio is Y, height

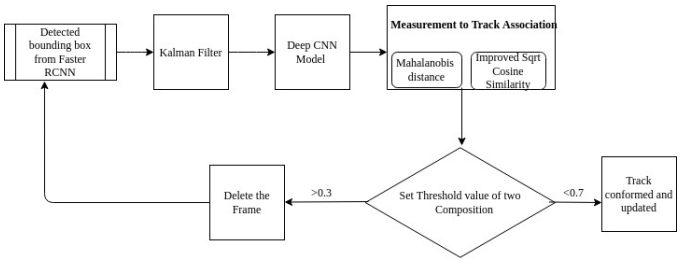


Fig. 7. Modified Deep Sort.

is h and the respective velocity coordinator are (a,b,c,d). The entire count of each track frame since the last successful measurement association of Tn is given by n for each track. The Kalman channel predictions occurred, the counter is incremented, or when the track has been assigned with the previous list, the counter is reset to 0. Suppose a new track prediction is started for each detection that cannot be assigned to any of the current lists. The first three frames of the new track are classified as tentative. These frames are kept for further processing when the association measurement is found at every timestamp, otherwise deleted.

1) Association of the Current Frame and Predicted Frame: The Hungarian algorithm solves the measurement to track the association between the predicted Kalman state and the next frame. By creating two relevant measures, we were able to combine motion and appearance data. We utilized Mahalanobis Distance between the predicted frame and the next frame to get motion information.

$$d^{(1)}(p,q) = (d_p - y_p)^T S_i^{-1}(d_q - y_p) \tag{5}$$

Equation 5 denotes the projection of the pth track distribution into measurement space (yp, Sp), and the detection of the qth bounding box by dq. The Mahalanobis Distance removes the state estimation uncertainty between the protected and newly arrived state mean track location. Further, it is possible to provide false Association by the threshold value at 95% confidence interval computed from the inverse distribution we denote

$$b^{(1)}_{(p,q)} = 1[d^{(1)}(p,q) \leq t^{(1)}] \tag{6}$$

Mahalanobis Distance is suitable only when motion uncertainty is less for the association metric. Our image space Kalman filtering framework provides the approximate value of the predicted object location. The Mahalanobis Distance is an uninformed metric for tracking in occlusion when the rapid displacement of the image plane. Therefore we integrated the second metric for tracking each bounding box. With the help of a protected CNN to extract the appearance feature of the bounding boxes, the architecture of the appearance feature network is shown in Fig. 8. To detect the space of the pth and qth track proposes most of the authors used cosine similarity. The cosine similarity is derived from the Euclidean distance. However, the Euclidean distance is not good for dealing with the probability-based approach. Zhu *et al.* [29] proposed the Sqrt cosine similarity. It is

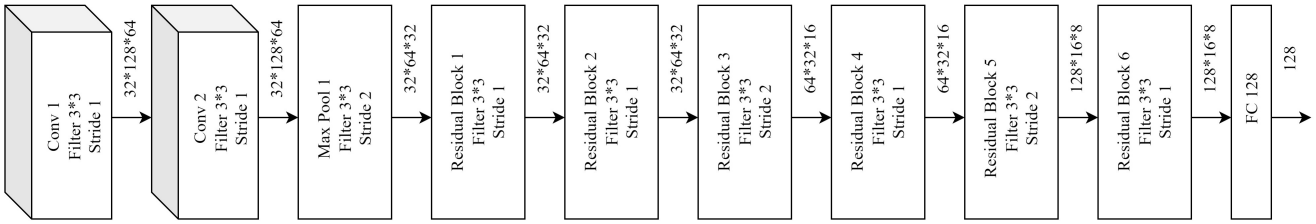


Fig. 8. Architecture of Appearance Feature Extractor Network.

derived from Hillinger distance, but there is conflict to define the similarity measures in some cases. Hence we adopted improved sqrt cosine similarity [30] to find the pth and qth track detection in appearance space.

$$d^{(2)}(p,q) = \frac{\sum_{i=1}^{T} \sqrt{p_i q_i}}{\sqrt{\sum_{i=1}^{T} p_i} \sqrt{\sum_{i=1}^{T} q_i}} \tag{7}$$

The binary variable is used to indicate if the association is good according to this metric.

$$b_{(p,q)}^{(2)} = 1[d^{(2)(p,q)} \leq t^{(2)}] \tag{8}$$

By addressing different aspects of the association method in a combination of both matrices. Mahalanobis Distance gives the information of the object position based on the motion for short-term prediction. The improved sqrt cosine distance is used to recover identities after long-term occlusion. Based on the Association problem, we combine what the metric is using a weighted sum.

$$C_{(p,q)} = \lambda d^{(1)}(p,q) + (1-\lambda) d^{(2)}(p,q) \tag{9}$$

Where we call the association appearance if it is within the range of both metrics.

$$b(p,q) = \prod_{i=1}^{2} b_{(p,q)}^{(i)} \tag{10}$$

In the equation 10, value is 1 it indicates both the metrics are equal, if zero metrics are not equal. It also indicates (p,q) is a true match between appearance and spatial Information. video sequence the next new frame detection is effectively associated with the present track OK then the track is continued as long as It is successfully associated and that track is confirmed and tracking update else deleted immediately.

IV. Evaluations and Results

The proposed system used a self-generated dataset to perform the multi-object detection and tracking. We use Google collaboratory for evaluation and python programming language and Detectron 2 for the experimental setup.

A. Dataset

To train the detection algorithm, we used a self-generated dataset containing 5000 pictures of people, jam, bread, spoons, coffee cups, and coffee makers. The dataset images are collected from different weather conditions, different viewpoints, different lighting of day and nights, blurring of images, crowded places and malls. Some images are collected from surveillance cameras and the internet. After collecting the images, we annotated images in different classes using the Labelimg tool. We classify images into 6 different classes: person, jam, bread, spoon, coffee cup, and coffee maker shown in Table I. The LabelImg tool after annotating it generates the XML files after we convert them into JSON format.

TABLE I. Proposed Dataset Classes

Classes	Number of Instances	Number of Images
Person	2100	1000
Bread	878	700
Jam	900	750
Spoon	1212	800
Coffee Maker	1278	900
Coffee Cup	1211	850

B. MOT Benchmark Dataset

We have evaluated our proposed system in the MOT benchmark dataset [35]. This dataset contains a combination of 21 different datasets. The dataset has contained 645 second video and it is the combination of 21 different sequences with proper annotation.

C. Results

The proposed system has two stages: stage 1, Multi-object detection and stage 2, Multi-object tracking. Multi-object detection purpose we compared different Deep learning based object detection methods based on the accuracy and loss. After evaluating, we discovered that ResNet-101 based Faster RCNN is giving better performance. The evaluating different detection algorithm results are shown in Fig. 9. In Fig. 9(a) shows the multiple object detection output. (b) Gives an accuracy comparison of the Faster RCNN, Mask RCNN and Retinanet detection methods. Compared to all Faster RCNN gives better accuracy. (c) Shows the False negative results of Faster RCNN, Mask RCNN and Retinanet detection methods. Mask RCNN gives less False negative but Faster RCNN also gives better Results. (d) Shows the regression time box loss of Retinanet and Faster RCNN, (e) Gives the class name loss and (f) Gives a comparison of the total loss of Retinanet and Faster RCNN. Faster RCNN gives less detection loss compared to Retinanet and Faster RCNN. Compared to different detection methods, Faster RCNN gives good Accuracy and less detection

TABLE II. EVALUATED TRACKING RESULTS OF MOT BEHCHMARK DATASET

		MOTA	MOTP	MT	ML	ID	FM	FP	FN	RunTime
Batch	KBNT [31]	68.2	79.4	41.00%	19.00%	933	1093	11479	45605	0.7Hz
	LMPp [2]	71	80.2	46.90%	21.90%	434	587	7880	44564	0.5Hz
	MCMOT HDM [32]	62.4	78.3	31.50%	24.20%	1394	1318	9855	57257	35Hz
	NOMTwSDP16 [33]	62.2	79.6	32.50%	31.10%	406	642	5119	63352	3Hz
Online	EAMITT [34]	52.5	78.8	19.00%	34.90%	910	1321	4407	81223	12Hz
	POI [31]	66.1	72.5	34.00%	20.80%	805	3093	5065	55914	10Hz
	SORT[27]	59.8	79.6	25.40%	22.70%	1423	1835	8698	63245	60Hz
	DEEP SORT[28]	61.4	79.1	32.80%	18.20%	781	2008	12853	56668	40Hz
	Proposed System	71.2	80.1	33.40%	17.90%	825	1225	4115	54724	41Hz

loss. Hence we used Faster RCNN for detection purposes. It has variable performance on different classes of Self-Generated Dataset.

We proposed a Multi-object tracking method based

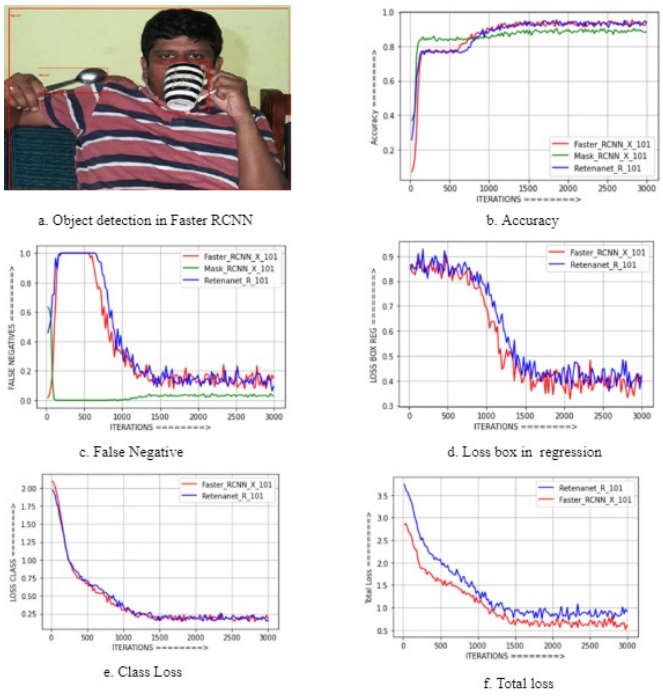


Fig. 9. Object Detection Output and Comparision Graphs of Different Deep Learning Detection Methods.

on the appearance and motion features. Firstly we trained the dataset using the Faster RCNN and then we run modified deep sort on the same evaluation dataset. Fig. 10 shows the tracking of the different frames with identity numbers.

In Table II shows the results of our proposed system's assessment on the MOT dataset. We used an object detection threshold of 0.8and further fine-tuned it with the different parameters to produce an efficient model. The following parameters are used for evaluation.

Multi-Object Tracking Accuracy (MOTA) - Summarizes the tracking accuracy terms of identity switches, false negative and false positive.

Fig. 10. Tracking with Unique Identity Number.

Multi-Object Tracking Precision (MOTP) - Provides the tracking precision for bounding boxes average dissimilarity between ground-truth value and predicted location.

Mostly Tracked (MT) - The percentage of ground-truth tracks has had the same label for at least 80% of their life span.

Mostly Lost (ML) - The proportion of ground-truth tracks that are tracked for at minimum 20% of their lifespan.

Identity Switches (ID) - It gives the number of times identity number changes the ground truth track.

Fragmentation (FM)- Identifies the number of times tracks have been interrupted due to missed detection.

Table II shows the results of our proposed work. It increases accuracy from 61.4 to 71.2 and also reduces the fragmentation problem. At the same time, identity switches increase slightly due to occlusion. We have seen a significant increase in the mostly tracked object. Hence our proposed model is suitable for online tracking.

Existing techniques were not efficient in reducing identity switching, fragmentation, and accuracy. However, using the modified deep sort technique it is possible to reduce identity switching and fragmentation reasonably better. Proposed research work contribute in reducing false identification due to

the presence of frequent identity switching and fragmentation. It helps to track the multi objects in real time environment with better accuracy.

V. Conclusion

In this paper, we presented an improved multi-object detection and tracking technique that will reduce identity switches and fragmentation in the presence of occlusions. The proposed technique employed the Faster RCNN method to detect multiple objects and to overcome occlusion challenges. We also presented a modified deep sorting technique for multi-object tracking to reduce identity switch and fragmentation issues. The technique is simulated on a series of experiments using real and synthetic datasets has yielded better accuracy, reduced identity switching, and fragmentation compared to existing techniques. The specified objectives are achieved in this work. Improvement of performance in dark environmental conditions is yet to be investigated.

References

[1] R. W. Storm and Z. W. Pylyshyn, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism," *Spatial Vision*, vol. 3, no. 3, pp. 179–197, 1988.

[2] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele, "A Multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects," 2016. [Online]. Available: http://arxiv.org/abs/1607.06317

[3] G. Khan, Z. Tariq, M. U. G. Khan, P. Mazzeo, S. Ramakrishnan, and P. Spagnolo, "Multi-person tracking based on faster r-cnn and deep appearance features," 2019.

[4] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 3047–3055, 2015.

[5] A. Milan, K. Schindler, and S. Roth, "Detection- and trajectory-level exclusion in multiple object tracking," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3682–3689, 2013.

[6] A. K. Jain, Z. Yu, and S. Lakshmanan, "Object matching using deformable templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 267–278, 1996.

[7] J. L. Mundy, "Object Recognition in the Geometric Era: A Retrospective," pp. 3–28, 2006.

[8] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," vol. 1, pp. 886–893, 2005.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3952 LNCS, pp. 589–600, 2006.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[13] G. Khan, M. U. Ghani, A. Siddiqi, Z. ur Rehman, S. Seo, S. W. Baik, and I. Mehmood, "Egocentric visual scene description based on human-object interaction and deep spatial relations among objects," *Multimedia Tools and Applications*, vol. 79, no. 23-24, pp. 15 859–15 880, 2020.

[14] R. Girshick, J. Donahue, T. Darrell, J. Malik, U. C. Berkeley, and J. Malik, "1043.0690," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 5000, 2014. [Online]. Available: http://arxiv.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[16] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1440–1448, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," pp. 779–788, 2016.

[19] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2304–2311, 2013.

[20] A. Bewley, L. Ott, F. Ramos, and B. Upcroft, "Alextrac: Affinity learning by exploring temporal reinforcement within association chains," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 2212–2218, 2016.

[21] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, "Bayesian multi-object tracking using motion context from multiple objects," pp. 33–40, 2015.

[22] D. Reid, "An algorithm for tracking multiple targets," *IEEE transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[23] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association," *IEEE Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, 1983.

[24] G. Kayumbi, P. L. Mazzeo, P. Spagnolo, M. Taj, and A. Cavallaro, "Distributed visual sensing for virtual top-view trajectory generation in football videos," pp. 535–542, 2008.

[25] M. Yang and Y. Jia, "Temporal dynamic appearance modeling for online multi-person tracking," *Computer Vision and Image Understanding*, vol. 153, pp. 16–28, 2016.

[26] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 4705–4713, 2015.

[27] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2016-Augus, pp. 3464–3468, 2016.

[28] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2017-Septe, pp. 3645–3649, 2018.

[29] S. Zhu, L. Liu, and Y. Wang, "Information retrieval using Hellinger distance and sqrt-cos similarity," *ICCSE 2012 - Proceedings of 2012 7th International Conference on Computer Science and Education*, no. Iccse, pp. 925–929, 2012.

[30] S. Sohangir and D. Wang, "Improved sqrt-cosine similarity measurement," *Journal of Big Data*, vol. 4, no. 1, 2017.

[31] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "POI: Multiple object tracking with high performance detection and appearance feature," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, pp. 36–42, 2016.

[32] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, no. Mcmc, pp. 68–83, 2016.

[33] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 3029–3037, 2015.

[34] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, pp. 84–99, 2016.

[35] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking," pp. 1–15, 2015. [Online]. Available: http://arxiv.org/abs/1504.01942

OBEInsights: Visual Analytics Design for Predictive OBE Knowledge Generation

Leona Donna Lumius, Mohammad Fadhli Asli*
Faculty of Computing and Informatics
Universiti Malaysia Sabah, Malaysia

***Abstract*—Gaining traction in modern higher education, outcome-based education (OBE) focuses on strategizing pedagogical approaches to help the student achieve specified learning outcomes. In the context of Malaysia, OBE is oriented towards holistic development of graduates to ensure readiness towards the working sector. To empower OBE implementation, standardized measuring instrument iCGPA was introduced to higher education institutions nationwide. With lower dependency on provided curriculum, graduate abilities and values development are also attainable via extracurricular activities. However, analyzing the curriculum results in hand with extracurricular activities can be a daunting task, albeit the potential enriched performance assessment. In addition, the current iCGPA instrument employs radar map that restricts data exploration despite its capability in visualizing multivariate information. This study aims to enable predictive knowledge generation on understanding the relationship between learning activities and performance in OBE. Therefore, a predictive visual analytics system namely OBEInsights is proposed to facilitate education analysts in assessing OBE. The system development started with the identification of crucial design and analytic requirements via a domain expert case study. The system is then built with deliberate considerations of guiding factors of a design framework conceptualized from the case study. Subsequently, the system was then evaluated in usability testing with 10 domain experts that consist of usability rating and expert validation. The evaluation and expert validation results demonstrated the effectiveness and usability of OBEInsights in facilitating OBE predictive assessment. Several design insights on constructing visual analytics for OBE assessment were also discovered in terms of effective visualization, predictive modeling, and knowledge generation. Analytic designers and builders can leverage the reported design insights to enhance knowledge generation tools for OBE assessment.**

Keywords*—Visual analytics; visualization; learning analytics; outcome-based education (OBE)

I. Introduction

Outcome-based education (OBE) has become a prominent higher education strategy and pedagogical approach around the globe. OBE focuses on organizing teaching and assessment that help students achieve specified outcomes or goals [1]. Many countries have adopted the OBE approach in their higher education structures and initiatives along with additional unique goals. In Malaysia, OBE in higher education is oriented towards the development of holistic graduates and readiness towards the working sector. To achieve this goal, an initiative and instrument namely Integrated Cumulative Grade Point (iCGPA) were introduced to Malaysian higher education institutions [2]. iCGPA helps the institutions in determining the graduate's achievement based on the program learning outcomes (PLO) set by the faculty.

The instrument also eases the recording of graduates' ability and values attainment throughout the study program duration. The recorded data is then visualized in the form of a radar map, indicating the graduate's improvement in terms of abilities and acquisition of values. However, graduate abilities and values attainments are also attainable via extracurricular activities with less dependency to the provided curriculum. Analyzing the curriculum results with extracurricular activities could provide enriched understandings towards assessing the student performance.

Despite these potentials, merging curriculum results with extracurricular activities in an analysis can cause information overload. In addition, the current radar map representation of iCGPA restricts data exploration despite its capability in visualizing multivariate information. Visual analytics is a data exploration method supported by interactive visualization [3], allowing the analyst to pursue new inquiry throughout the exploration [4]. The main motivation of this study is to facilitate education analysts in performing predictive analysis on OBE learning activities and performance. Prior studies on educational visual analytics were observed to primarily focus on visualization tool creation and system features [5], [6]. Furthermore, this study found limited visualization work that discusses OBE-specific domain users, analysis tasks, and visualization design.

This paper reports our empirical investigation on the visualization design for supporting OBE knowledge generation with regard to design requirements, development, and evaluation. To address the gap, this study proposes a predictive visual analytic namely OBEInsights that enables education analysts and practitioners to perform predictive OBE assessment. This study firstly explores the design requirements and analytic practices by interviewing 5 domain experts in a domain characterization case study. A design framework is conceptualized based on the identified requirements and analysis tasks from the case study. Next, the visual analytic OBEInsights was designed and developed with design consideration guidance of the framework. Subsequently, this study evaluated and demonstrated the effectiveness of OBEInsights in usability testing with 10 domain experts. The major contributions of this paper are as follows:

1) A novel visual analytics design framework for supporting OBE knowledge generation.
2) A visualization system namely OBEInsights for facilitating OBE predictive learning analysis.
3) Empirical evaluation report to demonstrate the effectiveness of OBEInsights in supporting knowledge generation.

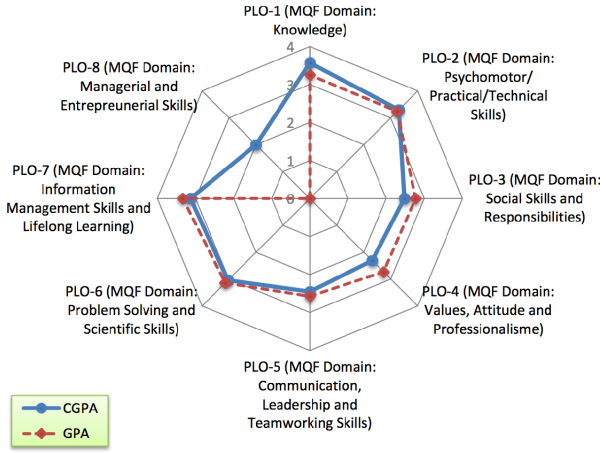


Fig. 1. iCGPA's Radar Map of Student's Achievement based on Specified LOD.

The remaining content of this paper is organized as follows. Section 2 introduces the fundamentals of OBE predictive analysis and visual analytics. Next, Section 3 describes the methodology in deriving the design framework, visualization system development, and design evaluation. Section 4 then presents and discusses the results of empirical design evaluation in a user case study. Finally, the conclusions and future work are presented in Section 5.

II. Background

This section introduces the background of this study by summarizing the underlying principles and fundamentals of outcome-based education (OBE), predictive analytics, and visual learning analytics. The adoption of OBE in modern higher education calls for efforts in developing new analytics to support OBE-oriented analysis.

A. Outcome-Based Education and Predictive Analytics

Outcome-based education (OBE) is an educational approach or pedagogical perspective that emphasizes curriculum choices based on student performance [7]. Apart from the core approach, OBE implementations in many countries can be different with additional goals or structures. OBE implementation in Malaysia's higher education is oriented towards producing graduates equipped with relevant expertise, morals, and social skills.

Malaysia practices top-down approach for organizational management and decision-making process [8], [9]. To standardize OBE in Malaysian higher education institutions, the Malaysian Qualifications Agency (MQA) has set 8 specific Learning Outcome Domains (LOD). In 2015, the Malaysian Ministry of Higher Education introduces the Integrated Cumulative Grade Point Average (iCGPA), an evaluation instrument for Malaysian OBE [2]. iCGPA records and reports on the students' holistic attainment throughout the study duration as per specifications of LOD set by the MQA. Currently, iCGPA report of students' holistic achievement throughout each semester is visualized as a radar map as shown in Fig. 1.

Education analysts can analyze the OBE datasets like iCGPA by using predictive modeling to identify future patterns and trends [10]. Parameters like the students' knowledge levels, performance, scores, or marks are commonly analyzed and assessed in higher education [11]. Predictive models are applied against the dataset to automatically learn and generate predictions based on recorded historical data. However, there are limited advances in OBE evaluation platforms and tools, particularly in integrating student learning performance with extracurricular activities. In addition, there is limited work that clearly describes the OBE analytic practices and users specifically in Malaysia.

B. Visual Learning Analytics

Visual analytics combines the prowess of automated analysis with interactive visualization, allowing the user to gain efficient understanding, reasoning, and decision-making on large and complex datasets [12]. Applied into the education domain, visual learning analytics is defined as the use of computational tools and methods for understanding educational phenomena via interactive visualization [6]. Educators leverage visual analytics to understand or measure students' progress for diagnostic pedagogical decision-making in real-time [13].

Literature indicates many advances of visual analytics approaches in the educational context especially in enhancing learning analytics and decision-support tools [6]. In addition, prior work also explores the educational analytic practices via design study to characterize analysis scenarios, target users, and viable visualization. Several design studies have investigated visualization design for different focus like online classroom [14], [15] and massive open online courses (MOOC) [16], [17], [18]. Prior work also reported that the visual form of graphs, maps, and dashboards have greater potential to generate effective visualization especially for performance assessment [19].

Despite many advances in visual learning analytics, there are limited discussions for facilitating knowledge generation of specific learning pedagogy like OBE. Therefore, this study was carried out to investigate and offer insights in constructing predictive visual analytics for OBE assessment. This study then designed and developed OBEInsights, a visual analytics system for supporting OBE assessment as described in the following section.

III. Visual Analytics System: OBEInsights

This section presents OBEInsights, a visual analytics system to support education analysts in performing OBE predictive analysis. Detailed descriptions of requirement analysis leading towards the conceptualization of its design framework are also presented. Furthermore, the specifications and features of OBEInsights including data sources, automated data analysis, and visualization are explained.

A. Domain Characterization and Design Framework

To identify the visualization design requirements for facilitating OBE analysis, this study adopted part of the design study methodology [20], specifically the problem characterization. Related to requirements analysis in software engineering [21], problem characterization serves as a discovery stage in understanding domain-specific analysis tasks. Understanding the domain analysis helps the designer to translate and fit it

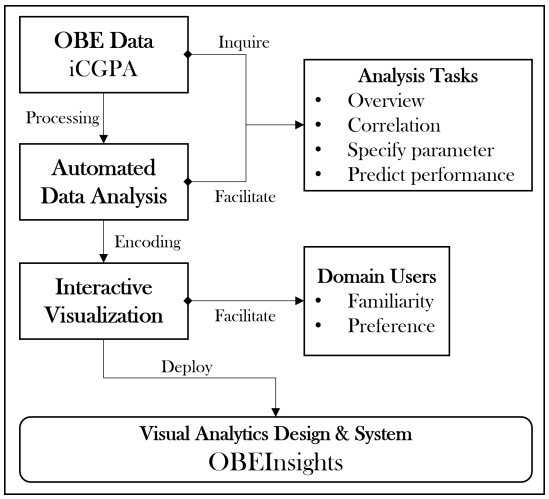


Fig. 2. Visual Analytics Design Framework for Supporting OBE Predictive Analysis.

TABLE I. BACKGROUND OF DOMAIN EXPERTS

Expert	Qualification	Teach Experience	Expertise Area
Exp1	PhD	30 years	Natural Science
Exp2	PhD	20 years	Tourism, Entrepreneurship
Exp3	PhD	30 years	Higher Education
Exp4	PhD	35 years	Engineering
Exp5	PhD	20 years	Accounting

into generic visualization language and create actionable visual metaphors. 5 domain experts from the higher education field with different expertise but have considerable proficiency and experience in OBE were interviewed as described in Table I. A semi-structured interview was carried out with the aim to learn and discover OBE-specific analysis tasks and nature from the experts. The interview helps the study to understand these major requirements: domain analytic practices, domain user's visualization knowledge, and domain user's analysis interests.

Based on the identified requirements, a visual analytics design framework was conceptualized for facilitating OBE predictive assessment as shown in Fig. 2. The purpose of this framework is to guide visualization designers in translating domain analysis tasks and considering appropriate visualization to facilitate OBE analysis. Referring to the visual analytics model [22], this study mapped the visual analytics components for the context of predictive OBE assessment. The interview results reveal interesting insights on OBE analytic practices in Malaysia, specifically in pedagogical decision-making and analysis interests. The experts explained the major factors that need to be considered in determining the faculty's program learning outcomes: curriculum revisit, external feedback, and standard compliance. Based on the interview, 3 major analysis tasks in OBE predictive analysis were identified:

1) Overview on student curriculum results and activities.
2) Correlation between student performance and extracurricular activities.
3) Prediction on student progression and achievement based on activities.

In terms of domain users, the experts stated that most education analysts and practitioners possess limited knowledge and familiarity with visualization. Therefore, the factors of familiarity and preference need to be considered in generating appropriate visualization. The presented framework is a refined iteration of prior work [23], added with additional focus on design development. Driven by the identified requirements and factors, a visual analytics system was designed and developed for facilitating OBE predictive analysis. The following subsection presents detailed descriptions of the visual analytics system including specifications and development process.

TABLE II. INTEGRATED RELEVANT ATTRIBUTES IN ICGPA DATASET

Attribute	Remarks
`student_id`	Identifier for student information
`sem`	Identifier for semester sessions
`igpa`	Grade point average
`icgpa`	Cumulative grade point average
`po_value`	Program learning outcomes (PLO)
`po_this_sem_course`	PLO per semester
`po_igpa`	Student PLO achievement per semester
`po_courses`	No. of courses contribute to PLO
`po_grade_pointer`	Grade point obtained per PLO
`po_achievement`	Achievement distribution

B. Visual Analytics Design and Development

Guided by the conceptualized framework, this study designed and built OBEInsights, a predictive visual analytic system for supporting OBE assessment. The system was built with specifications and considerations as follows.

Data Sources: In this study, two different OBE datasets generated by iCGPA consisting of the 'Accounting' program of our institution that contains students' PLO information were used. These datasets were selected after thorough consideration of easier access and its uniformity with many OBE datasets. The dataset consists of students' PLO results and extracurricular activities in each semester. The datasets were prepared for automated analysis by integrating key relevant attributes as shown in Table II.

Automated Data Analysis: Based on identified analysis requirements, this study designed the automated data analysis particularly for facilitating student achievement prediction. Literature recommends the following algorithms for modeling prediction due to model flexibility and customizable parameters: random forest, gradient boosting, neural network, and support vector machines [24], [25]. An experiment was conducted upon these algorithms in terms of lower and upper intervals to determine the most precise predictive modeling for OBE assessment. Before the experiment, the variable correlation within the dataset was firstly identified by using best subset classification. The classification result shows that `igpa`, `icpga`, `po_igpa`, `po_value`, and `sem` to be the impacting variables on the prediction modeling. Furthermore, the classification result indicates linear dependency towards `po_grade_pointer` that could lead to poor accuracy measures. Next, the prediction accuracy of the trained models was measured in an experiment with additional validation by using misclassification error. The results of the prediction models' performance are shown in Table III.

The experiment result shows that the random forest algorithm to be the best model for predicting student achievement.

TABLE III. LOWER AND UPPER INTERVALS OF PREDICTION ACCURACY

Algorithms	Accuracy (Low)	Accuracy (High)
Random Forest	91.0%	94.9%
Gradient Boosting	88.2%	93.0%
Neural Network	45.2%	53.2%
Support Vector Machines	46.4%	54.4%

This study infers that the lower performance by neural network and support vector machine is due to sharpened model flexibility that overfits prediction. Moreover, random forest and gradient boosting use sequential model training, simulating and refining decision tree based on prior accuracy. Therefore, the random forest prediction technique was incorporated into our automated data analysis design for OBE prediction assessment.

Interactive Visualization: With the constructed automated data analysis component, this study needs to assess the appropriate visualizations and interfaces to display it. Upon deliberate consideration of the critical factors of user familiarity and preference, simple visual metaphors and layouts are selected. The visual analytics is then deployed in a form of simple dashboard front-end integrated with predictive modeling back-end. The front-end was developed using HTML, CSS, and Javascript, while the back-end was developed using Phyton. OBEInsights incorporates the existing radar map iCGPA to ensure continuous familiarity towards further OBE prediction analysis. The interfaces of OBEInsights are as shown in Fig. 3.

The user can select students to observe from the student list on the left sidebar, and the main window updates the visualization as per students' information. The main window can be scrolled down to show more analysis items like student activities in Fig. 4 and the prediction panel in Fig. 5. The user can access 'Predict & Analyze' tab to open the prediction assessment panel for specific program outcomes (PO). User can customize the parameters like student achievements or extracurricular activities, and the predictive model generates results. The prediction result for the selected PO then can be interpreted based on achievement distribution categories.

IV. DESIGN EVALUATION WITH DOMAIN EXPERTS

This section presents and discusses the design evaluation of our OBEInsights visual analytics system. This study demonstrates and evaluates the effectiveness of OBEInsights

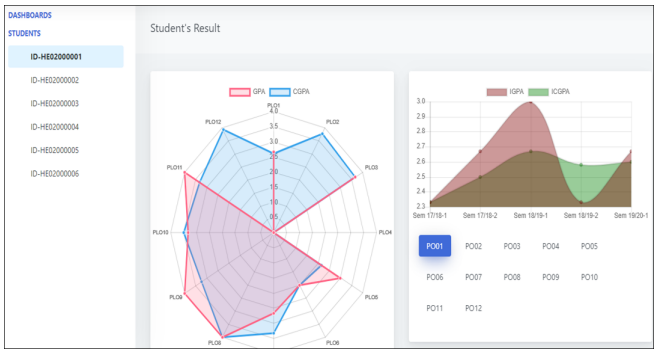


Fig. 3. Overview Layout of OBEInsights.

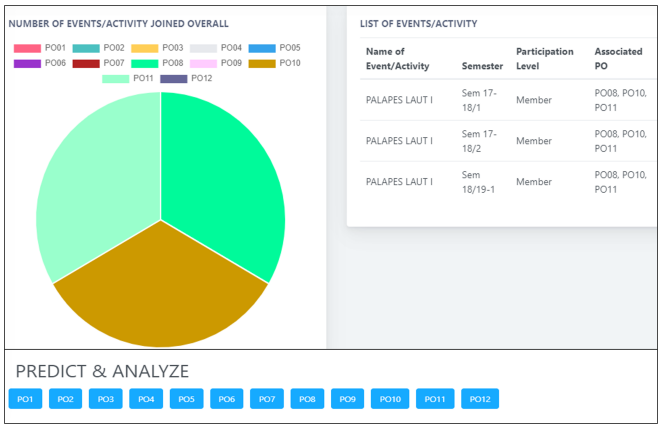


Fig. 4. Student Activities and Achievements from Extracurricular Activities.

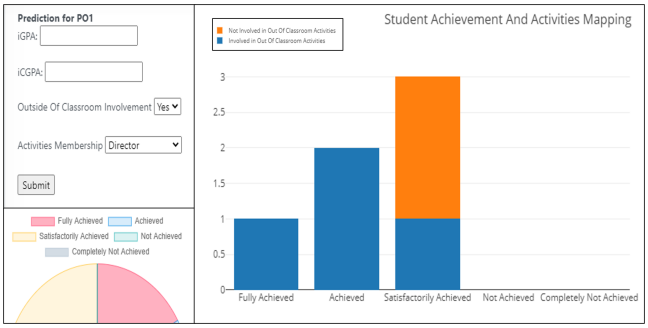


Fig. 5. Prediction Assessment Panel for Student Achievement and Activities.

in visualization system usability testing with domain experts. Subsequently, this section discusses the significant findings from usability testing in terms of effective visualization, modeling, and knowledge generation.

A. Case Study with Domain Experts

To evaluate and demonstrate the effectiveness of the system, this study conducted a usability testing with the domain experts. This test specifically measures OBEInsights' usability effectiveness in visualizing OBE predictive assessment. 10 domain experts from the higher education field with great proficiency and expertise in OBE participated in the test. The test was performed in an observation session with each expert for approximately 1 hour. The session starts with a brief introduction of the features of OBEInsights with an allocated 15 minutes for the experts to use the system.

After the introduction, the test began with inquiring the experts to perform several OBE analyses tasks as followings.

1) Display the overall performance of a student.

TABLE IV. SUS SCORE INTERPRETATION

SUS	Grade	Adjective
>80.3	A	Excellent
68 - 80.3	B	Good
68	C	Okay
51 - 68	D	Poor
<51	F	Awful

TABLE V. USABILITY SCALE RATING OF OBEINSIGHTS

Usability Scale	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
I will use this system frequently.	5	3	5	4	3	5	3	4	4	5
The system is unnecessarily complex.	2	2	1	2	2	3	1	1	1	2
The system is easy to use.	4	4	5	4	4	4	4	3	4	4
I will need technical support to use this system.	2	2	1	2	2	3	2	2	2	1
The system functions are well integrated.	4	3	4	4	4	3	4	3	4	3
The system has too many inconsistencies.	3	3	2	3	3	2	2	2	3	1
I can learn to use the system quickly.	4	4	5	4	3	4	4	5	3	4
The system is cumbersome to use.	2	1	1	2	2	1	2	1	3	2
I feel confident in using the system.	4	4	4	4	4	4	4	4	3	4
I need to learn many things to use the system.	2	2	1	2	5	3	4	2	1	1
Total Scale (x/50)	32	29	29	31	32	32	30	27	28	27
SUS Score (%)	**75**	**70**	**75**	**72.5**	**70**	**75**	**70**	**77.5**	**70**	**82.5**

TABLE VI. VISUALIZATION SCALE RATING OF OBEINSIGHTS

Visualization Scale	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Featured visualizations offers different perspectives.	4	4	5	5	5	5	4	4	4	4
Interactivity features helps my ability in data exploration.	4	3	4	4	4	5	4	4	5	3
Visualizations displays OBE performance data effectively.	5	4	4	5	5	4	4	3	5	4
Featured visualizations allows me to discover new patterns.	4	4	5	4	5	3	4	5	4	5
Visualizations enables me to explore and discover insights.	4	5	4	3	5	4	4	5	3	5
Total Scale (x/25)	21	20	22	21	24	21	20	21	21	21
SUS Score (%)	**84**	**80**	**88**	**84**	**96**	**84**	**80**	**84**	**84**	**84**

2) Find the relationship between the student's PLO achievement and extracurricular activities.
3) Review a student's iCGPA progression for specific PLO over the study duration
4) Predict a student's future performance for specific PLO with customized parameters for extracurricular activities, position, iGPA and iCGPA.

The purpose of inquiring about these tasks is to allow the users to practically utilize the featured interactive visualization to perform major OBE analysis. Furthermore, the effectiveness and limitation of OBEInsights in facilitating the analysis from the direct perspective of the domain user can be observed. Upon completing the tasks, the experts were then asked to answer two sets of questionnaires that were designed based on the system usability scale (SUS) [26]. The resulting scores can be interpreted based on rating as shown in Table IV. The questionnaires inquire the experts' perception towards the usability and visualization of OBEInsights by scale rating. The results of the usability testing are shown in Table V and Table VI.

B. Discussion

Based on the presented results, this study further discusses and reflects on the significant findings from the conducted design evaluation. This study found several visualization design insights that can be explored in three spectrums: visualization, predictive modeling, and knowledge generation as discussed as follows.

1) Effective Visualization Design for Student's Holistic Development: Referring to earlier domain characterization, this study identified three major analyses in OBE assessment: (1) Overview of student results and activities, (2) Student's PLO results over the semesters, and (3) Relativity of student extracurricular activities with performance. The domain characterization helps specify the analysis interests and practices from OBE analyst and practitioners to identify crucial design requirements. The factors of visualization familiarity and preferences plays a role in enhancing the learnability of the system [27]. The evaluation result in Table V and Table VI indicates that the featured interaction and visualization encodings applied into OBEInsights are effective in facilitating OBE predictive analysis. With average SUS score of 84.8%, the experts' rating demonstrates and validates the effectiveness of the design in visualizing OBE predictive assessment. The inclusion of visualization literacy into the design consideration greatly affects the perceived effectiveness of presented visualization to the users [28]. Despite the positive results, the implemented design only employs basic visualization that can be limited when visualizing advanced or complex analysis scenario. This study are interested to explore the implementation of advanced visualization into OBEInsights to facilitate complex OBE analysis. Furthermore, we intend to reiterate and refine the current design especially on main interfaces and deployment by incorporating user experience design.

2) Modeling Accuracy on Predicting Student's Performance: Based on the earlier accuracy experiment, we learn that the 'Random Forest' algorithm yields the highest accuracy rate on low-prediction (91%) and high-prediction (94%) for OBE prediction assessment. The prediction results were also validated with misclassification errors. Our findings correlate with other studies that also found 'Random Forest' and 'Gradient Boosting' to produce accurate prediction results. However, the predictive algorithms that were investigated in this study were typical in handling structured datasets. Future research can investigate further on predictive machine learning algorithms with regards to flexibility in handling unstructured datasets. Despite the positive accuracy result, our study has not included human-user confidence in its design consideration and evaluation. Further investigation on human-user confidence towards the automated OBE analysis output helps curate the

prediction reliability. In addition, the datasets used in this study only pertain information limited to one study program and limited extracurricular activity variables. Analyzing many different OBE datasets from other institutions using the visual analytics system may yield different results or reveal interesting insights.

3) Framework Supports on Knowledge Generation: OBEInsights was built based on deliberate guided consideration towards many factors in the conceptualized design framework in Fig. 1. Inspired by the visualization mantra "Overview, zoom and filter, then details-on-demand" [29], the framework was designed to support analysis tasks in each step in the top-down decision-making process [30]. Starting from raw data, the domain user first overviews the entire OBE information from the dataset with no specific analysis subject. Next, the user specifies their analysis by selecting subjects and parameters for the automated data analysis to process. The automated analysis design needs to facilitate the user's potential OBE analysis interest throughout their data exploration. Finally, the generated output needs to be encoded into appropriate visualization and interaction features to be displayed to the user. Currently, OBEInsights facilitate knowledge generation in a linear way, having the user select or customize specific analysis subjects and parameters. We intend to reiterate the design framework to support multifaceted analysis that allows for the inclusion of multiple subjects and parameters. Enabling multifaceted analysis could enrich the knowledge generation on pedagogical consideration for attaining learning outcomes. The presented design framework and system only cover the OBE assessment scenario that involves standardized parameters and specific stylized pedagogical decision-making in Malaysia. Other researchers can pursue further investigation on designing visual analytics for many different OBE analyses with regards to unique geographical contexts or variants.

V. Conclusion and Future Work

This paper has presented the empirical investigation in designing visual analytics for supporting OBE predictive assessment on learning activities and performance. The findings from this study offer insight into visual analytics design development for OBE analysis particularly in automated prediction and interactive visualization. This paper also presented the predictive visual analytics system namely OBEInsights along with elaborate descriptions of its design conceptualization and development. The effectiveness of OBEInsights is demonstrated and validated via a design evaluation study with domain experts. Design evaluation results indicate the system's effectiveness in facilitating OBE predictive assessment with great usability rating. Moreover, this study also learns the impact of crucial factors like visualization familiarity and preference towards the system usability. For constructing automated analysis components, several recommended algorithms were experimented with to determine the best predictive modeling for OBE assessment. The experiment result reveals the Random Forest model as the highest performing technique for OBE predictive modeling. In terms of supporting knowledge generation, a visual analytics design framework was conceptualized to guide visualization development for OBE assessment. The design evaluation results also demonstrated the framework capability via the deployed OBEInsights in supporting major OBE analysis tasks. The scope of this study is limited to OBE predictive assessment with specific parameters and pedagogical decision-making. This study recommends further investigation on visualizing many advanced OBE analysis that handles different context, variables, or users. Future research can also explore the potential of incorporating human-confidence design and unstructured handling capabilities. In our future works, we intend to reiterate the design and explore its performance against many different datasets that possess different structures and complexity. Furthermore, this study intends to enable multifaceted analyses to augment the knowledge generation process for OBE predictive assessment.

Acknowledgment

This work is supported by the Ministry of Education Malaysia under the research grant FRG0499-2018.

References

[1] N. Rao, "Outcome-based education: An outline," *Higher Education for the Future*, vol. 7, no. 1, pp. 5–21, 2020.

[2] A. Yusof, N. Naim, M. Latip, N. Aminuddin, and N. Ya'acob, "Implementation of integrated cumulative grade point average (iCGPA)," in *Towards academic excellence in Malaysia. 2017 IEEE 9th International Conference on Engineering Education (ICEED*, 2017, p. 106–109.

[3] J. J. Thomas, *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[4] P. Caserta and O. Zendra, "Visualization of the static aspects of software: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 7, p. 913–933, 2010.

[5] R. Therón, "Visual Learning Analytics for a Better Impact of Big Data," in *Lecture Notes in Educational Technology*, 2020, p. 99–113.

[6] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Computers & Education*, vol. 122, p. 119–135, 2018.

[7] R. Harden, "Outcome-based education: the future is today," *Medical Teacher*, vol. 29, no. 7, p. 625–629, 2007.

[8] M. Asli and M. Hamzah, "Mobile visualization: A framework to assist fund management decision-making process," in *Proceedings of the 24th International Business Information Management Association Conference*, 2014, p. 1888–1895.

[9] N. Siddiquee, J. Xavier, and M. Mohamed, "What works and why? Lessons from public management reform in Malaysia," *International Journal of Public Administration*, vol. 42, no. 1, p. 14–27, 2019.

[10] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski, "The State-of-the-Art in Predictive Visual Analytics," *Computer Graphics Forum*, vol. 36, no. 3, p. 539–562, 2017.

[11] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[12] W. Jentner and D. A. Keim, "Visualization and visual analytic techniques for patterns," *High-Utility Pattern Mining*, pp. 303–337, 2019.

[13] R. Vatrapu, C. Teplovs, N. Fujita, and S. Bull, "Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making," in *ACM International Conference Proceeding Series*, 2011, p. 93–98.

[14] H. He, O. Zheng, and B. Dong, "VUSphere: Visual Analysis of Video Utilization in Online Distance Education," in *IEEE Conference on Visual Analytics Science and Technology, VAST 2018*, 2018, p. 25–35.

[15] C. Y. Sung, X. Y. Huang, Y. Shen, F. Y. Cherng, W. C. Lin, and H. C. Wang, "Exploring Online Learners' Interactive Dynamics by Visually Analyzing Their Time-anchored Comments," in *Computer Graphics Forum*, vol. 36, no. 7. Wiley Online Library, 2017, pp. 145–155.

[16] M. Asli, M. Hamzah, A. Ibrahim, and A. Jimat, "Visual Analytics: Design Study for Exploratory Analytics on Peer Profiles, Activity and Learning Performance for MOOC Forum Activity Assessment," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 1, p. 66–72, 2019.

[17] Q. Chen, X. Yue, X. Plantaz, Y. Chen, C. Shi, T. Pong, and H. Qu, "ViSeq: Visual Analytics of Learning Sequence in Massive Open Online Courses," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 3, p. 1622–1636, 2020.

[18] J. Wong and X. Zhang, "MessageLens: A Visual Analytics System to Support Multifaceted Exploration of MOOC Forum Discussions," *Visual Informatics*, vol. 2, no. 1, p. 37–49, 2018.

[19] A. Jääskeläinen and J. Roitto, "Visualization techniques supporting performance measurement system development," *Measuring Business Excellence*, 2016.

[20] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, p. 2431–2440, 2012.

[21] C. Knoll, A. Cetin, T. Möller, and M. Meyer, "Extending recommendations for creative visualization-opportunities workshops," in *IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*. IEEE, 2020, pp. 81–88.

[22] D. Keim, G. Andrienko, J. D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information Visualization*. Springer, 2008, pp. 154–175.

[23] L. Lumius, M. Hamzah, C. Yee, V. Pang, and G. Leng, "Visual Analytics Design To Support Knowledge Generation: The Case Of Outcome Based Education Assessment in Malaysia," in *IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2020, p. 135–140.

[24] A. Endert, W. Ribarsky, C. Turkay, B. Wong, I. Nabney, I. Blanco, and F. Rossi, "The state of the art in integrating machine learning into visual analytics," *Computer Graphics Forum*, vol. 36, no. 8, p. 458–486, 2017.

[25] M. Hamzah and T. Vu, "A taxonomy of twitter data analytics techniques," in *Proceedings of the 32nd International Business Information Management Association Conference*, 2018, p. 3434–3459.

[26] J. Lewis, "The system usability scale: past, present, and future," *International Journal of Human–Computer Interaction*, vol. 34, no. 7, p. 577–590, 2018.

[27] M. Asli, M. Hamzah, A. Ibrahim, and E. Ayub, "Problem characterization for visual analytics in MOOC learner's support monitoring: A case of Malaysian MOOC," *Heliyon*, vol. 6, no. 12, p. 05733, 2020.

[28] S. Lee, B. Kwon, J. Yang, B. Lee, and S. Kim, "The correlation between users' cognitive characteristics and visualization literacy," *Applied Sciences (Switzerland)*, vol. 9, no. 3, p. 488, 2019.

[29] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The craft of information visualization*. Elsevier, 2003, p. 364–371.

[30] M. Hamzah, A. Sobey, and A. Koronios, "Supporting decision making process with information visualisation: A theoretical framework," in *2nd IEEE International Conference on Information Management and Engineering*, 2010, p. 267–271.

Modelling and Simulating Exit Selection during Assisted Hospital Evacuation Process using Fuzzy Logic and Unity3D

Intiaz Mohammad Abir, Ali Ahmed Ali Moustafa Allam, Azhar Mohd Ibrahim*
Department of Mechatronics Engineering
International Islamic University Malaysia
Kuala Lumpur, Malaysia

***Abstract*—Evacuation procedures are an integral aspect of the emergency response strategy of a hospital. Evacuation simulation models help to properly evaluate and improve evacuation strategies. However, the issue of exit selection during evacuation is often overlooked and oversimplified in the evacuation simulation models. Moreover, most of the available evacuation simulation models lack integration of movement devices and assisted evacuation features. However, finding a solution of these limitations is a necessity to properly evaluate evacuation strategies. To tackle this problem, we propose an effective approach to model exit selection using a fuzzy logic controller (FLC) and simulate assisted hospital evacuation using Unity3D game engine. Our research demonstrates that selecting exits based on distance only is not sufficient for real life situation because it ignores the unpredictability of human behavior. On the contrary, the use of the proposed FLC for exit selection makes the simulation more realistic by addressing the uncertainty and randomness in an evacuee's decision-making process. This research can play a vital role in future developments of evacuation simulation models.**

Keywords*—*Evacuation simulation; exit selection; fuzzy logic; unity3D

I. Introduction

Hospitals are generally called a safe building and an emergency care centre for individuals. Planning for disasters is essential to reduce the number of disaster-related deaths and injuries. However, Planning a well-timed evacuation of complex buildings such as hospitals is difficult. Hospitals have many patients with fractures or injuries. Majority of them face evacuation difficulties as they must use wheelchairs or crutches and can't move normally.

It can be asserted that the behavioural reaction of the evacuees throughout the movement plays a vital role in the evacuation process. The modelling of the movement of the evacuees which is correlated with behavioural reaction has been given importance while designing evacuation models. For safe evacuation, the importance of designing evacuation models to predict the evacuation process is paramount [1] Some of these models are; SGEM [2], [3], [4], SIMULEX [5], [6] EXODUS[7] , EGRESS [8], [9]. Behavioural reaction plays a major impact in the evacuation process and choosing an exit is a comparatively complex factor while evacuating. The term exit not only refers to the final exit which leads outside the structure but also internal exits which lead the evacuees from one confined area to another. As there are several exit options available in a multi-exit structure, the evacuees face a major dilemma while choosing the right exit. In a situation where evacuation is necessary, everything occurring in the environment works as stimuli and shapes the reaction of the evacuees (e.g., the activity of other evacuees [1]).

Today's evacuation simulation software allows designers to easily evaluate evacuation performance for various conditions and designs of the building's internal structure. Over the years, a vast number of models have been developed for general building evacuation simulation. However, it could be impossible or inappropriate to use most of these models for simulating hospital evacuation as it is necessary for hospital environments to include movement devices (e.g., wheelchair and crutches) and assisted evacuation features (i.e., the ability to assist wheelchair users to evacuate). Moreover, proper modelling of exit selection is needed to make the simulation process smoother. Hence, we propose an effective approach to select a proper exit for each evacuee agent (doctors, nurses, patients, or visitors) using a fuzzy logic controller (FLC) and develop an assisted hospital evacuation simulation model using Unity3D game engine.This model can be used by the hospital authorities to calculate the total evacuation time of the evacuees. In this way, the efficiency of evacuation measures will also be quantified which will lead to a safer evacuation planning and designing.

The rest of the paper proceeds as follows: Section II explores previous research related to our work. Section III explains the methodology and implementation of this research. Section IV analyses and discusses the results. Finally, Section V provides the conclusion of the research.

II. Related Work

There are many evacuation models that can simulate the evacuation process of a general building. However, most of these models lack proper integration of movement devices and assisted evacuation features as these models were developed to simulate ambulant evacuees. Some of the current evacuation models which are not explicitly designed to simulate assisted evacuation but are sufficiently flexible to accomplish this purpose indirectly. Disabled people were included in certain models by reducing their speeds. For example, FDS+Evac [10] model simulates the evacuation of elderly agents at a lower speed than normal agents. The assisted evacuation of hospitals and wheelchairs was only investigated by a small range of studies. For example, Hunt et al. [11], Alonso et al. [12],

Ursetta et al. [13], and Rahouti et al. [14] simulated assisted evacuation procedures in a hospital environment. However, these models lack proper modelling of exit selection.

It is important to take exit selection on a serious note as the behavior of an individual has considerable impact in the evacuation process and also has the potential to cause another emergency [15]. Complications arise when there is an imbalance in the evacuee crowd. If the structure has uneven design of exits it will result in an imbalance in the number of evacuees in the exits [16]. Researchers demonstrated that some components have direct impact on the evacuees' decision making such as, exit distance, the decision of following other individuals, the exit being in the range of vision, availability of light in the surrounding, crowd distribution, human psychology, width of the doors, the capacity of a particular exit , obstacle position, the queue length in the exits, familiarity with the structure, density of the crowd in particular exits, exit familiarity, angle made by the exit in respect to current movement direction, movement direction of other evacuees and social influence [1], [17], [18], [19], [20], [21], [22], [23], [24] . A study by Lovreglio et al. [25] demonstrated that in certain circumstances evacuees crowd up to particular exits avoiding other available exits. Wang et al. [26] analyzed the behavior of evacuees during a panic situation by using a combination of automata and multi-agent based model and demonstrated that the selection of exit behavior varies based on the crowd around that exit. These findings were also validated by Xu et al. [27].

Many models have been developed by researchers in past decades in order to regulate exit selection during evacuation [28], [29], [30], [31], [32]. Dynamic background field [33], [34], bayesian game theory [35], simulated annealing (SA) and depth-first search (DFS) [36] are some of the proposed methods to solve this issue. The selection of exit is made by considering the reciprocity of the group of evacuees in a game theory based model proposed by Lo et al. [1] which can be considered as one of the most notable works done in this research field.

In recent years, researchers have adopted methods like fuzzy logic, least effort algorithm, game theory, random utility theory, modified multinomial logit model, reinforcement learning to model exit selection. For example, Yang et al. utilized fuzzy logic to model exit selection using two input parameters: Normalized distance and normalized density [37]. Liu et al. followed a similar approach like Yang et al. However, they combined the output of the fuzzy logic system with exit width and herding behavior to determine the target exit [38]. Wang et al. utilized game theory to model exit selection. He considered factors like distance, visual range and choice firmness [39]. Zhang et al. proposed a multi-exit selection model considering three factors: distance, density and exit width [40]. Fu et al. proposed Two multi-exit selection models based on the social force model to analyze the dynamic change in exit selection by considering the effect of the exit distance, exit width, crowd number and crowd distribution [41]. Xu et al. proposed deep reinforcement learning based exit selection model named MultiExit-DRL [42]. Cao et al. implemented exit selection model based on random utility theory [15]. Edrisi et al. proposed three different exit choice models which are: the shortest path exit choice, the multinomial logit model, and the modified multinomial logit model with revising decisions. Their research demonstrated that the modified multinomial logit model with revising decisions outperforms the other two models [43]. Ma et al. modified the social force model and proposed an integrated exit selection model where evacuees can observe nearby evacuees and choose the appropriate exit by calculating the shortest estimated evacuation time [44]. Fu et al. combined least effort algorithm with a cellular automaton model to model exit selection while considering the distance to exits and crowd density around exits [45].

The movement of the evacuees can be distributed in three levels and among them, exit choice belongs to the highest strategic level. Mostly it is presumed that the exit selection is done entirely based on shortest distance optimization [46], [47], [48]. However, considering only distance is not enough to properly model the exit selection problem as in real time situations evacuees consider other relevant factors as well. Moreover, from the above review we can conclude that most of the existing exit selection models considered environmental factors (e.g., exit width, crowd density) while ignoring psychological factors (e.g., exit familiarity). Previous works/experiments suggest that factors like familiarity [18], [24] and visibility [22], [23], [24] can play a vital role in exit selection as human behavior is unpredictable and does not depend on one single factor. Thus, a simpler method is necessary for helping the evacuee agents in quick decision making. This paper proposes a new approach which considers psychological factor (exit familiarity) and environmental factors (exit distance and exit visibility) of exit selection by using a Fuzzy Logic Controller (FLC). Hence, it will be able to address the unpredictability of human behavior to some extent. These three factors are taken from the available literature that were discussed previously. This method will be able to provide a more realistic evacuation simulation compared to other methods where exit selection is done solely based on distance or on random.

III. Methods

Evacuation simulation models are important for investigating various evacuation strategies suitable for different scenarios. Reviews of general evacuation simulation models indicate that these models were mainly developed for normal building environments where evacuees can usually move unaided. Only a few studies are available on assisted evacuation.

In this paper, we propose an assisted evacuation simulation model to simulate the evacuation process of a hospital where assistants transfer non-ambulant patients from a risky place to a safe place using hospital devices such as wheelchairs and other ambulant evacuees (e.g., Doctors, Nurses, patients, and visitors) evacuate on their own. The proposed model utilizes a FLC to regulate the exit selection behavior of the ambulant evacuees (e.g., Doctors, Nurses, patients, and visitors). This section describes the details of the proposed system design.

A. Software

The simulation was carried out using Unity3D which is a game development engine developed by Unity Technologies. This software is capable of building 3D, 2D, Virtual Reality games, simulations, and other interactions. Fuzzy logic

modelling was done using MATLAB and later implemented in Unity3D. Finally, 3DS Max, which is a professional 3D graphics computer tool designed to create 3D animations, models, and games was used to design the agents and the movement devices

B. Types of Agents

Occupants with different levels of dependence will be expected at health facilities. To better simulate the process of evacuation in this form of setting, it is important to introduce two types of agents: ambulant agents such as patients, employees or visitors who can move without help, and non-ambulant agents such as dependent and highly dependent patients. For the simulation, six types of agents were considered. These are: doctors, nurses, visitors, patients with crutches, dependent patients with wheelchairs, independent patients with wheelchairs. Fig. 1 illustrates the different agents.

C. Movement Devices

Explicit representation of movement devices is a necessary requirement to simulate the evacuation of health facilities. Two types of movement devices were considered for this simulation. These are wheelchairs and crutches. The wheelchairs can be operated in two ways. In the first case the patient is disabled but independent. Which means the patient can operate the wheelchair on its own. In the second case the patient is both disabled and dependent, so assistance is required from ambulant agents (e.g., staff, nurse) to move the wheelchair. On the other hand, crutch users can move on their own without requiring any assistance. The usage of movement devices is illustrated in Fig. 1.

D. Layout Design

In this paper, the simulated hospital area is a W x H hall with 40m x 60m dimension. The considered floor has twenty-one rooms for patients including beds, one room for nurses, cafe, one room for dentist, one room for X-ray, one room for general practitioner and reception. Fig. 2 shows the layout of the simulation area.

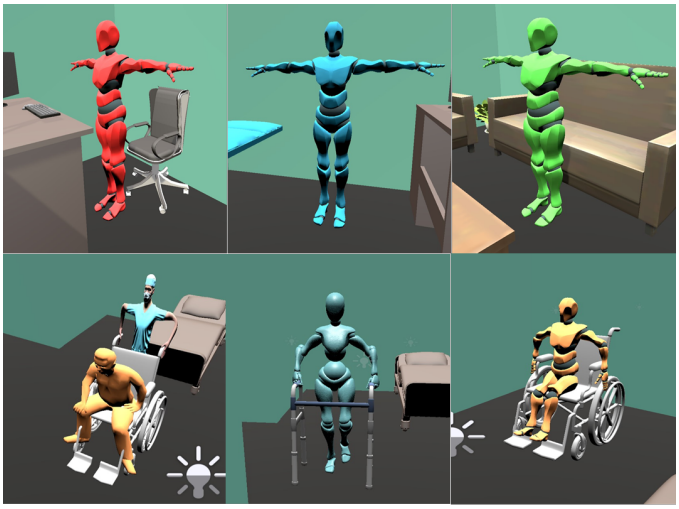


Fig. 1. Agents and Movement Devices used for Simulation. Doctors(red), Nurses(blue), Visitors(green), Patients with Crutches, Independent Patients with Wheelchair, Dependent Patients with Wheelchair.

E. Agent's Navigation System

For this research, we used Unity's built-in navigation system named NavMesh AI [49] which is divided into two main components. The first one is NavMesh (Navigation Mesh) area, which is used to define navigable areas in the environment. This component specifies areas where agents can walk, as well as the position of obstacles that the agent needs to avoid. This system is used for pathfinding and AI-controlled navigation. The second one is the NavMesh Agent component which is used to define an object as agent and also to set the agents' characteristics and features. This NavMesh AI helps agents to avoid each other, move around the environment towards a goal/target (e.g., exit). The agents can only move in the walkable surfaces while avoiding the obstacles and other agents. In Fig. 2 the blue area is the NavMesh area which is the agents' walkable surface. No agent can navigate in the remaining area. The walls and the furniture are obstacles for agents and the agents must avoid them while navigating.

F. Modelling Exit Selection

Most of the classical mapping or classification techniques, regardless of the number of classes or sets can be normalized down to two sets or classes namely 0 (false) and 1 (true). When an element belongs to a set, binary mapping evaluates it as 1 or true. If an element does not belong to a set, it is evaluated as 0 or false. Because of their objective nature traditional mapping methods only apply to exact correspondence. But in real life there are many relationships which are not so black and white. So, these relationships can't be described using the traditional binary mapping techniques. Fuzzy logic however can be utilized to solve this issue to some extent. It is a mathematical modelling method based on fuzzy sets and the related membership functions. In fuzzy logic "degrees of truth" is taken into consideration rather than exact true or false (1 and 0). Unlike crisp sets, a fuzzy set allows partial belonging to a set, that is defined by a degree of membership, denoted by μ, that can take any value from 0 (element does not belong at all in the set) to 1 (element belongs fully to the set).

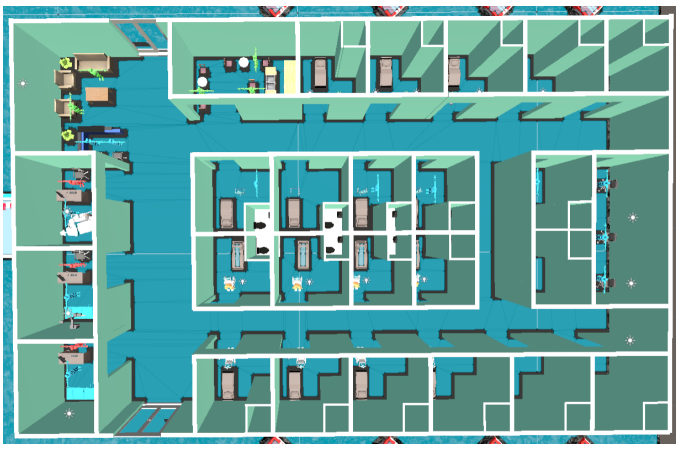


Fig. 2. Layout of the Simulation Environment.

Everything in between 0 and 1 denotes the extent to which the element belongs to one set or the other. A fuzzy system is a repository of the fuzzy expert knowledge that can interpret data in subjective terms instead of precise Boolean logic. The expert knowledge is a collection of fuzzy membership functions and a set of fuzzy rules, known as the rule-base, having the form: IF (conditions are fulfilled) THEN (consequences are inferred)

Fuzzy logic is proposed in this paper to deal with human behavior uncertainty while selecting the proper exit during evacuation. As these parameters are hard to quantify using crisp calculation. The integration of fuzzy logic into computer models produced positive results where perception, feelings and judgment play a significant role. Therefore, human thinking and environmental adjustments are integrated in the proposed model using fuzzy laws to model the evacuee agent's exit selection behavior during the evacuation.

There are four major parts of a typical Fuzzy Logic Controller, namely fuzzifier, rule-base, inference engine and defuzzifier.

The fuzzifier performs fuzzification which is the process of converting crisp input values into fuzzy data points. In this process fuzzifier obtains the membership degree of the fuzzy set from the specific input according to the membership function. For this research, we have utilized triangular membership functions for all the linguistic variables.

After attaining the degree of membership for each input value from the fuzzifier, a rule-base is needed which contains the if - then rules/conditions that are required to form the inference process of the output variable. Generally, the rules are expressed by the fuzzy natural language. The number of fuzzy rules needed can be determined by multiplying the number of linguistic variables of all the inputs, for example, our proposed FLC has 3 input variables: Distance (four linguistic variables), visibility (three linguistic variables) and familiarity (three linguistic variables) so a total of (4 x 3 x 3) = 36 if -then rules are required for the rule base. The rule-base of our proposed FLC is demonstrated in Table I.

The Inference Engine provides the decision-making logic of the controller. By applying the fuzzy rules of inference, it evaluates the fuzzy input values and the provided rules to deduce the fuzzy output values. For this research, we have adopted a typical Mamdani fuzzy inference engine [50], [51].

Finally, the defuzzifier converts the fuzzy output into quantifiable and objective crisp output. In this paper the output of the defuzzifier is the probability of choosing a certain exit. The centroid defuzzification method is used for the defuzzification process of our proposed FLC.

Our proposed FLC has three input variables and one output variable. These are described below:

1) Inputs:

Visibility: Visibility represents the visibility of the exits which may vary due to different smoke conditions or crowd density. The visibility of each exit is assigned randomly. The variable has three linguistic variables: Low Visibility (LV), Medium Visibility (MV), High Visibility (HV). The range of visibility is from 0-10.

Distance: This variable provides the distance between an agent and an exit. The variable has Four linguistic variables: Very Near (VN), Near (N), Far (F), Very Far (VF). The range of distance is from 0-45 meter.

Familiarity: Familiarity represents how familiar an agent is with a specific exit. Same as visibility, familiarity is also selected randomly. The range is from 0-10. The variable has three linguistic variables: Not Familiar (NF), Familiar (F), Very Familiar (VF). Figure 3 visualizes the input variables

2) Output:

Probability of Exit Selection : This variable determines the selection probability of an exit. The variable has 8 membership functions: Zero Probability (Z), Very Low (VL), Medium Low (ML), Low (L), Medium High (MH), High (H), Very High (VH), One (O). Fig. 4 visualizes the output variable.

IV. Results And Discussion

This section describes the result of the simulation and discusses the findings of this research. At the beginning, the agents were placed randomly in the simulation area. When the simulation begins, the unity NavMesh system allows the agents to move into the target point (e.g., exit). The FLC determines the exit for each agent. The simulation ends when all the agents evacuate from the floor. Fig. 5 shows the initial and the final position of the agents. The simulation was carried out by considering two cases.

In the first case the proposed FLC was not included. In the second case the process was again repeated with the inclusion of the FLC. A total of 44 agents were simulated in the environment (3 doctors, 15 nurses, 5 patients with crutches, 11 patients with nurses, 5 patients with wheelchairs and 5 visitors).

For simplicity and ease of visualization only the results of 10 randomly picked agents were included in the tables (1 doctor, 2 nurses, 2 patients with crutches, 2 patients with nurses, 2 patients with wheelchair, 1 visitor). The simulation was conducted 10 times for each case. The details of the simulations are provided below.

A. Results

1) Case 1 (Without using FLC): The simulation is conducted without the FLC for the first experiment. The speed of each agent was set to 1.4m/s. The agent will choose the nearest exit as we have only considered distance in this case. Table II provides the simulation results of 10 randomly picked agents for the first experiment. The table shows the evacuation time (in seconds) of each agent for each simulation run and the standard deviation of the evacuation time

2) Case 2 (Using FLC): For the second experiment, the appropriate exit selection was done by the proposed FLC instead of selecting the nearest exit. The FLC takes the distance, visibility, and familiarity of each of the exits from each agent to the exits and calculates the probability of selection for each exit. The exit with the highest probability of selection is chosen as the target for each agent. Similar to case 1 The speed of each agent was set to 1.4m/s. Table III provides the simulation results for the second experiment.

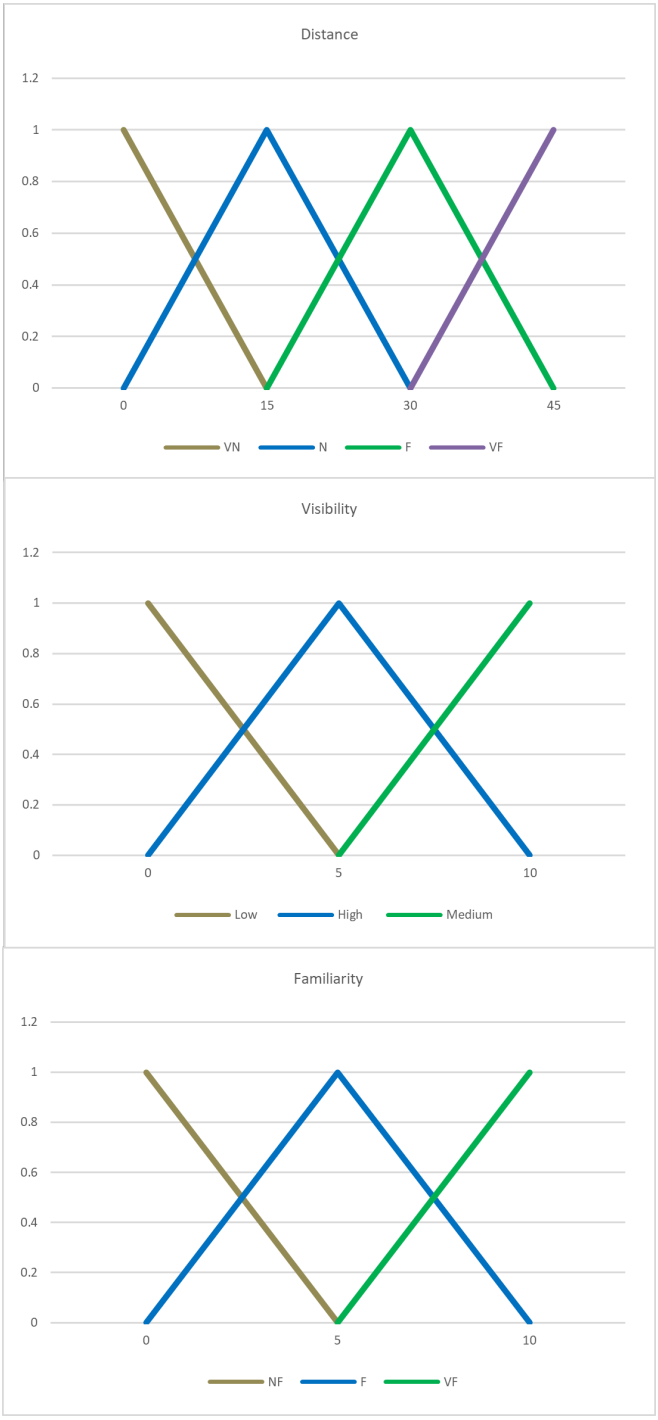


Fig. 3. Membership Function for Input Variables (Visibility, Distance, and Familiarity).

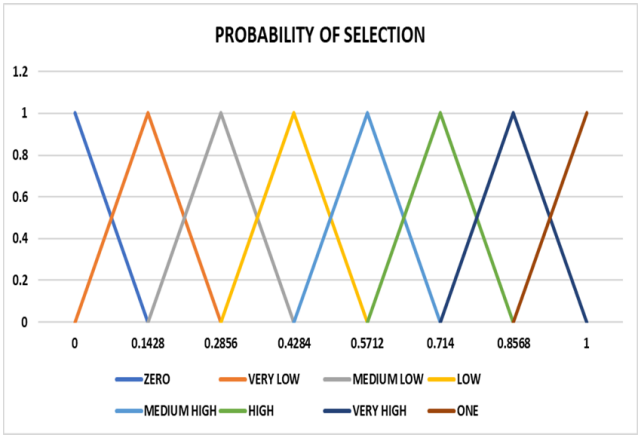


Fig. 4. Membership Function for Output Variable (Probability of Exit Selection).

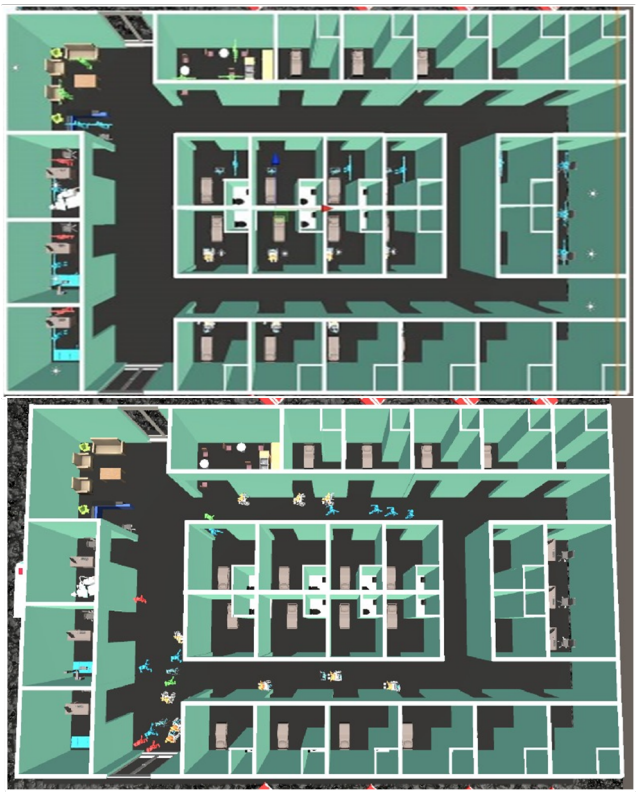


Fig. 5. The Initial Position of Agents (Top), The Final Position (Bottom) of the Agents During Simulation.

B. Discussion

When the exit is selected based on only distance the simulation fails to reflect human uncertainty and randomness. From Table II, it can be seen that the evacuation time of the agents remained almost the same for each test run and the standard deviation of the evacuation time is close to 0. On the contrary, when the FLC is utilized, the agents selected the exit based on distance, familiarity and visibility which impacted their evacuation time. From Table III it can be seen the standard deviation of the evacuation time of each agent is higher than experiment 1. This result provides an important finding. While simulating without using the FLC the exit is selected solely based on the distance. So, the agents always select the nearest exit without considering other factors which is not the case in real life evacuations. From the reviewed literature it can be seen that humans consider many factors (e.g., the exit being in the range of vision, availability of light in the surrounding, familiarity with the structure, exit familiarity) while choosing an exit for evacuation [1], [17]–[22]. This issue

is solved when the FLC is used as the FLC considers not only the distance but also the visibility and the familiarity of the exits. Our proposed method is able to reflect and demonstrate human uncertainty and randomness to some extent. Hence, this research demonstrates that considering only distance is just not enough to select the exit. Including other factors (e.g., visibility, familiarity) help to make the evacuation simulation more realistic and increase the evacuation performance.

V. Conclusion

The objective of this research is to model and simulate assisted hospital [15] evacuation by modelling exit selection for individual agents (doctors, nurses, patients, or visitors) using FLC. For this research, only three types of transport devices have been simulated. Our research findings demonstrate that the proposed FLC can reflect human randomness and uncertainty to some extent.

However, the research has some limitations. For example, the exit is selected only once for each agent. This approach is not very feasible as the visibility of the exits are constantly changing in a practical situation. A more dynamic solution is needed for the selection procedure to properly reflect the change in visibility and the effect it causes in exit selection. Another limitation of the proposed FLC is that it only reflects the effect of three factors. However, in reality humans consider many other factors while selecting the proper exit which are not included in the proposed FLC.

Other movement devices should be considered during simulation to imitate real life situations such as stretchers and rescue sheets. Moreover, multiple floors and intermediate exits should be added to the simulation to analyze its effect in evacuation performance. Finally, the effect of social distancing should also be considered in the simulation model as social distancing is a must in the time of pandemic.

Acknowledgment

This research is supported by the FRGS 2019 Grant: FRGS/1/2019/ICT02/UIAM/02/2 awarded by the Ministry of Education Malaysia. The first author would like to thank IIUM-TFW-2020 Scheme for sponsoring his postgraduate study

References

[1] S. M. Lo, H.-C. Huang, P. Wang, and K. K. Yuen, "A game theory based exit selection model for evacuation," *Fire Safety Journal*, vol. 41, no. 5, pp. 364–369, 2006.

[2] S. M. Lo and Z. Fang, "A spatial-grid evacuation model for buildings," *Journal of Fire Sciences*, vol. 18, no. 5, pp. 376–394, 2000.

[3] G. S. Zhi, S. M. Lo, and Z. Fang, "A graph-based algorithm for extracting units and loops from architectural floor plans for a building evacuation model," *Computer-Aided Design*, vol. 35, no. 1, pp. 1–14, 2003.

[4] S. M. Lo, Z. Fang, P. Lin, and G. S. Zhi, "An evacuation model: the sgem package," *Fire Safety Journal*, vol. 39, no. 3, pp. 169–190, 2004.

[5] P. A. Thompson and E. W. Marchant, "A computer model for the evacuation of large building populations," *Fire safety journal*, vol. 24, no. 2, pp. 131–148, 1995.

[6] ——, "Computer and fluid modelling of evacuation," *Safety Science*, vol. 18, no. 4, pp. 277–289, 1995.

[7] M. Owen, E. R. Galea, and P. Lawrence, "Advanced occupant behavioural features of the building-exodus evacuation model," *Fire Safety Science*, vol. 5, pp. 795–806, 1997.

[8] N. Ketchell, S. Cole, D. M. Webber, C. A. Marriott, P. J. Stephens, I. R. Brearley, J. Fraser, J. Doheny, and J. Smart, *The EGRESS code for human movement and behaviour in emergency evacuations*. University of Edinburgh, Artificial Intelligence Applications Institute, 1993.

[9] N. Ketchell, G. J. Bamford, and B. Kandola, "Evacuation modelling: a new approach," 1995, pp. 499–505.

[10] T. Korhonen, S. Hostikka, S. Heli, and H. Ehtamo, "Fds+evac: An agent based fire evacuation model," 0.

[11] A. Hunt, "Simulating hospital evacuation," no. January, p. 326, 2016.

[12] V. Alonso-Gutierrez and E. Ronchi, "The simulation of assisted evacuation in hospitals," *Femtc 2018*, 2018.

[13] D. Ursetta, A. D'Orazio, L. Grossi, G. Carbotti, S. Casentini, and L. Poggi, "Egress from a hospital ward: A case study," *Fire and Evacuation Modelling Technical Conference*, 2014.

[14] A. Rahouti, R. Lovreglio, C. Dias, and S. Datoussaïd, "Simulating assisted evacuation using unity3d," 2019, pp. 265–275.

[15] S. Cao, L. Fu, and W. Song, "Exit selection and pedestrian movement in a room with two exits under fire emergency," *Applied Mathematics and Computation*, vol. 332, pp. 136–147, 2018.

[16] J. Gao, J. He, and J. Gong, "A simplified method to provide evacuation guidance in a multi-exit building under emergency," *Physica A: Statistical Mechanics and its Applications*, vol. 545, p. 123554, 2020.

[17] S. Liu, L. Yang, T. Fang, and J. Li, "Evacuation from a classroom considering the occupant density around exits," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 9, pp. 1921–1928, 2009.

[18] M. Kinateder, B. Comunale, and W. H. Warren, "Exit choice in an emergency evacuation scenario is influenced by exit familiarity and neighbor behavior," *Safety science*, vol. 106, pp. 170–175, 2018.

[19] R. Alizadeh, "A dynamic cellular automaton model for evacuation process with obstacles," *Safety Science*, vol. 49, no. 2, pp. 315–323, 2011.

[20] D. C. Duives and H. S. Mahmassani, "Exit choice decisions during pedestrian evacuations of buildings," *Transportation Research Record*, vol. 2316, no. 1, pp. 84–94, 2012.

[21] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transportation Research Part B: Methodological*, vol. 40, no. 8, pp. 667–687, 2006.

[22] E.-W. Augustijn-Beckers, J. Flacke, and B. Retsios, "Investigating the effect of different pre-evacuation behavior and exit choice strategies using agent-based modeling," *Procedia Engineering*, vol. 3, pp. 23–35, 2010.

[23] S. Cao, L. Fu, P. Wang, G. Zeng, and W. Song, "Experimental and modeling study on evacuation under good and limited visibility in a supermarket," *Fire Safety Journal*, vol. 102, pp. 27–36, 12 2018.

[24] R. Y. Guo and H. J. Huang, "Logit-based exit choice model of evacuation in rooms with internal obstacles and multiple exits," *Chinese Physics B*, vol. 19, no. 3, p. 030501, 3 2010.

[25] R. Lovreglio, A. Fonzone, L. Dell'Olio, and D. Borri, "A study of herding behaviour in exit choice during emergencies based on random utility theory," *Safety Science*, vol. 82, pp. 421–431, 2016.

[26] J. Wang, L. Zhang, Q. Shi, P. Yang, and X. Hu, "Modeling and simulating for congestion pedestrian evacuation with panic," *Physica A: Statistical Mechanics and its Applications*, vol. 428, pp. 396–409, 2015.

[27] Y. Xu, H.-j. Huang, and L.-j. Tian, "Simulation of exit choosing in pedestrian evacuation using a cellular automaton model based on surrounding pedestrian density," 2011, pp. 1109–1112.

[28] H.-J. Huang and R.-Y. Guo, "Static floor field and exit choice for pedestrian evacuation in rooms with internal obstacles and multiple exits," *Physical Review E*, vol. 78, no. 2, p. 21131, 2008.

[29] S. Gwynne, E. R. Galea, P. J. Lawrence, M. Owen, and L. Filippidis, "Adaptive decision-making in response to crowd formations in building exodus," *Journal of Applied Fire Science*, vol. 8, no. 4, pp. 301–325, 1999.

[30] V. Schneider, "Modelling of human response and behaviour in complex surroundings," 2004.

[31] H. Zhao and Z. Gao, "Reserve capacity and exit choosing in pedestrian evacuation dynamics," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 10, p. 105001, 2010.

TABLE I. RULE-BASE OF THE PROPOSED FLC

Familiarity	NF				F				VF			
Distance / Visibility	VND	ND	FD	VFD	VND	ND	FD	VFD	VND	ND	FD	VFD
LV	H	MH	L	ML	VH	H	MH	L	O	VH	H	MH
MV	MH	L	ML	VL	H	MH	L	ML	VH	H	MH	L
HV	L	ML	VL	Z	MH	L	ML	VL	H	MH	L	ML

TABLE II. EVACUATION TIME OF THE SELECTED AGENTS (IN SECONDS) WITHOUT USING THE PROPOSED FLC

	Agent Name									
	doctor 2	nurse 11	nurse 13	patient with crutches 1	patient with crutches 5	patient with nurse 8	patient with nurse 9	patient with wheelchair 2	patient with wheelchair 5	visitor 3
Iteration 1	9.800118	5.635603	10.97213	9.990309	24.4834	10.97213	15.17973	17.25896	28.21373	11.28892
Iteration 2	9.716181	5.587833	10.84643	9.989389	24.49657	10.97705	15.10169	16.99366	28.18499	10.97705
Iteration 3	9.770281	5.758615	10.89974	10.05894	24.57523	11.02906	15.22316	17.1451	28.33407	11.1489
Iteration 4	9.711522	5.607227	10.89141	10.14683	24.72237	11.03564	15.1421	17.17188	28.43073	11.19462
Iteration 5	9.894517	5.7282	10.82855	9.894517	24.79922	11.01106	15.20018	17.237	28.28679	11.49648
Iteration 6	9.603971	5.838264	10.90553	9.714827	24.3032	11.01818	15.00109	16.95982	28.04262	10.90553
Iteration 7	9.65391	5.683241	10.84098	9.987705	24.38693	10.96639	15.11619	17.12051	28.19271	11.31287
Iteration 8	9.61427	5.75701	10.80946	9.976979	24.27111	10.95858	15.07268	16.99114	28.01997	10.95858
Iteration 9	9.618695	5.666526	10.93444	10.05021	24.62155	10.93444	15.21281	17.07364	28.23475	11.1718
Iteration 10	9.673256	5.837012	10.91276	9.827637	24.40374	11.05625	15.12542	17.03004	28.07324	11.05625
Standard Deviation	0.088822	0.08433	0.048783	0.117239	0.165068	0.037428	0.065905	0.100012	0.123954	0.174194

TABLE III. EVACUATION TIME OF THE SELECTED AGENTS (IN SECONDS) USING THE PROPOSED FLC

	Agent Name									
	doctor 2	nurse 11	nurse 13	patient with crutches 1	patient with crutches 5	patient with nurse 8	patient with nurse 9	patient with wheelchair 2	patient with wheelchair 5	visitor 3
Iteration 1	9.599457	5.165013	10.43153	16.64089	23.92897	19.12175	23.48237	16.52794	27.73574	10.57593
Iteration 2	14.29306	5.245591	15.24546	9.359359	23.90145	10.49932	14.70391	24.97406	27.9697	11.22695
Iteration 3	13.67541	5.218663	10.47915	16.58311	29.334	19.2174	14.64088	16.58311	27.39561	10.2729
Iteration 4	9.539491	5.217213	10.69946	16.53998	23.81956	10.49263	14.64291	16.53998	27.79864	10.36195
Iteration 5	9.663138	5.343233	14.33399	9.289529	29.65339	10.46342	23.68378	16.85125	34.72333	13.04313
Iteration 6	9.554058	5.295016	10.10942	9.554058	24.06516	10.382	23.66674	16.69532	34.06012	11.33598
Iteration 7	9.626184	5.279238	10.14894	9.505378	24.01919	10.49212	14.84437	16.70228	27.86506	10.73832
Iteration 8	9.560124	5.197507	10.42999	9.716187	24.26485	19.21355	23.93615	16.82404	34.94345	12.48743
Iteration 9	9.534514	14.91878	15.64993	9.717838	24.51029	10.43473	14.91878	16.78667	34.48695	11.68569
Iteration 10	14.40427	5.121847	15.47817	9.444882	29.55148	19.27566	14.97395	25.05231	34.71779	14.81874
Standard Deviation	2.089058	2.906826	2.37543	3.2450102	2.500984	4.2851	4.364833	3.331497	3.426042	1.360217

[32] W. Yuan and K. H. Tan, "An evacuation model using cellular automata," *Physica A: Statistical Mechanics and its Applications*, vol. 384, no. 2, pp. 549–566, 2007.

[33] T. Huan-Huan, D. Li-Yun, and X. Yu, "Influence of the exits' configuration on evacuation process in a room without obstacle," *Physica A: Statistical Mechanics and its Applications*, vol. 420, pp. 164–178, 2015.

[34] D. Li-Yun, C. Li, and D. Xiao-Yin, "Modeling and simulation of pedestrian evacuation from a single-exit classroom based on experimental features," *Acta Physica Sinica*, vol. 64, no. 22, 2015.

[35] B. L. Mesmer and C. L. Bloebaum, "Incorporation of decision, game, and bayesian game theory in an emergency evacuation exit decision model," *Fire Safety Journal*, vol. 67, pp. 121–134, 2014.

[36] H. A. Kurdi, S. Al-Megren, R. Althunyan, and A. Almulifi, "Effect of exit placement on evacuation plans," *European Journal of Operational Research*, vol. 269, no. 2, pp. 749–759, 2018.

[37] X. Yang, X. Yang, and Q. Wang, "Pedestrian evacuation under guides in a multiple-exit room via the fuzzy logic method," *Communications in Nonlinear Science and Numerical Simulation*, vol. 83, p. 105138, 4 2020.

[38] T. Liu, X. Yang, Q. Wang, M. Zhou, and S. Xia, "A fuzzy-theory-based cellular automata model for pedestrian evacuation from a multiple-exit room," *IEEE Access*, vol. 8, pp. 106 334–106 345, 2020.

[39] W.-L. Wang, F.-F. Wan, and S.-M. Lo, "Game theory model of exit selection in pedestrian evacuation considering visual range and choice firmness*," *Chinese Physics B*, vol. 29, no. 8, p. 084502, 7 2020.

[40] D. Zhang, G. Huang, C. Ji, H. Liu, and Y. Tang, "Pedestrian evacuation modeling and simulation in multi-exit scenarios," *Physica A: Statistical Mechanics and its Applications*, vol. 582, p. 126272, 11 2021.

[41] Y. Fu, W. Shi, Y. Zeng, H. Zhang, X. Liu, and Y. Liu, "Simulation study on pedestrian evacuation optimization in a multi-exit building," vol. 1780, no. 1, 2021, p. 12024.

[42] D. Xu, X. Huang, J. Mango, X. Li, and Z. Li, "Simulating multi-exit evacuation using deep reinforcement learning," *Transactions in GIS*, vol. 25, no. 3, pp. 1542–1564, 6 2021.

[43] A. Edrisi, B. Lahoorpoor, and R. Lovreglio, "Simulating metro station evacuation using three agent-based exit choice models," *Case Studies on Transport Policy*, vol. 9, no. 3, pp. 1261–1272, 9 2021.

[44] G. Ma, Y. Wang, and S. Jiang, "Optimization of building exit layout: Combining exit decisions of evacuees," *Advances in Civil Engineering*, vol. 2021, 2021.

[45] L. Fu, J. Fang, S. Cao, and S. Lo, "A cellular automaton model for exit selection behavior simulation during evacuation processes," *Procedia Engineering*, vol. 211, pp. 169–175, 1 2018.

[46] S. P. Hoogendoorn and P. H. L. Bovy, "Pedestrian route-choice and activity scheduling theory and models," *Transportation Research Part B: Methodological*, vol. 38, no. 2, pp. 169–190, 2004.

[47] G. Santos and B. E. Aguirre, "A critical review of emergency evacuation simulation models." NIST, Gaithersburg, USA, 2004, p. 339.

[48] H. L. Kluepfel, "A cellular automaton model for crowd movement and egress simulation," Ph.D. dissertation, 2012.

[49] Z. He, M. Shi, and C. Li, "Research and application of path-finding algorithm based on unity 3d," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*, 8 2016.

[50] MAMDANI and E. H., "Applications of fuzzy algorithms for control of simple dynamic plant," *Proc. IEE*, vol. 121, pp. 1585–1588, 1974.

[51] L. Fu, W. Song, and S. Lo, "A fuzzy-theory-based behavioral model for studying pedestrian evacuation from a single-exit room," *Physics Letters A*, vol. 380, no. 34, pp. 2619–2627, 8 2016.

Lattice-based Group Enlargement for a Robot Swarm based on Crystal Growth Models

Kohei Yamagishi[1]
Graduate School of Advanced Science and Technology
Tokyo Denki University
5 Senju Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan

Tsuyoshi Suzuki[2]
Department of Information and communication Engineering
Tokyo Denki University
5 Senju Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan

***Abstract*—Swarm robotic systems control multiple robots in a coordinated manner for using this flexible coordination to solve complex tasks in various environments. Such systems can utilize the individual capabilities of robots scattered within the swarm as well as the collective capabilities of the assembled robots. By coordinating these capabilities, swarms can solve tasks with a range of purposes, including carrying out rough sweeps of the overall environment using scattered robots or detailed observation of a part of the environment using assembled robots. This study developed a self-organization method for constructing regular groups of robots from scattered robots to achieve coordination between individual and collective states. An approach that integrates elements of self-organization with different input information requires centralized control to manage them. To provide this self-organization without centralized control, we focus on using the phase-field method and cellular automata to facilitate crystal growth that produces ordered structures from scattered particles. We formulate a method for arranging robots in a self-organizing manner based on the geometrical regularities of tile-able lattices (honeycomb, square, and hexagonal lattices) on a two-dimensional plane, demonstrate the process undertaken in carrying out the proposed method, and quantitatively evaluate the effectiveness of the lattice-based geometrical regularity approach. The proposed method contributes to carrying out tasks with a range of purposes by organizing states with either individual or collective capabilities of robot groups.**

***Keywords*—Multi-robot systems; self-organization; distributed control; crystal growth**

I. Introduction

Swarm robotic systems, which apply swarm intelligence through the control of multiple homogeneous robots, have the features of scalability, flexibility, and robustness [1]. In recent years, researchers have assessed techniques for the practical application of such systems [2], [3]. Changing the manner in which the swarm is embodied can enable the robots to process multiple small tasks using parallel capabilities as well as medium- to large-scale tasks using collective capabilities. This gives the robot swarm advantages relative to single robots in carrying out large-scale/wide-area tasks such as surveillance and environment exploration and cooperative tasks such as multi-shape object transportation [4]. In this study, we examined a robot swarm-based-transporting application that manages individual and collective capabilities simultaneously to enable parallel transportation of small objects that depends on the performance of a single robot and cooperative transportation of heavy or large objects that exceeds their performances by robot groups.

When a swarm system performs multiple similar tasks, the swarm divides into multiple robot groups to carry out the tasks in parallel. However, only this simple group structure is not suitable for performing some tasks. Considering the transporting task as an example, groups with scales, shapes, and structures should be constructed such that they satisfy the size, shape, and weight of an transported object, and the system should self-organize its groups based on the given conditions.

The self-organization of a swarm robot system involves aggregation, pattern formation, and self-assembly [5]. Aggregation is a method for allocating robots to several groups from a set of scattered robots or dividing a robot swarm into several groups. Several methods have been proposed for the allocation of robots based on external factors such as task value and distance [6], [7] and the division of robot swarms based on internal factors such as the number of tasks given by the host system [8], [9]. Allocation approaches allow for the flocking of scattered robots after information relevant to the tasks has been gathered from the environment; division approaches enable robots to work in rough groups to explore an environment. Pattern formation is a method for arranging robots in pre-designed shapes. In this approach, the robot swarm will often construct designed formations from pre-aggregated arbitrary shapes. The robots to be added will then search for the edge of the target group or region and converge to positions suitable for enlarging the pattern designed on a 2D plane [10] or in 3D space [11], or the edge of a designed region [12]. Recently, a method focusing on reaction-diffusion systems for morphogenesis through growth was proposed as a pattern strategy [13], [14]. This approach is expected to facilitate large-scale distributed patterning because it can adapt to changing self-healing defects caused by the partial failure of the robot swarm or changes in the environmental geometry. Self-assembly is the third method for maintaining either physical or cyber positioning between organized robots. Physical connections using grippers [15], magnets [16], and welding [17] as well as virtual connections using networks and non-contact sensors [18], [19] have been proposed to facilitate self-assembly through the fabrication of rigid or elastic body-like swarms without physical constraints, respectively. In pattern formation and self-assembly, the environment-adaptive structure [14], [17] produces groups suitable for foundation shapes and structural loads that self-organize via flexible coupling by controlling the reinforcement around heavily loaded robots. By contrast, lattice-based structures [11], [18] produce groups with geometrical regularities among robots. Depending on the geometrical conditions, dense or sparse groups can be constructed,

allowing a lattice-based structure to adapt to changes in robot density, that is, the number of robots required to carry out tasks such as coordinative transportation or observation. Thus, the implementation of self-organization requires the integration of each elemental method. This complicates the configuration of robots and systems.

In robot groups for transporting—which is the objective of this study—ordered arrangement enables the robots to efficiently support heavy or light objects on agerage by changing their density. Self-organization through these elemental technologies can be used to produce such an arrangement from scattered robots. However, this approach requires centralized control management because the information that must be used differs by task. To solve this self-organization challenge without centralized control, we focus on crystal growth, a natural phenomenon that produces lattice-based structures such as snowflakes, salt, and ores from scattered particles. This paper proposes a self-organization method for constructing groups of crystalline (ordered) robot arrangements from scattered robots without the use of centralized control. To this end, we formulate an autonomous distributed control model for introducing crystal growth into swarm robotic systems to induce enlargement.

In the process of crystal growth, particles find the surfaces of existing crystals through solidification and deposition and then adhere to positions on those crystals determined by existing meteorological conditions and molecular properties, thereby continually increasing the number of crystal layers and enlarging the crystal structure [20]. In this manner, crystal growth constitutes a self-organizational process. There are two existing mathematical tools for predicting and reproducing crystal growth: the phase-field method and cellular automata. The phase-field method describes the changes in a crystal surface during growth by calculating the state transitions of particles using scalar values that denote the stochastic state between the solid and liquid phases instead of independent thermodynamic states. Cellular automata reproduce the complex systems underpinning phenomena such as crystal growth and the formation of traffic jams. In this approach, rulesets for updating the states of cells in terms of discrete neighboring states are applied over discrete space-time intervals to reproduce macroscopic phenomena. By combining the state transitions of particles under the phase-field method with the growth rules of cellular automata, we seek to apply crystal growth to swarm robotic systems. To this end, we propose a self-organization method (Fig. 1) for constructing groups with a crystalline structure from scattered robots based on local information obtained from contact between robots.

The remainder of this paper is structured as follows. Section II defines the configuration of robots and robot swarms and sets the problem of the self-organizing task. Sections III and IV describe, respectively, a behavioral model for robots based on state transition using the phase-field method and a layer-forming method based on cellular automata under several lattice conditions and with different coverages. In Section V, we describe the results of a robot swarm self-organization simulation based on the proposed method and quantitatively verify that the robot swarm can construct lattices with layers at arbitrary scales. Finally, Section VI concludes the paper.

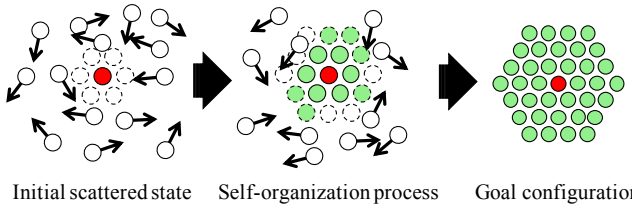


Fig. 1. Self-organizing a Regularly Arranged Group from Scattered Robots.

II. Problem Statement

In this study, we considered a swarm comprising N-unit homogeneous mobile robots. Each robot is equipped with a ranging sensor and local wireless communicator and can move in any direction on a two-dimensional plane within an upper velocity limit of v [m/s]. Each mounted sensor and communicator interfaces with the sensors and communicators on other robots and can be used to observe obstacles directly on lattices tiled on the two-dimensional plane, as shown in Fig. 2. To prevent errors in measuring the distances between robots, the robots cluster into groups with circular perimeters of diameter σ [m]. Under this condition, we define r_{ij} as the relative distance vector between the i-th robot and the j-th neighboring robot it observes. Each robot updates its velocity control value and communicable state based on the relative distance vectors and the exchanged state for a given interval using asynchronous timing. The robots move according to the calculated velocity control values.

For constructing a group of robots to navigate various lattices from randomly arranged states on a two-dimensional plane, a landmark robot is designated to collect other robots around an observed target. The other robots find the landmark robot through environment exploration and converge on positions that enlarge the group uniformly. The resulting organized group is a regular structure based on the geometrical regularities of honeycomb, square, or hexagonal lattices that can tile a two-dimensional plane at equally spaced intervals. In this process, if the robots are not oriented so that they face in the same direction, the group cannot construct the lattice recursively; therefore, the angular references of the robots are all assumed to be aligned along the same direction.

III. Behavioral Control of Robot

To carry out the proposed self-organization process, the robots must utilize the following functionalities: environmental exploration to move individually, surface exploration to find

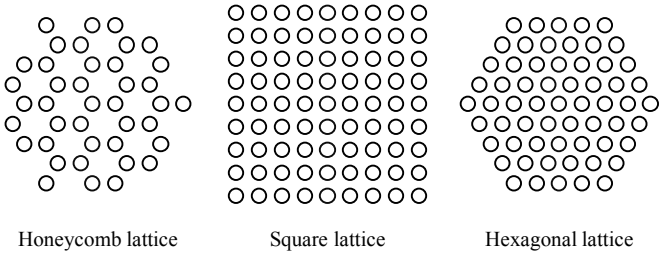


Fig. 2. Structures of Tile-able Lattices on a Two-dimensional Plane.

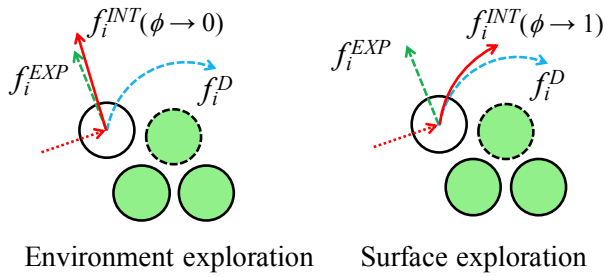


Fig. 3. Illustration of Output Behavioral Vector According to the Order Parameter Value when (left) ϕ is Close to Zero and (right) ϕ is Close to One.

adherable sites on the organized group, and maintenance of the lattice. In addition, the magnitudes of these controls must be altered according to the individual robot's progress. To achieve these behavioral transitions, we apply a phase-field method to represent the state probability between the solid and liquid phases.

The phase-field method involves the application of a reaction-diffusion equation that combines diffusion based on the state difference between a particle and its neighbors with a reaction based on the particle's state to formulate the path of surface movement in crystal growth. The following simple surface movement model is used:

$$\frac{d\phi}{dt} = \nabla^2\phi + 8\phi(1-\phi)(\phi - 0.5 + \beta) \tag{1}$$

where ϕ is an order parameter denoting the phase of the particle, which is a continuum value between zero (liquid phase) and one (solid phase), and β is the reaction rate, which is a constant parameter that solidifies under the condition $\beta > 0$.

When a robot exchanges order parameter information with other organized robots, it updates its internal state as reflected by this equation to proceed with its solidification and transition, depending on its order parameter value, from environment exploration via object-reflecting to surface exploration via edge-following, as shown in Fig. 3. Through this solidification process, the robots exploring an environment find sites of adherence on the organized group.

Each robot recognizes other robots that have been organized based on the construction state of the lattice, s_{j0}, which is obtained from the cellular automata produced by the j-th neighboring robot. If $s_{j0} > 0$, the neighboring robots are recognized as an organized group, with the set of recognized organized robots defined as CS_i, which is used to calculate the output of the phase-field method (in the next section, we will explain how the detailed cellular automata states are derived).

The order parameter that represents the state between the solid and liquid phases controls the i-th robot's behavior. When this parameter contacts the j-th organized robots, it is updated according to (1) as follows:

$$\phi_i \leftarrow \phi_i + \left(\frac{\sum \phi_j - \phi_i}{|CS_i|} + 8\phi_i(1-\phi_i)(\phi_i - 0.5 + \beta) \right) \Delta t \tag{2}$$

where Δt is the interval of the algorithm. According to this order parameter, the controller that decides the priority of environment and surface explorations is represented as follows:

$$f_i^{INT}(\phi_i) = \begin{cases} (1-\phi_i)\,\hat{f}_i^{EXP} + \phi_i f_i^D & ; CS_i \neq \emptyset \\ \hat{f}_i^{EXP} & ; otherwise \end{cases}. \tag{3}$$

The order parameter transitions from zero to one when the crystal growth produces a solidified state. This model prioritizes surface exploration, f_i^D, when close to the solidification state and environment exploration, f_i^{EXP}, otherwise. Note that a robot that does not yet neighbor an organized group uses only environment exploration to find a group. These environment and surface exploration behaviors must involve interactive reflection to avoid other robots and obstacles and edge-following drift of the outline of the organized group; these behaviors are described as follows.

The avoidance behavior used in environmental exploration is generated by constructing a reflection vector between the pre-behavioral vector and the point of collision with a target as follows:

$$f_i^{EXP} = \begin{cases} \sum\limits_{j \in TS_i} f_i^{INT} - 2\left(f_i^{INT} \cdot \hat{r}_{ij}\right)\hat{r}_{ij} & ; TS_i \neq \emptyset \\ f_i^{INT} & ; otherwise \end{cases} \tag{4}$$

where TS_i is the set of reflected objects, which contains obstacles that can be contacted by the robot and neighboring robots within a balanced potential distance, r_0. To compare the magnitudes of environment and surface exploration behavior, the calculated reflection vector is normalized and integrated into the behavioral vector.

The drift behavior for surface exploration used to search for a position of adherence to an organized group is described using a vector that rotates around the neighboring robot group while maintaining the potential, i.e., the distance needed to construct the lattice. It is calculated as follows:

$$f_i^D = f_i^P - \begin{cases} R\left(+\frac{\pi}{2}\right)\hat{r}_{id} & ; |f_i^{INT} \times r_{id}| \geq 0 \\ R\left(-\frac{\pi}{2}\right)\hat{r}_{id} & ; |f_i^{INT} \times r_{id}| < 0 \end{cases} \tag{5}$$

where the first and second terms on the right-hand side are the potential functions used to maintain a constant distance between robots as an surface and the rotation surrounding the neighbor closest to the pre-behavioral vector among organized robots, $d = \forall \arg\max\left(f_i^{INT} \cdot r_{ij} | j \in CS_i\right)$, respectively. The boundary conditions of the model determine the direction of this rotation along the pre-behavioral vector. By combining these, the robot can explore around an organized group at a certain distance.

The inter-robot potential works not only for drift but also for positioning. This potential generates attraction and repulsion forces that maintain the distance between robots needed to construct a group based on the lattice. As this approach focuses on a particle system, we incorporate the Lennard-Jones potential [21] as the distance potential and a simple sinusoidal potential as an angular potential, which is represented as follows:

$$f_i^P = \sum_{j \in CS_i} \frac{1+\cos(L(\theta_{ij}+l_i\pi))}{2|CS_i|} f^{LJ}(r_{ij}) + \sin(2L\theta_{ij}) R\left(\frac{\pi}{2}\right)\hat{r}_{ij} \quad (6)$$

where L is the number of neighboring robots depending on the target lattice. f^{LJ} is the Lennard-Jones potential adjusted for the attraction, as shown follows:

$$f^{LJ}(r) = \frac{169\sqrt[6]{\frac{13}{7}}}{63\cdot 2^{\frac{5}{6}}}\frac{1}{r}\left(12\left(\frac{\sigma}{r}\right)^{12} - 6\left(\frac{\sigma}{r}\right)^{6}\right). \quad (7)$$

This potential outputs a maximum attraction value of one to maintain a constant balanced distance r_0 (where $f^{LJ}(r_0)=0$) that depends on the diameter of the robot. In addition to this potential, the robot is attracted in a direction that satisfies the geometric regularity of the lattice according to the period of the angular potential. By these interactions, the robot moves to the position that places the the neighbors in the direction according to the regularity of the lattice. If the robot does not maintain the lattice-based structure, such as during drift, L is given zero to ensure that the angular potential does not work.

The i-th robot moves according to the movement vector, $v_i = vf_i^{INT}$, based on the behavioral vector calculated using these equations.

IV. Enlarging Algorithm

According to the behavioral model described in the previous section, a robot can find an adherence position suitable for further construction of the lattice that forms the group. We propose a ruleset of cellular automata to provide this position to the robot and a self-organizing potential function to maintain its tiled position. Fig. 4 shows the neighborhoods of these lattices, which can exchange information with neighboring robots in the same layer (unfilled layer), and the neighboring relations for constructing the lattice. The transition function of the cellular automata for these neighborhoods is represented as follows:

$$s_{i0} \leftarrow \begin{cases} \delta(s_{i0}, s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5}, s_{i6}, s_{i7}, s_{i8}, s_{i9}, s_{i10}, s_{i11}, s_{i12}) & ;\text{Honeycomb lattice} \\ \delta(s_{i0}, s_{i1}, s_{i2}, s_{i3}, s_{i4}) & ;\text{Square lattice} \\ \delta(s_{i0}, s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5}, s_{i6}) & ;\text{Hexagonal lattice} \end{cases} \quad (8)$$

where s_{i0} is the lattice construction state and s_{i1} to s_{iN} are the neighbor states, which depend on the number of neighbors in the lattice, with N = 12, 4, and 6 denoting a honeycomb, square, and hexagonal lattice, respectively. Because each lattice has rotational symmetry around its central robot, the surrounding states can be defined by any neighboring robot [22].

To build a group, a set of transition functions, called a ruleset, that repeatedly fills and enlarges the layers of the group by counting the number of filled corners in the outer layers is applied. The construction states s_{i0} to s_{iN} are therefore defined as continuous natural numbers from zero, indicating that the liquid phase searches for a position of adherence of to up to S_{max} and the solid phase fills the layers. S_{max} is at least eight, six, and eight for honeycomb, square, and hexagonal lattices, respectively, with adhered and layer-filled states included in the number of corners of the layer formed by each lattice.

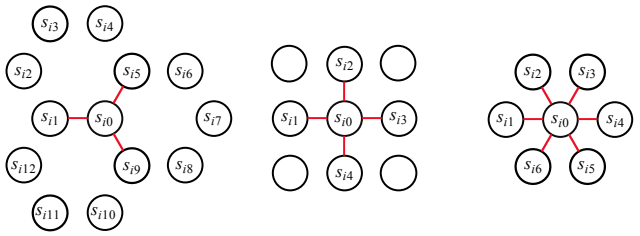


Fig. 4. Directly Observable Neighbors and Coupling Neighbors in Lattices.

Counting and sharing the values in the outer layers requires the neighbor values of the left and right neighbors, s_{il} and s_{ir}, respectively, in the layer inhabited by the i-th robot. The left and right reference directions of a robot can be toward either the inner or outer layers; therefore, to recognize its neighbors within its layer, the i-th robot calculates the layer l_i at the current position based on the layer l_j of the neighboring robots as $l_i = \min(l_j | j \in LS_i)+1$. For reference, the layer of the landmark robot is set to zero. LS_i, the set of neighboring robots within the range $\sqrt{2}r_0$, can be used to observe the robots in the inner layer from a constructible position in the outer layer. Consequently, the i-th robot obtains the neighbors from the layer adjacent to the inner layer as s_{il} and s_{ir}, respectively. The ruleset for synchronously enlarging the layers of a lattice based on these neighborhoods is then

$$s_{i0} \leftarrow \begin{cases} 1 & ;s_{i0}=0,\ \exists S_{max} \in s_{in} | 1 \le n \le N \\ 1 & ;s_{i0}=0,\ 0<s_{ir}<S_{max} \text{ or } 0<s_{il}<S_{max} \\ s_{ir}+1 & ;0<s_{i0}, s_{ir}, s_{il} < S_{max} \text{ and } \theta_{ril} < \frac{N-0.5}{2N}\pi \\ s_{ir} & ;0<s_{i0}, s_{ir}, s_{il} \end{cases} \quad (9)$$

This ruleset is constructed from the top two enlarging rules and the bottom two filling rules that produce the process shown in Fig. 5. By applying these enlarging rules, robots that are adjacent to either a filled inner layer or a neighborhood in the same layer will converge to that position and be organized into the group. The robots applied this rule complete the exploration by (3) to construct the group, and apply the potential model according to the number of neighbors of the organized lattice. The robots in the outer layer then count the number of corners in that layer based on the filling rules. If the angle between s_{il} and s_{ir} is less than π rad, the corresponding robot is in a corner and the adjacent robots are located on the sides. The robots identified as being in corners then transition to a state that adds one to s_{ir}, whereas the others continue sharing s_{ir}. By repeating this process, the robots in the outer layer reach their maximum state, S_{max}, allowing the robot swarm to synchronously enlarge the layers of the organized group.

V. Evaluation of Synchronistic Enlargement of Robot Aggregation

Using the proposed self-organization method, we confirmed that the self-organization process is integrated and that scattered robots can synchronously enlarge the layers of a lattice-based structure to self-organize a group. To this end,

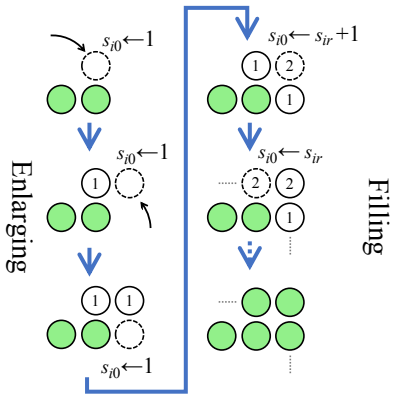


Fig. 5. Filling Process of the Robots in the Outer Layer by the Proposed Ruleset. The Robots Construct a New Layer on the Outside of the Green-colored Robots.

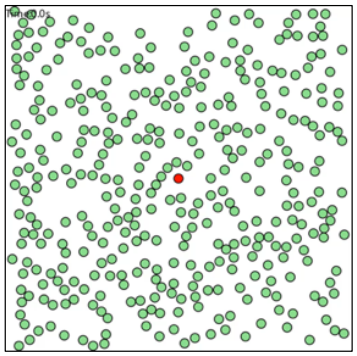


Fig. 6. Example of Field Generated by Simulation, with Green Robots Self-organizing Around a Red Landmark Robot.

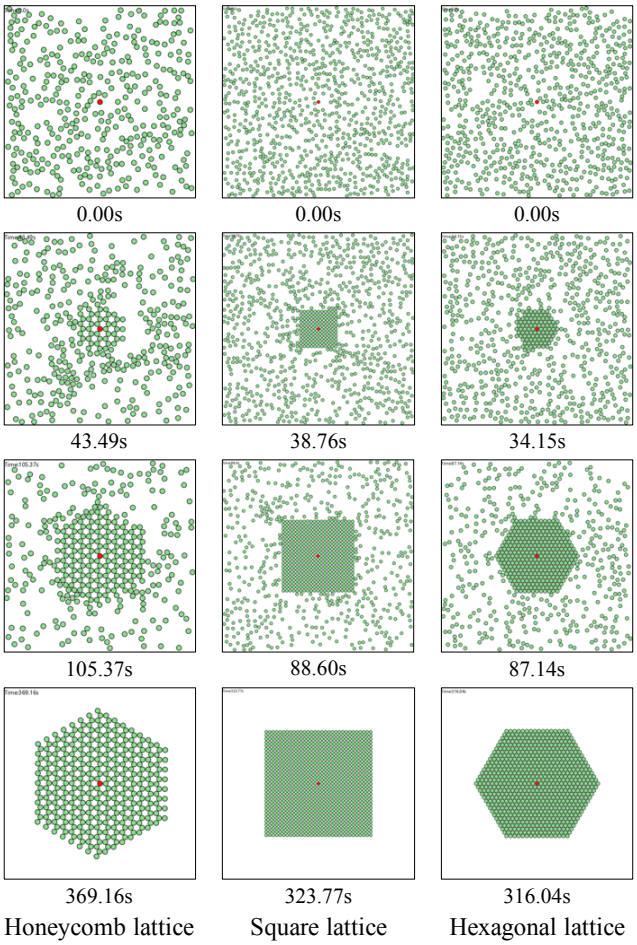


Fig. 7. Group Enlargements up to Self-organization of 15-layer Robot Groups on Different Lattice Types.

we simulated the process of building self-organizing lattice-based groups based on honeycomb, square, and hexagonal lattices. We also evaluated the geometrical regularities of the organized groups relative to their collective centers. Transportation requires a robot swarm comprising several tens to several hundred robots, depending on the size, weight, and shape of a transported object. In our simulation, we confirmed that the proposed method can be used to direct robots in the construction of a group to perform transporting tasks for large objects and that these groups can arrange the robots regularly to distribute the load for the weight of the transported objects.

We evaluated the dynamics of mobile robots under the following conditions: the robots were generated at random positions within a square region based on the average exploring range of an individual robot, with the landmark robot placed at the center of the region, as shown in Fig. 6. Note that this region size affects the search time of the robots. Each initialized robot had a diameter of 20 pixels, could move in all directions within an upper-velocity limit of 50 pixel/s, and was able to begin moving in a random direction. The robots also updated their behavioral vectors and communicable information computed using the proposed method with a reaction coefficient β of 0.1 at intervals of 10 ms. We simulated the movement of the robots for each lattice and layer condition.

A. Appearance of the Self-organization Process

To verify that a robot swarm could synchronously enlarge a layer based on the proposed cellular automata ruleset, we simulated self-organization of 15-layer groups on honeycomb, square, and hexagonal lattices using the required number of robots (361, 481, and 721, respectively) for each case and setting the average exploration area to 40×40 pixel2. The self-organization results obtained under these conditions are shown in Fig. 7.

From top to bottom, the rows in the figure show the groups constructed at 0 (initial state), 5, 10, and 15 (completed state) layers, respectively. In each case, the robots, which had been initialized with individual behaviors, converged to regular positions on the applied lattice and enlarged the organized group in all directions around the landmark robot. The shapes of the in-progress and completed organized groups were, in general, similar, and there were no defects in the filled layers or overgrowth in the outermost layers of the organized groups. These results indicate that the layer-filling process based on the proposed cellular automata ruleset worked as designed, with all robot swarms applying the proposed method achieving the construction of ordered arrangements from scattered robots.

The robot behavioral paths overlap and increase depending

on the random reflection vectors of the other robots, therefore, this paper cannot compare and discuss their time requirements. In the enlargement process, a robot exploring its environment will have to move in the path with little overlap to find an organized group. We expect to improve by incorporating behavioral models, such as random walk [23] and Lévy flight [24], that can effectively explore the overall environment into the proposed method.

B. Convergence Location of Scalable Self-organization

We then confirmed that the self-organization by a scalable robot swarm can satisfy the geometric regularity of a lattice. To evaluate this, we compared the differences between the simulated collective centers and the coordination of the landmark robots. Here, by "collective center" we mean the geometric condition-dependent coordination of the landmark robot as the organized robots converge to their ideal positions. Based on this difference, we could evaluate the geometrical regularity of the self-coordination process. The evaluation index for an individual robot is given by

$$O_c = \frac{1}{100}\sum_{try=1}^{100}\sqrt{\left(\sum_{i=1}^{N}\frac{x_i - x_c}{N}\right)^2 + \left(\sum_{i=1}^{N}\frac{y_i - y_c}{N}\right)^2} \tag{10}$$

where x_c and y_c are the coordinates at which the landmark robot is initially placed, x_i and y_i are the coordinates of the i-th robot when self-organization is complete, and N is the number of robots required under each experimental condition. To account for the randomness of the sequence of enlargement and the robot behavioral paths, the index was measured 100 times and the average value and standard deviation were calculated.

The index was measured for the first- to twentieth-layer group of each lattice with the average exploring area set to 60×60 pixel2. Fig. 8 shows the measurement results for each lattice. In each case, the obtained difference was shorter than the distance at which the potential model converged over time, indicating that the differences satisfied geometrical regularity. In addition, the indices for the series of honeycomb lattices were larger than those for the other lattices because, in the honeycomb lattice, the supported range of the proposed angular potential has wider than those of the others, and half of the robots in the outermost layer only construct with one robot. The indices also increased and scattered as the number of robots neighboring only one robot with weak convergence forces and wide support ranges increased with the number of layers.

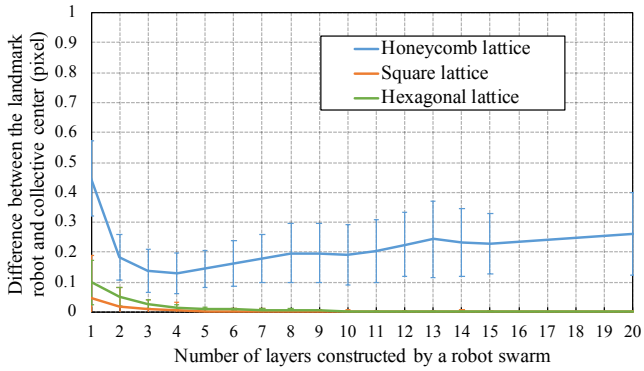


Fig. 8. Error in Measured Geometric Regularity as a Function of Number of Layers. The Solid Lines and Error Bars are the Average Values from 100 Trials and the Corresponding Standard Deviations, respectively.

VI. Conclusion

This paper proposed a self-organization method that integrates the self-organization process to utilize the cooperative capability of swarm robot systems. To achieve this, we focused on crystal growth and developed a distributed control/algorithm that combines a phase state transition based on the Phase-Field method and a group enlargement based on Cellular Automata. We demonstrated the self-organization process and evaluated the geometrical regularity of organized groups based on honeycomb, square, and hexagonal lattice-based arrangement conditions. From this result, we have achieved to produce a collective state from a parallel distributed state of a swarm robot system based on local robot positions and information exchange between the local robots.

Because the proposed method is a self-organizing approach based on the use of identical robots to fulfill a given lattice condition, it is limited by its inability to be used to construct flexible shapes (e.g., amorphous structures) based on long-range ordered coupling. Nevertheless, under the proposed method various group shapes can be represented by increasing the number of robots, i.e., the resolution. This will allow us to develop an approach for constructing designed group shapes by controlling the landmark robot and the layer growth speed. Furthermore, the proposed method has been shown to be effective at fixed-point observation; for the robots to engage in dynamic observation or the cooperative transportation of large objects, they will have to move in organized groups. Therefore, in the future, we will address the development of a dynamic approach in which the geometric conditions of the lattice are maintained.

Acknowledgment

This work was partially supported by the Research Institute for Science and Technology of Tokyo Denki University, grant number Q20D-08 / Japan. We would like to thank Editage (www.editage.com) for English language editing.

References

[1] D. Tarapore, R. Gros, and K-P. Zauner, "Sparse Robot Swarms: Moving Swarms to Real-World Applications," *Front. Robot. AI*, vol. 7, no. 83, pp. 1–9, 2020, DOI: 10.3389/frobt.2020.00083.

[2] D. Carrillo-Zapata, E. Milner,J. Hird, G. Tzoumas, P.J. Vardanega, M. Sooriyabandara, M. Giuliani, A.F.T. Winfield, and S. Hauert, "Mutual Shaping in Swarm Robotics: User Studies in Fire and Rescue, Storage Organization, and Bridge Inspection," *Front. Robot. AI*, vol. 7, no. 53, pp.1–19, 2020, DOI: 10.3389/frobt.2020.00053.

[3] S. Batra, J. Klingner and N. Correll, "Augmented Reality for Human-Swarm Interaction in a Swarm-Robotic Chemistry Simulation," in *SWARM 2021,* Online, 2021, pp. 250–263.

[4] Y. Zhou, B. Rao, and W. Wang, "UAV Swarm Intelligence: Recent Advances and Future Trends," *IEEE Access*, vol. 8, pp. 183856–183878, 2020, DOI: 10.1109/ACCESS.2020.3028865.

[5] M. Schranz, M. Umlauft, M. Sende, and W. Elmenreich, "Swarm Robotic Behaviors and Current Applications," *Front. Robot. AI*, vol. 7, no. 36, pp. 1–20, 2020, DOI: 10.3389/frobt.2020.00036.

[6] A. Jevtic, A. Gutierrez, D. Andina, and M. Jamshidi, "Distributed Bees Algorithm for Task Allocation in Swarm of Robots," *IEEE Syst. J.*, vol. 6, no. 2, pp. 296–304, 2012, DOI: 10.1109/JSYST.2011.2167820.

[7] H. Zhao, Z. Nie, and X. Wang, "Design and Analysis of Multi-robot Grouping Aggregation Algorithm," *J. Robot. Netw. Artif. Life*, vol. 6, no. 1, pp. 2352–6386, 2019, DOI: 10.2991/jrnal.k.190602.002.

[8] S. Yamashita, D. Kurabayashi, and Y. Hattori, "Autonomous Division and Integration of Anonymous Agents by Using Interaction between Oscillators," *Trans. JSME*(in Japanese), vol. 84, no. 857, pp. 1–14, 2018, DOI: 10.1299/transjsme.17-00338.

[9] A. Barci, and C. Bettstetter, "Sandsbots: Robots That Sync and Swarm," *IEEE Access*, vol. 8, pp. 218752–218764, 2020, DOI: 10.1109/ACCESS.2020.3041393.

[10] M. Rubenstein, A. Cornejo, and R. Nagpal, "Programmable Self-Assembly in a Thousand-Robot Swarm," *Science*, vol. 345, no. 6198, pp. 795–799, 2014, DOI: 10.1126/science.1254295.

[11] Y. Zhu, D. Bie, X. Wang, Y. Zhang, H. Jin, and J. Zhao, "A Distributed and Parallel Control Mechanism for Self-Reconfiguration of Modular Robots using L-systems and Cellular Automata," *J. Parallel Distrib. Comput.*, vol. 102, pp. 80–90, 2017, DOI: 10.1016/j.jpdc.2016.11.016.

[12] M. Alhafnawi, S. Hauert, and P. O'Dowd, "Self-Organised Saliency Detection and Representation in Robot Swarms," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1487–1494, 2021, DOI: 10.1109/LRA.2021.3057567.

[13] A. R. Shirazi and Y. Jin, "A Strategy for Self-Organized Coordinated Motion of a Swarm of Minimalist Robots," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 1, no. 5, pp. 326–338, 2017, DOI: 10.1109/TETCI.2017.2741505.

[14] D. Carrillo-Zapata, J. Sharpe, A.F.T. Winfield, L. Giuggioli, and S. Hauert, "Toward Controllable Morphogenesis in Large Robot Swarms," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3386–3393, 2019, DOI: 10.1109/LRA.2019.2926961.

[15] S. Nouyan, R. Gross, M. Bonani, F. Mondada, and M. Dorigo, "Teamwork in Self-Organized Robot Colonies," *IEEE Trans. Evol. Comput.*, vol. 13, no. 4, pp. 695–711, 2009, DOI: 10.1109/TEVC.2008.2011746.

[16] R. Zongwei and Z. Yanhe, "Self-Reconfiguration Planning of Self-Reconfigurable Robot based on Windmill-like Crystal Cell Group," in *MSIE 2011,* Harbin, CN, 2011, pp. 1218–1222, DOI: 10.1109/MSIE.2011.5707641.

[17] P. Swissler and M. Rubenstein, "FireAnt3D: A 3D Self-Climbing Robot Towards Non-Latticed Robotic Self-Assembly," in *IROS 2020,* Las Vegas, NV, USA, 2020, pp. 3340–3347.

[18] G. Lee and N.Y. Chong, "A Geometric Approach to Deploying Robot Swarms," *Ann. Math. Artif. Intell.*, vol. 52, pp. 257–280, 2008, DOI: 10.1007/s10472-009-9125-x.

[19] H. Yang, S. Cao, L. Bai, Z. Zhang, and J. Kong, "A Distributed and Parallel Self-Assembly Approach for Swarm Robotics," *Robot. Auton. Syst.*, vol. 118, pp. 80–92, 2019, DOI: 10.1016/j.robot.2019.04.011.

[20] A.V. Redkov, S.A. Kukushkin, and A.V. Osipov, "Spiral Growth of A Crystal Due to Chemical Reaction," *J. Phys.: Conf. Ser.*, vol. 1124, pp.1–6, 2018, DOI:10.1088/1742-6596/1124/2/022006.

[21] N. Yu and A.A. Polycarpou, "Adhesive Contact Based on The Lennard–Jones Potential: A Correction to The Value of The Equilibrium Distance as Used in The Potential," *J Colloid Interface Sci.*, vol. 278, pp.428–435, 2004, DOI:10.1016/j.jcis.2004.06.029.

[22] N.H. Packard, "Lattice Models for Solidification and Aggregation," in *Science on Form,* Tsukuba, JP, 1986, pp. 95–100.

[23] R.J. Alitappeh and K. Jeddisaravi, " Multi-Robot Exploration in Task Allocation Problem," *Appl Intell*, 2021, DOI: 10.1007/s10489-021-02483-3.

[24] Y. Katada, A. Nishiguchi, K. Moriwaki, and R. Watakabe, "Swarm Robotic Network Using Levy Flight in Target Detection Problem," *Artif. Life Robot.*, vol. 21, pp. 295–301, 2016, DOI: 10.1007/s10015-016-0298-1.

Secure and Efficient Proof of Ownership Scheme for Client-Side Deduplication in Cloud Environments

Amer Al-Amer[1], Osama Ouda*[1,2]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Saudi Arabia[1]
Department of Information Technology, Faculty of Computers and Information, Mansoura University, Egypt[2]

***Abstract*—Data deduplication is an effective mechanism that reduces the required storage space of cloud storage servers by avoiding storing several copies of the same data. In contrast with server-side deduplication, client-side deduplication can not only save storage space but also reduce network bandwidth. Client-side deduplication schemes, however, might suffer from serious security threats. For instance, an adversary can spoof the server and gain access to a file he/she does not possess by claiming that she/he owns it. In order to thwart such a threat, the concept of proof-of-ownership (PoW) has been introduced. The security of the existing PoW scheme cannot be assured without affecting the computational complexity of the client-side deduplication. This paper proposes a secure and efficient PoW scheme for client-side deduplication in cloud environments with minimal computational overhead. The proposed scheme utilizes convergent encryption to encrypt a sufficiently large block specified by the server to challenge the client that claims possession of the file requested to be uploaded. To ensure that the client owns the entire file contents, and hence resist collusion attacks, the server challenges the client by requesting him to split the file he asks to upload into fixed-sized blocks and then encrypting a randomly chosen block using a key formed from extracting one bit at a specified location in all other blocks. This ensures a significant reduction in the communication overhead between the server and client. Computational complexity analysis and experimental results demonstrate that the proposed PoW scheme outperforms state-of-the-art PoW techniques.**

***Keywords*—Client-side deduplication; proof of ownership; convergent encryption; could storage services**

I. Introduction

Cloud computing is the provision of on-demand access to different computing services and resources, such as storage space, servers, networks, databases, and software, over the internet [1]. The different models of cloud computing can provide a wide range of capabilities, adapted with different business goals, to diverse clients and/or consumers [2, 3]. The rapid development and integration of cloud computing have led organizations, institutions, and individuals to increasingly turn to utilize services provided over the cloud [4]. Consequently, an increasing number of individuals and organizations tend to move their data on cloud storage services (e.g., Dropbox, SkyDrive, Google Drive, iCloud, Amazon S3). This resulted in rapid growth in the volume of data that is stored on the cloud storage servers [5, 6, 7].

To increase efficiency as well as reduce the storage space required on storage servers, cloud storage providers tend to avoid downloading and uploading duplicated data [5, 8]. Data deduplication is an effective mechanism that aims at reducing the required storage space of the cloud storage servers by avoiding storing several copies of the same data. There are two main types of deduplication [9, 10] namely, server-side deduplication and client-side deduplication. The server-side deduplication schemes [11, 12] remove duplicated copies of the same files after uploading them to the server. On the other hand, In client-side deduplication [10, 13], duplicated copies are identified on the client side and not uploaded to the server. Hence, in contrary to server-side deduplication, client-side deduplication can not only save storage space and uploading time but also reduce network bandwidth.

However, client-side deduplication schemes might suffer from serious security threats [13, 14, 15]. For instance, an adversary can spoof the server and gain access to a file that he/she does not possess by claiming that he/she owns it. To thwart such a threat, Halevi et al. [16] proposed a cryptographic primitive, referred to as "proof of ownership" (PoW), to allow the server to verify whether a client owns a file. They pointed out that a robust PoW scheme should alleviate potential security threats without introducing I/O and computational overhead at both client and server sides. Since the introduction of the proof-of-ownership concept, several PoW schemes have been proposed in the past few years [14, 16, 17, 18, 19, 20, 21, 22]. However, the security of such schemes cannot be assured without affecting the computational complexity of client-side deduplication.

An efficient PoW scheme should satisfy several requirements. First, the chances that an adversary successfully passes a PoW run should be negligible if the adversary does not possess the file in its entirety. Second, a small fixed amount of information should be loaded on the server-side regardless of the file size. Third, the amount of processed information on both the client and server sides should be minimal. In addition, the amount of transmitted data between the client and the server should be reduced to minimize the bandwidth. Unfortunately, the PoW schemes discussed above do not fulfill all of these requirements. Thus, new techniques that can resolve the security-efficiency trade-off and reduce communication and storage overhead should be proposed.

This paper proposes a secure and efficient proof-of-ownership scheme for client-side deduplication in cloud environments that fulfills the requirements mentioned above. The proposed technique utilizes convergent encryption to encrypt a sufficiently large block specified by the server to challenge the client that claims possession of the file requested to be uploaded. To ensure that the client owns the entire file contents, and hence resists collusion attacks [23, 24], The server challenges the client by requesting him to split the file he/she asks to upload into fixed-sized blocks and then encrypts

a randomly chosen block using a key formed from extracting one bit at a specified location in all other blocks. This ensures a significant reduction in the communication overhead between the server and client. Moreover, the proposed scheme resists attacks of honest-but-curious servers [25, 26, 27].

The rest of the paper is structured as follows. Section II provides a brief background on the concepts of data deduplication, convergent encryption, and proof-of-ownership. Section III describes the proposed PoW scheme in detail. Section IV presents a computational complexity analysis of the proposed scheme and provides a comparison with the state-of-the-art schemes. Section V describes the experimental results and discussion. Finally, Section VI concludes the paper.

II. Background

A. Data Deduplication

Data deduplication, also called single-instance storage, techniques aim to eliminate duplicate copies of the same data to improve storage utilization and reduce the unnecessary cost of storage capacity needs [13]. A prominent example of the usefulness of data deduplication is redundant file attachments in email systems. Consider a typical email system containing 50 copies of the same 20 megabyte (MB) file attachment. Saving or archiving this email platform requires 1000 MB of storage space. The storage demand can drop to only 20 MB if data deduplication is employed. The example shown in Fig. 1 demonstrates the main concept of data deduplication. There are two main types of data deduplication in cloud environments: server-side deduplication and client-side deduplication. Server-side deduplication techniques identify repeated data after uploading them to the server, whereas client-side deduplication techniques identify duplicate copies of data before they are uploaded to the server. Therefore, client-side deduplication techniques can reduce network data transfers in addition to storage capacity.

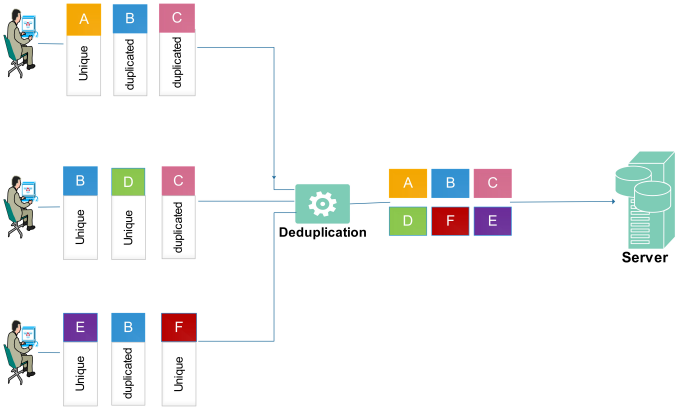


Fig. 1. Data Deduplication Technology.

B. Convergent Encryption

Data deduplication techniques can benefit from convergent encryption (CE) to achieve security smoothly and more easily [10, 15]. Convergent encryption is a cryptographic concept that ensures security in the cloud by achieving confidentiality and data privacy. The main idea behind CE is to create similar ciphertexts from similar plaintexts (see Fig. 2). Unlike traditional cryptography, in which data encryption and decryption are carried out using cryptographic keys that are independent of the data being encrypted, and hence different ciphertexts are obtained from the same plaintexts, CE ensures that the same key is used for the same plaintexts. In CE, the data digest or hash is used as a key to encrypt the data. Encrypting data in CE undergoes the following three steps: 1) the digest (hash value) of the plaintext in question is computed, 2) the plaintext is then encrypted using its digest as a key, and 3) finally, the hash is encrypted with a key chosen by the user and stored along with the obtained ciphertext. These steps ensure that identical data copies will generate the same key and the same ciphertext.

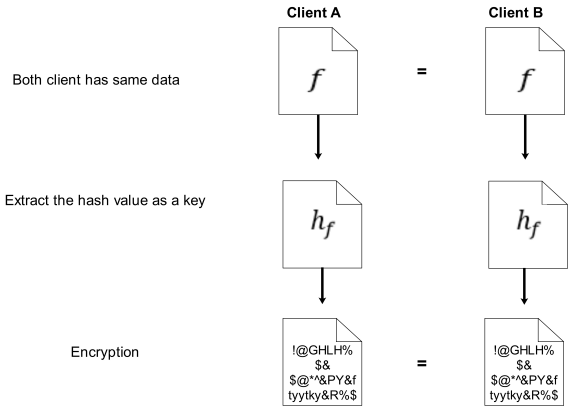


Fig. 2. Convergent Encryption Concept.

C. Proof of Ownership (PoW)

In client-side deduplication techniques, the hash value of the file requested to be uploaded by the client is firstly computed and sent to the server. If this hash value exists in the list of previously uploaded files to the server, the server will request the client not to upload the file again to avoid storing redundant data. However, in order to append the client to the list of owners of that file, the server has to verify that the client owns the entire file and not try to spoof it. Traditional proof of ownership protocols, Such as the one proposed by Halevy et al. [5] oud storage provider (CSP) has access to the original file. In other words, such protocols depend on trust between the cloud storage provider and the client. However, this trust might generate many potential security risks since cloud storage providers (CSPs) should not be fully trusted. The process of adapting PoW protocols so that they can work properly on encrypted data is an open problem so far [13].

Several PoW schemes have been proposed over the past few years. Gonzalez-Manzano et al. [14] proposed an attribute-based symmetric encryption proof of ownership scheme, referred to as ase-PoW, for hierarchical environments. The main goals of this scheme are to resist honest-but-curious servers and to provide flexible access control to ensure that users have access to sensitive files with the right and real privileges. The

idea behind this scheme is based on recursively encrypting parts of the file being uploaded to the server to assure its possession by the user. The ase-PoW scheme has the advantage of its ability to resist guessing attacks on the content and reduce the cloud workload. However, it does not take into account the issues of user revocation and key updating.

Dave et al. [17] proposed a secure proof of ownership scheme based on utilizing Merkle trees. The idea is based on calculating the responses of challenges in advance at the server-side to avoid computational overhead while uploading the file. The cloud server does not need to hold over the resources until the response is received, which is preferable to the utilization of stateless protocols. The user on the client-side is requested to encrypt the file to be uploaded to the server using its digest as a key (a.k.a. convergent encryption (CE)). Afterward, the user computes a file tag (usually a cipher-text hash) to check the file's existence on the server. The file uploading process consists of four stages (metadata generation, challenge generation, response generation, and response verification). If the file tag does not exist in the FileList kept by the server, the client will be requested to upload the file with the tag. In the case of subsequent uploads, the server sends an unused precomputed PoW challenge to the client. If the client owns the entire file, then the client will correctly respond to the server.

Islam et al. [18] proposed a secure and reliable storage scheme for cloud environments with client-side deduplication (SecReS). The authors combined convergent encryption and secret sharing techniques to achieve data confidentiality. They used the Reed-Solomon erasure code to achieve fault tolerance through distributing data to multiple storage servers. Moreover, Merkle trees are utilized to verify the ownership of data and to perform secure data deduplication. Mishra et al. [19] proposed a merged PoW scheme (MPoWS) for block-level deduplication in cloud storage. By employing a random test approach, MPoWS meets the requirements of client-side and server-side mutual verification. In MPoWS, large files are uploaded to the servers, and then their duplication is checked using various blocking flags. The authors used a random test approach to increase security and make it difficult to predict which block will be validated.

Fan et al. [20] proposed a secure deduplication scheme based on a trusted execution environment (TEE), which provides secure key management by using convergent encryption with cloud users' privileges. Trusted execution environments improve cryptographic systems' capability to resist chosen-plaintext and chosen-ciphertext attacks. The authors proposed assigning a set of privileges to every cloud user. Therefore, data deduplication can be performed if and only if the user of the cloud has the right and valid privileges. In [21], Ouda proposed a secure and effective proof of ownership scheme for client-side deduplication in the cloud. This scheme verifies if the client owns the entire file for which he/she claims possession. In other words, the proposed scheme does not allow an adversary to engage in a successful proof of ownership without fully owning the file's content. This can be achieved by requesting the client to encrypt the entire file using the file hash as the key before uploading the file to the server. This prevents the curious server from disclosing the file content.

Du et al. [22] proposed a proof of ownership and retrievability framework (PoOR) in which the cloud client can prove ownership of files to the server without uploading or downloading the files. The proposed framework consists of the pre-processing two phase, proof of ownership phase, and retrievability phase. The proof of ownership phase depends on the Merkle Tree protocol and comprises three steps: prove, challenge, and verify. Cui et al. [28] proposed a new attribute-based storage system that supports secure and efficient deduplication. The proposed system runs on a hybrid cloud environment where the private cloud is responsible for detecting identical copies for storage management. Ma et al. [29] demonstrated how attribute-based encryption can be used to minimize storage space and share data efficiently. In this technique, if the attributes of certain user is matched, then the user is given the right to decipher the encrypted data.

Blasco et al. [30] proposed a PoW scheme, called bf-PoW, that utilizes the Bloom filters to mitigate the server-side overhead. The main drawback of the bf-PoW scheme is that it does not consider data privacy. Di Pietro and Sorniotti [9] introduced another scheme, referred to as s-PoW in which the server requests clients to send bit-values of randomly selected bit positions of files requested to be uploaded by those clients. Although this scheme is computationally efficient at the client-side, it is not efficient on the server-side. Manzano and Orfila [31] proposed a PoW scheme, called ce-PoW, employing the concept of convergent encryption to encrypt file chunks before uploading them to the server. This scheme reduces issues related to key management. However, since the encryption is applied at the chunk level, the number of encryption keys increases linearly with the number of requested chunks. This can put a significant burden on both storage space and bandwidth as the security parameters increase.

Huang et al. [32] proposed a bidirectional and malleable proof-of-ownership scheme for large files in cloud storage (BM-PoW). The proposed BM-PoW protocol allows the server and user to interact to ensure ownership of the file to be uploaded even if the file is updated. Thus, secure and efficient deduplication for large files in static and dynamic archives is guaranteed. Miao et al. [33] proposed a novel PoW protocol that benefits from the distinguishable properties of chameleon hashing. Although this protocol is more efficient than existing PoWs based on Merkle hash tree, it is vulnerable to brute-force attack (BFA) due to its limited keyspace [34].

Although some solutions have been proposed to improve the efficiency at the server-side, other solutions tend to enhance the computational cost at the client-side. Besides, most of the existing schemes cannot satisfy all the security requirements in terms of resisting honest-but-curious servers and collusion attacks without affecting the efficiency and/or communication bandwidth requirements. Therefore, it is promising to study how to develop PoW schemes that can balance the trade-off between the security and efficiency requirements.

III. Proposed PoW Scheme

As we previously mentioned, the main goal of our proposed PoW scheme is to provide an efficient means to prove the ownership of files in client-side deduplication environments securely. Precisely, we aim at minimizing the exchanged

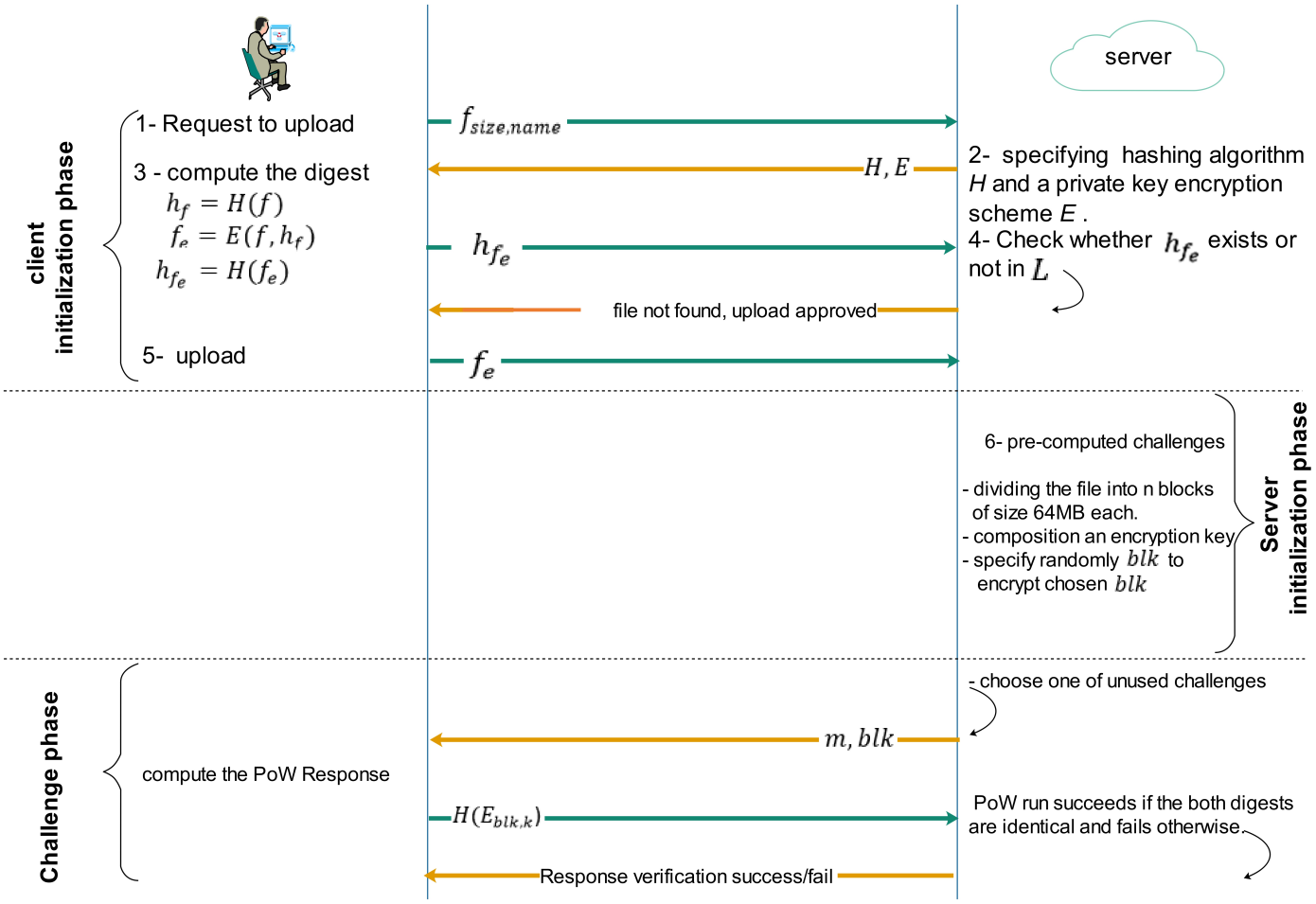


Fig. 3. Proposed PoW Scheme.

information between the client and server, reducing the data uploaded in memory during a PoW session, and decreasing the likelihood that a malicious client can successfully respond to the challenge sent to him/her by the server via increasing the amount of exchanged information between a malicious client and a legitimate owner of the file. Definition of abbreviations and symbols are given in Table I.

As illustrated in Fig. 3, the proposed PoW scheme consists of three phases: the client initialization phase, server initialization phase, and challenge phase. In the client initialization phase (see Algorithm 1), the client initiates a file (f) upload request to the server by simply sending general attributes of the file, such as its name and size. The server responds to the client by sending a message specifying a robust hashing algorithm $\mathcal{H}$ (e.g., MD5, SHA1, etc.) as well as a private key encryption scheme E (e.g., DES, AES, etc.). The client uses the specified hashing algorithm to compute the file digest, $h_f = \mathcal{H}(f)$, which is then used as a key to encrypt the file using the encryption algorithm specified by the server to obtain an encrypted file $f_e = E(f, h_f)$. The encrypted file f_e is then hashed using $\mathcal{H}$ to obtain its digest $h_{f_e} = \mathcal{H}(f_e)$. Finally, the client sends h_{f_e} to the server so that it can decide whether the file has been uploaded before by a different user.

TABLE I. DEFINITION OF ABBREVIATIONS AND SYMBOLS

Abbreviation	Definition
f	File to be uploaded to the server
f_e	Encrypted file
$\mathcal{H}(f)$	Hash of the file f
blk	A block of f
j	Block number
n	Number of non-overlapping blocks
κ	Encryption key
m	Bit position
$\mathcal{L}$	List of uploaded files
$\mathcal{C}$	The client
$\mathcal{S}$	The server

Assuming that the server stores the digest of each previously uploaded encrypted file, the server can decide whether f has been uploaded before by searching for h_{f_e} in the list $(\mathcal{L})$ of the stored digests. It is worth noting that our scheme

Algorithm 1 Client Initialization Phase

Input: File f
Output: $\mathcal{H}(f_e)$, f_e

Client $\mathcal{C}$ sends a request to server $\mathcal{S}$ to upload the file f
$\mathcal{S}$ sends to $\mathcal{C}$ the name of the hashing algorithm $\mathcal{H}$ and encryption algorithm E
$\mathcal{C}$ computes the digest $\mathcal{H}(f)$ of f using $\mathcal{H}$
$f_e \leftarrow E(f, \mathcal{H}(f))$
$\mathcal{C}$ computes $h_{f_e} = \mathcal{H}(f_e)$ and sends it to $\mathcal{S}$
$\mathcal{S}$ searches for h_{f_e} in the list ($\mathcal{L}$) of uploaded files
if $\mathcal{H}(f_e)$ is found in $\mathcal{L}$ **then**
    Go to the Challenge Phase
else
    Allow $\mathcal{C}$ to upload f_e to $\mathcal{S}$
    Go to the Server Initialization Phase
end if

Algorithm 2 Server Initialization Phase

Input: File f_e
Output: The entry $\mathcal{L}[h_{f_e}]$

Divide f_e into n blocks of size 64MB each
for $i \leftarrow 1$ to c **do** /* c = No. of challenges */
    Choose two integers m and $j < n$
    Extract the m-th bit of each block in f_e
    Generate a cryptographic key κ by concatenating the n extracted bits
    Encrypt the j-th block in f_e using κ to obtain $E(blk_j, \kappa)$
    Compute the digest of $E(blk_j, \kappa)$
    $Challange[i] =< m, j, \mathcal{H}(E(blk_j, \kappa)) >$
end for
Create a new entry for f_e consisting of all the generated challenges and append it to $\mathcal{L}$

assumes that files are encrypted before uploading them to the server to prevent honest-but-curious servers from disclosing the contents of the uploaded files. If h_{f_e} is not found in the list of the stored digests, the server sends a message to the client to start uploading f_e; otherwise, the server initiates the challenge phase.

After the file f_e is uploaded, the server initialization phase (see Algorithm 2) starts by creating a new entry for f_e. This entry consists of h_{f_e} and a pointer to a set of pre-computed challenges that will be used to prove the ownership of f_e by other clients who might request to upload the same file in the future. A challenge is created by dividing the file into n non-overlapping blocks sufficiently large to resist collusion attacks. It is assumed that sharing data $\geq$ 64MB among colluders would discourage them launch collusion attacks [30]. Thus, we recommend setting the block size at such levels. Then, the encryption key (κ) is composed by concatenating all bits at a specific position (m) across all blocks. For instance, if $m = 3$ and $n = 128$, then the third bit in each block is extracted, and the set of the 128 extracted bits are concatenated to create a 128-bit cryptographic key κ. An example that illustrates the key generation process, where $m = 3$ and the file and block sizes are 8 GB and 64 MB, respectively, is shown in Fig. 4. The generated key is then used to encrypt a randomly chosen block (blk_j) of the encrypted file f_e using the AES

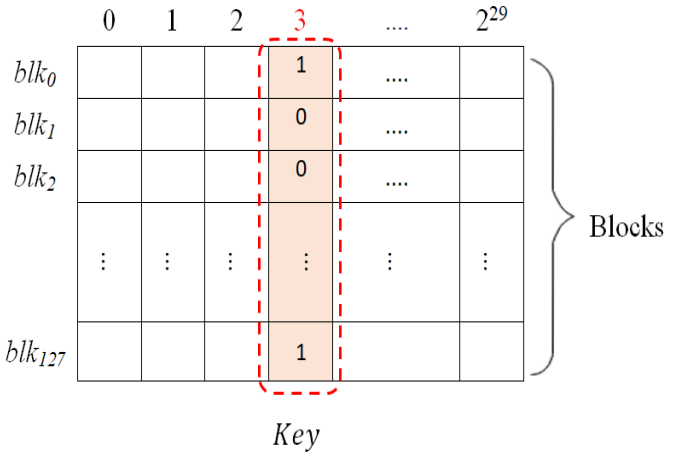


Fig. 4. An Example that Illustrates the Key Generation Process.

encryption algorithm to obtain $E(blk_j, \kappa)$. Finally, the digest of the encrypted block $\mathcal{H}(E(blk_j, \kappa))$ is obtained and stored along with m and j as one challenge for f_e. The previous challenge creation process is then repeated using different values of m and j to encrypt different blocks of f_e in order to generate as many challenges as needed.

The challenge phase is initiated when an entry is found for f_e in the list of uploaded files kept by the server. In this case, the server chooses one of the available challenges in $\mathcal{L}[h_{f_e}]$ to verify that the client possesses the file he/she requests to upload. The server challenges the client by sending m and j corresponding to the chosen challenge. Note that with just these two values, the amount of information sent from the server to the client is minimized. This satisfies an important design objective of the proposed PoW scheme. The client responds to the challenge by dividing the file, after encrypting it using the same procedure described in the client initialization phase, into n blocks, generating κ by extracting bits at position m of all blocks, encrypting the block blk_j specified by the server using κ, and finally calculating the digest of the encrypted block $\mathcal{H}(E(blk_j, \kappa))$ and sends it to the server as a response to the received challenge. The server matches the received block digest with the corresponding digest stored in $\mathcal{L}$. The PoW run succeeds if both digests are identical and fails otherwise.

It is important to note that all design goals of the proposed PoW scheme are satisfied. The data exchanged between the client and server are minimized. In the challenge phase, the server sends two small pieces of information; namely, the bit position index m used to generate the key κ and the index of the block to be encrypted. Similarly, the client is required to respond to the challenge received from the server by just sending the specified block digest. From the security perspective, the block size is set to 64MB because a PoW scheme is considered secure if the amount of exchanged information between a legitimate owner of f and a malicious client required to pass a PoW run is not smaller than 64MB [30]. Moreover, since one bit per block is used to generate the key (κ), a large number of challenges can be generated for each uploaded file. Precisely, more than 2^{29} different challenges can be generated if the block size is set to 64MB.

TABLE II. COMPARISON BETWEEN THE PROPOSED SCHEME AND FIVE RELATED POW SCHEMES CONCERNING SPACE AND COMPUTATIONAL COMPLEXITY. κ: SECURITY PARAMETER, n: NUMBER OF PRE-COMPUTED CHALLENGES, l: TOKEN LENGTH, F: FILE SIZE, B: BLOCK SIZE AND p_f: FALSE POSITIVE RATE (BF-POW SCHEME)

	ase-PoW	ouda-PoW	ce-PoW	bf-PoW	s-PoW	Proposed
Client computation	$O(B).Sym.n_r.hash$	$O(F).CE.hash$	$O(B).CE.hash.hash$	$O(F).hash$	$O(F).hash$	$O(B).CE.hash$
Server init computation	$O(B).hash$	$O(F).hash$	$O(B).hash.hash$	$O(F).hash$	$O(F)$	$O(B).hash$
Server init I/O	$O(F)$	$O(F)$	$O(F)$	$O(F)$	$O(F)$	$O(F)$
Server regular I/O	$O(0)$	$O(0)$	$O(0)$	$O(0)$	$O(n.k)$	$O(0)$
Server memory usage I/O	$O(n.l.k)$	$O(n.l)$	$O(n.l.k)$	$O(\frac{log(l/p_f)}{l})$	$O(n.k)$	$O(n.k)$
Bandwidth	$O(l.k)$	$O(l.k)$	$O(l.k)$	$O(\frac{l.k}{p_f})$	$O(k)$	$O(k)$

Algorithm 3 Challenge Phase

Input: m and j of an unused challenge
Output: PoW response

1: $\mathcal{S}$ sends m and j to $\mathcal{C}$
2: $\mathcal{C}$ divides f_e into n non-overlapping blocks of size 64MB each
3: $\mathcal{C}$ extract the m the bit of each block in f_e
4: $\mathcal{C}$ generate a cryptographic key κ by concatenating the n extracted bits
5: $\mathcal{C}$ encrypt the j the block in f_e using κ to obtain $E(blk_j, k)$
6: $\mathcal{C}$ send to the $\mathcal{S}$ the challenge $\mathcal{H}(E(blk, k))$
7: $\mathcal{S}$ reciving the challenge $\mathcal{H}(E(blk, k))$
8: **if** Clinet $\mathcal{H}(E(blk, k))$ = Server $\mathcal{H}(E(blk, k))$ **then**
9: PoW success
10: **else**
11: Fail to PoW
12: **end if**

IV. COMPLEXITY ANALYSIS

This section demonstrates how the proposed PoW scheme fulfills bandwidth and space efficiency requirements by comparing it to five well-known PoW schemes from the literature, as shown in Table II. Specifically, we compare the complexity of our proposed scheme by focusing on bandwidth, server memory usage, client/server computation, and I/O. It can be noticed from Table II that the complexity of the proposed scheme is similar to the complexity of the other schemes with respect to server initialization I/O as it primarily relies on the file size. For the regular server I/O, the complexity of the proposed scheme is also similar to the complexity of the other schemes except for s-PoW. Moreover, the complexity of the proposed scheme outperforms the complexity of the other schemes with respect to client computation and server initialization computation, mainly because the proposed scheme only encrypts a randomly chosen block in the file rather than encrypting the whole file. It is worth noting that this does not affect the security of the proposed scheme since the employed cryptographic key is extracted from all blocks of the file.

V. EXPERIMENTAL RESULTS

This section describes the experiments conducted to evaluate the performance of the proposed PoW scheme. All experiments were conducted on a personal computer with Intel Core i7-4770 CPU (2.4 GHz) and 8 GiB RAM. The performance of the proposed scheme was compared with the performance of the ce-PoW scheme proposed by Gonzalez-Manzano et al. [14], ase-PoW scheme proposed by Manzano and Orfila [31], Ouda-PoW scheme proposed by Ouda [21], bf-PoW scheme proposed by Blasco et al. [30], and s-PoW scheme proposed by Di Pietro and Sorniotti [9]. In all experiments, the schemes were evaluated using randomly generated test files of sizes ranging from 4MB to 2GB, doubling the size at each step. We used the C++ programming language for the implementation and utilized the OpenSSL cryptographic library [35] for the encryption and hashing operations, namely, AES (in counter mode) and SHA-256.

The clock cycles spent by the client to upload a file for the first time and to respond to the server challenge have been measured and compared with the corresponding clock cycles spent by the other tested PoW schemes. Fig. 5 shows the computational cost (clock cycles) spent in the client initialization phase for the four tested schemes. It can be noticed from the figures that the proposed PoW scheme outperforms all the other schemes. This is mainly because the s-PoW dealing with file level and the server requests from the clients to send bit-values of randomly selected bit positions of files requested to be uploaded. whereas in bf-PoW the clients compute a token for each segment index using the hash function, which in turn increments the executed operations. The ce-PoW scheme requires implementing multiple hashing and encryption operations separately for each file chunk. In ase-PoW, on the other hand, the client has to encrypt each part of the file chunks provided by the Attribute Certificate Service (ACS) symmetrically, whereas in the Ouda-PoW scheme, the entire file should be hashed twice and encrypted once. However, we can also notice that the performance of both the Ouda-PoW and ase-PoW schemes is close to the performance obtained by our proposed scheme with respect to the complexity of client

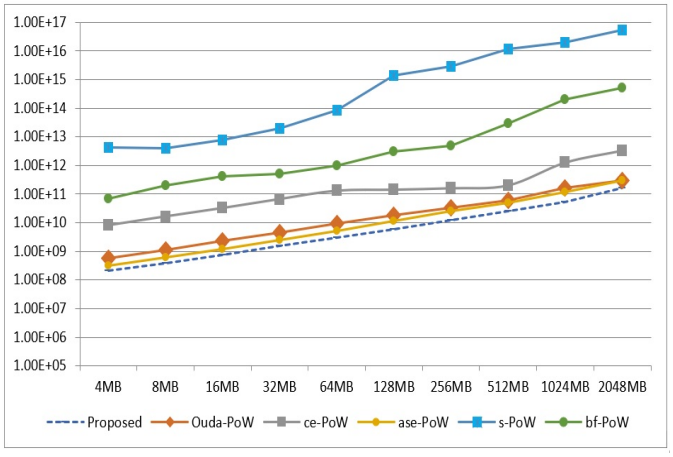


Fig. 5. Clock Cycles Required for the Client Initialization Phase.

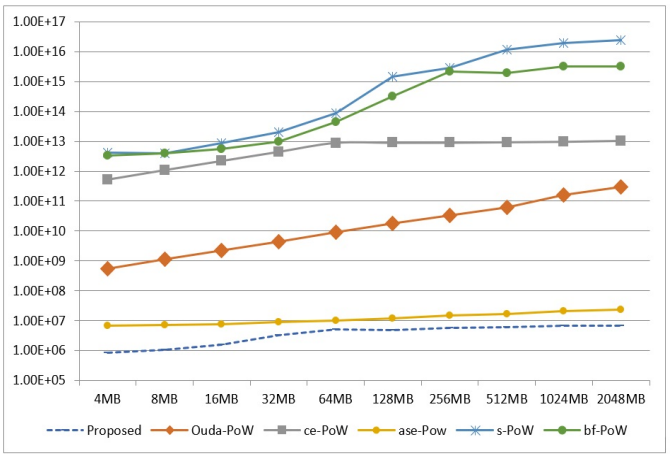


Fig. 6. Client Response Creation Clock Cycles.

initialization.

Fig. 6 shows the results obtained from comparing the proposed scheme with the other evaluated schemes regarding the time spent by the client responding to the server challenge. It is clear from the figure that the proposed scheme outperforms the other three schemes. This result is expected because the server in s-PoW uses a pseudorandom number generator to precompute the challenges that contain random file bits. In bf-PoW, the server initializes the bloom filter and divides the input files into chunks of fixed size, then creates the token chunks of each and inserts the function of each token into the bloom filters. In ce-PoW, the client encrypts all chunks specified by the server and then computes the hash over each encrypted chunk, increasing the computation time. The client in the ase-PoW scheme, on the other hand, performs several frequent encryptions of the designated chunks. In ouda-PoW, the client generates a random string of the same file size using the random seed received from the server, performs an XOR operation on the generated string with the CE file, and finally computes the hash of the resulting string. By contrast, in the proposed scheme, the client extracts the key from all blocks in the file and then encrypts only the block specified by the server.

VI. Conclusion

In this paper, we have proposed a secure and efficient proof-of-ownership scheme to thwart potential collusion attacks against client-side deduplication in cloud environments. The proposed PoW scheme's main idea is to divide the file to be uploaded into a number of fixed-sized blocks and then encrypt a randomly chosen block using a key formed by extracting one bit at a specified location in all other blocks. Unlike existing PoW schemes, the proposed scheme minimizes the exchanged information between the client and server and reduces the amount of data uploaded in memory during a PoW session. Moreover, it decreases the likelihood that a malicious client can successfully respond to the challenge sent to her by the server by increasing the amount of exchanged information between a malicious client and a legitimate file owner. The computational complexity of the proposed scheme was compared to five different PoW schemes, and experimental results showed that the proposed scheme outperforms the state-of-the-art PoW schemes concerning the time spent (clock cycles) for client initialization and response to the challenge received from the server.

Acknowledgment

The authors would like to thank the Deanship of Graduate Studies at Jouf University for funding and supporting this research through the initiative of DGS, Graduate Students Research Support (GSR) at Jouf University, Saudi Arabia.

References

[1] Ali Sunyaev. Cloud computing. In *Internet computing*, pages 195–236. Springer, 2020. doi:10.1007/978-3-030-34957-8_7.

[2] Chris Dotson. *Practical Cloud Security: A Guide for Secure Design and Deployment*. O'Reilly Media, 2019. URL https://www.oreilly.com/library/view/practical-cloud-security/9781492037507/.

[3] Srijita Basu, Arjun Bardhan, Koyal Gupta, Payel Saha, Mahasweta Pal, Manjima Bose, Kaushik Basu, Saunak Chaudhury, and Pritika Sarkar. Cloud computing security challenges & solutions-a survey. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 347–356. IEEE, 2018. doi:10.1109/CCWC.2018.8301700.

[4] Omer K Jasim Mohammad. Recent trends of cloud computing applications and services in medical, educational, financial, library and agricultural disciplines. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*, pages 132–141, 2018. doi:10.1145/3233347.3233388.

[5] Wen Xia, Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, Yu Hua, Min Fu, Yucheng Zhang, and Yukun Zhou. A comprehensive study of the past, present, and future of data deduplication. *Proceedings of the IEEE*, 104(9):1681–1710, 2016. doi:10.1109/JPROC.2016.2571298.

[6] Taek-Young Youn, Ku-Young Chang, Kyung-Hyune Rhee, and Sang Uk Shin. Efficient client-side deduplication of encrypted data with public auditing in cloud storage. *IEEE Access*, 6:26578–26587, 2018. doi:10.1109/ACCESS.2018.2836328.

[7] Shunrong Jiang, Tao Jiang, and Liangmin Wang. Secure and efficient cloud data deduplication with ownership management. *IEEE Transactions on Services Computing*, 2017. doi:10.1109/TSC.2017.2771280.

[8] Won-Bin Kim and Im-Yeong Lee. Survey on data deduplication in cloud storage environments. *Journal of Information Processing Systems*, 17(3): 658–673, 2021. doi:https://doi.org/10.3745/JIPS.03.0160.

[9] Roberto Di Pietro and Alessandro Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 81–82, 2012. doi:https://doi.org/10.1145/2414456.2414504.

[10] Weijing You, Lei Lei, Bo Chen, and Limin Liu. What if keys are leaked? towards practical and secure re-encryption in deduplication-based cloud storage. *Information*, 12(4):142, 2021. doi:https://doi.org/10.3390/info12040142.

[11] Dutch T Meyer and William J Bolosky. A study of practical deduplication. *ACM Transactions on Storage (ToS)*, 7(4):1–20, 2012. doi:https://doi.org/10.1145/2078861.2078864.

[12] Jian Liu, Nadarajah Asokan, and Benny Pinkas. Secure deduplication of encrypted data without additional independent servers. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 874–885, 2015. doi:10.1145/2810103.2813623.

[13] Youngjoo Shin, Dongyoung Koo, and Junbeom Hur. A survey of secure data deduplication schemes for cloud storage systems. *ACM computing surveys (CSUR)*, 49(4):1–38, 2017. doi:10.1145/3017428.

[14] Lorena González-Manzano, Jose Maria de Fuentes, and Kim-Kwang Raymond Choo. ase-pow: A proof of ownership mechanism for cloud deduplication in hierarchical environments. pages 412–428, 2016. doi:10.1007/978-3-319-59608-2_24.

[15] Taek-Young Youn, Nam-Su Jho, Keonwoo Kim, Ku-Young Chang, and Ki-Woong Park. Locked deduplication of encrypted data to counter identification attacks in cloud storage platforms. *Energies*, 13(11):2742, 2020. doi:https://doi.org/10.3390/en13112742.

[16] Shai Halevi, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg. Proofs of ownership in remote storage systems. pages 491–500, 2011. doi:https://doi.org/10.1145/2046707.2046765.

[17] Jay Dave, Avijit Dutta, Parvez Faruki, Vijay Laxmi, and Manoj Singh Gaur. Secure proof of ownership using merkle tree for deduplicated storage. *Automatic Control and Computer Sciences*, 54(4):358–370, 2020. doi:https://doi.org/10.3103/S0146411620040033.

[18] Tariqul Islam, Hassan Mistareehi, and D Manivannan. Secres: A secure and reliable storage scheme for cloud with client-side data deduplication. pages 1–6, 2019. doi:10.1109/GLOBECOM38437.2019.9013469.

[19] Shivansh Mishra, Surjit Singh, and Syed Taqi Ali. Mpows: Merged proof of ownership and storage for block level deduplication in cloud storage. In *2018 9th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–7. IEEE, 2018. doi:10.1109/ICCCNT.2018.8493976.

[20] Yongkai Fan, Xiaodong Lin, Wei Liang, Gang Tan, and Priyadarsi Nanda. A secure privacy preserving deduplication scheme for cloud computing. *Future Generation Computer Systems*, 101:127–135, 2019. doi:https://doi.org/10.1016/j.future.2019.04.046.

[21] Osama Ouda. A secure proof of ownership scheme for efficient client-side deduplication in cloud. *Journal of Convergence Information Technology*, 11:82–92, 2016. URL https://bit.ly/3sjSkxj.

[22] Ruiying Du, Lan Deng, Jing Chen, Kun He, and Minghui Zheng. Proofs of ownership and retrievability in cloud storage. pages 328–335, 2014. doi:10.1109/TrustCom.2014.44.

[23] Di Zhang, Junqing Le, Nankun Mu, Jiahui Wu, and Xiaofeng Liao. Secure and efficient data deduplication in jointcloud storage. *IEEE Transactions on Cloud Computing*, 2021. doi:10.1109/TCC.2021.3081702.

[24] VS Lakshmi, S Deepthi, and PP Deepthi. Collusion resistant secret sharing scheme for secure data storage and processing over cloud. *Journal of Information Security and Applications*, 60:102869, 2021. doi:10.1016/j.jisa.2021.102869.

[25] Shanshan Li, Chunxiang Xu, and Yuan Zhang. Csed: Client-side encrypted deduplication scheme based on proofs of ownership for cloud storage. *Journal of Information Security and Applications*, 46:250–258, 2019. doi:http://dx.doi.org/10.1016/j.jisa.2019.03.015.

[26] Jinbo Xiong, Fenghua Li, Jianfeng Ma, Ximeng Liu, Zhiqiang Yao, and Patrick S Chen. A full lifecycle privacy protection scheme for sensitive data in cloud computing. *Peer-to-peer Networking and Applications*, 8 (6):1025–1037, 2015. doi:10.1007/s12083-014-0295-x.

[27] Kangle Wang, Xiaolei Dong, Jiachen Shen, and Zhenfu Cao. An effective verifiable symmetric searchable encryption scheme in cloud computing. In *Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City*, pages 98–102, 2019. doi:10.1145/3377170.3377251.

[28] Hui Cui, Robert H Deng, Yingjiu Li, and Guowei Wu. Attribute-based storage supporting secure deduplication of encrypted data in cloud. *IEEE Transactions on Big Data*, 5(3):330–342, 2017. doi:10.1109/TBDATA.2017.2656120.

[29] Hua Ma, Ying Xie, Jianfeng Wang, Guohua Tian, and Zhenhua Liu. Revocable attribute-based encryption scheme with efficient deduplication for ehealth systems. *IEEE Access*, 7:89205–89217, 2019. doi:10.1109/ACCESS.2019.2926627.

[30] Jorge Blasco, Roberto Pietro, Agustin Orfila, and Alessandro Sorniotti. A tunable proof of ownership scheme for deduplication using bloom filters. *2014 IEEE Conference on Communications and Network Security, CNS 2014*, pages 481–489, 12 2014. doi:10.1109/CNS.2014.6997518.

[31] Lorena González-Manzano and Agustín Orfila. An efficient confidentiality-preserving proof of ownership for deduplication. *J. Netw. Comput. Appl.*, 50:49–59, 2015. doi:10.1016/j.jnca.2014.12.004.

[32] Ke Huang, Xiao-song Zhang, Yi Mu, Fatemeh Rezaeibagha, and Xiaojiang Du. Bidirectional and malleable proof-of-ownership for large file in cloud storage. *IEEE Transactions on Cloud Computing*, 2021. doi:10.1109/TCC.2021.3054751.

[33] Meixia Miao, Guohua Tian, and Willy Susilo. New proofs of ownership for efficient data deduplication in the adversarial conspiracy model. *International Journal of Intelligent Systems*, 36(6):2753–2766, 2021. doi:10.1002/int.22400.

[34] Angtai Li, Guohua Tian, Meixia Miao, and Jianpeng Gong. Blockchain-based cross-user data shared auditing. *Connection Science*, pages 1–21, 2021. doi:10.1080/09540091.2021.1956879.

[35] John Viega, Matt Messier, and Pravir Chandra. *Network security with OpenSSL: cryptography for secure communications*. " O'Reilly Media, Inc.", 2002. URL https://dl.acm.org/doi/10.5555/2167247.

A Secure Fog-cloud Architecture using Attribute-based Encryption for the Medical Internet of Things (MIoT)

Suhair Alshehri, Tahani Almehmadi
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah 21589, Saudi Arabia

***Abstract*—The medical internet of things (MIoT) has affected radical transformations in people's lives by offering innovative solutions to health-related issues. It enables healthcare professionals to continually monitor various medical concerns in their patients, without requiring visits to hospitals or healthcare professionals' offices. The various MIoT systems and applications promote healthcare services that are more readily available, accessible, quality-controlled, and cost-effective. An essential requirement is to secure medical data when developing MIoT architectures, as MIoT devices produce considerable amounts of highly sensitive, diverse real-time data. The MIoT architectures discussed in previous works possessed numerous security issues. The integration of fog computing and MIoT is acknowledged as an encouraging and suitable solution for addressing the challenges within data security. In order to ensure data security and to prevent unauthorized access, medical information is kept in fog nodes, and safely transported to the cloud. This paper presents a secure fog-cloud architecture using attribute-based encryption for MIoT to protect medical data. It investigates the feasibility of the proposed architecture, and its ability to intercept security threats. The results demonstrate the feasibility of adopting the fog-based implementation to protect medical data, whilst conserving MIoT resources, and the capability to prevent various security attacks.**

***Keywords*—MIoT; fog computing; cloud computing; ciphertext-policy; attribute-based encryption; security**

I. Introduction

Technology has transformed the lives of both individuals and organizations. Electronic healthcare services are considered to be an essential factor in the digital transformation of healthcare, helping to expand both the field and the kinds of healthcare services available, to improve the quality of healthcare services' delivery, and to reduce the cost associated with it [1].

The medical internet of things (MIoT) now leads technological advancements in healthcare, with the emergence of wearable and implantable medical devices and other technologies that enable the capture of medical data contributing to the considerable growth in this field. The value of the MIoT in the healthcare business is expected to increase to $534.3 billion by 2025, concurrent with an upsurge in the number of individuals with chronic conditions. According to the Grand View Research, this rise supports the need for technologically advanced medical devices [2].

The cost of healthcare services has increased significantly in recent years, although the costs concerned can be minimized via the industry's rapid digital transformation by improving operational efficiency [1]. Organizational and patient-care operations are enhanced when healthcare institutions are given real-time access to data generated by medical devices and MIoT-based devices. However, these advancements cause many challenges, particularly those related to the security and privacy of medical information [3].

Medical information is extremely susceptible to security threats [18], and the challenges imposed by privacy and security concerns hamper the widespread implementation of MIoT-based services. Thus, maintaining the confidentiality and safekeeping of information collected by MIoT devices remains a major research topic and must be prioritized, whether the information concerned is sent over the internet or stored in the cloud. Although the advent of fog computing resolved several issues that were apparent in traditional cloud-based architectures, medical data security remains a concern [4].

The current MIoT architectures that are developed in the literature to protect MIoT data fail to consider the limited capabilities of devices, such as storage and energy capacity, which affects the lifespan of the devices and their effectiveness in capturing and transmitting signals [4], [5], [7], [8], [9]. Thus, in order to ensure the adequate protection of medical information for MIoT services, a secure fog-cloud architecture is needed.

This paper develops a secure fog-cloud based architecture for the MIoT to ensure the security of medical data, while preserving the resources of these devices. It utilizes fog nodes to perform the encryption of the medical data, instead of MIoT devices using ciphertext-policy attribute-based encryption (CP-ABE). This is considered to be an ideal solution for protecting the privacy of medical data, as the data is encrypted by fog nodes before being stored in the cloud. The paper also considers minimizing the energy consumption of the processing overhead, improving the availability, and reducing the latency. The main contributions of this paper are as follows:

Our main contributions of this paper are as follows:

1) The design and development of a secure fog-cloud system using attribute-based encryption (ABE) for MIoT environments;

2) The definition of a set of requirements for achieving secure fog-cloud systems for MIoT environments;
3) The evaluation of the proposed system, in terms of a performance analysis, including network use, energy consumption, and security analysis.

This paper is organized as follows: Section II presents the fundamental details of the attribute-based encryption (ABE) employed in this paper. Section III discusses the current state of the art, focusing on the techniques that integrate ABE with fog computing. Section V-C defines a set of requirements for achieving secure fog-cloud architectures for MIoT environments. Section IV presents the proposed model, and Section V provides the evaluation results of the proposed model. Finally, Section VI concludes the paper.

II. Background

In widely distributed environments, especially the cloud environment, the symmetric encryption technologies with the same key for encrypting and decrypting suffer from key distribution and management issues. However, asymmetric encryption methods that use public and private keys lack computational efficiency [13], because the data owner is required to specify the identity of each recipient and their public key in advance, in order to implement the encryption algorithm, and to send the encrypted data to each recipient separately. This means the encryption process is repeated, according to the number of recipients. Since this type of individual scheme cannot be used to encrypt data once and send it to several users, ABE has emerged as a suitable solution for reducing the significant computational overhead of traditional encryption operations, while preserving data confidentiality and access control.

ABE is a novel and secure method for data sharing. It performs encryption and decryption while it obtains flexible access control [19], and was first introduced by Sahai and Waters [6]. It is an encryption mechanism that allows individuals to encrypt and decrypt data according to their attributes, such as job function, department, and specialty. ABE is an asymmetric cryptographic technique for one-to-many encryption that changed the traditional understanding of public-key encryption [6]. In traditional public-key encryption, the message is encrypted for a specific recipient using the recipient's public-key. In contrast, in ABE, one public-key is used to control access to encrypted data, using access policies and attributes [10].

Meanwhile, CP-ABE is a type of ABE that addresses the open challenge to organize access control and maintain the security of sensitive data, especially for internet of things (IoT) applications [13]. In a CP-ABE scheme, attributes are associated with the individual's secret key, and the encrypted message is associated with an access policy. Thus, authorized individuals can decrypt the message only if their secret keys and the associated attributes satisfy the decrypted message's access policy. This allows the storage of confidential data encrypted using CP-ABE on untrusted servers, without implementing authentication controls for the data access [13].

According to the extant literature, there are additional advantages to CP-ABE, compared to traditional cryptographic techniques [13], [20], [21], [22]. These advantages are as follows:

- It provides a high level of data confidentiality.
- It enables encrypted access control mechanism for access control applications.
- It reduces communication overload, because the generation of a user's secret key occurs only once.
- It achieves collusion resistance, because each attribute is associated with a polynomial or a random number that prevents legitimate users from colluding with each other.
- It supports user scalability; as the number of authorized users increases, the system can work efficiently.

A CP-ABE scheme consists of four fundamental algorithms: setup, encrypt, key generation, and decrypt. A detailed description of these algorithms can be found in [6].

III. Related Work

The model of the MIoT is believed to be tremendously valuable for remote health monitoring systems. The critical nature of the functions that use these systems requires a great degree of precision and accessibility. The lack of accessibility and punctuality, as well as the reliability of cloud-based IoT is a much debated topic, particularly regarding instances when the internet connection becomes undependable, and/or slower than expected. Furthermore, due to the centralized resource management and policies set by service providers that are related to the nature of cloud computing, the systems are susceptible to infiltration. An electronic healthcare system cannot be sustained with this vulnerability, due to the necessity to protect medical data. This necessitates the adoption of fog-based MIoT systems to overcome the limitations of cloud-based MIoT systems.

Fog computing is a distributed platform that generates a new layer between the cloud and MIoT devices that decreases the amount of processing done in the cloud, thereby allowing more efficient and effective service delivery. However, fog computing involves a number of security issues that are inherited from the cloud itself, including the ability to verify identities, to authenticate user inputs, to enforce access control, and to preserve privacy. Several procedures exist that can be applied to resolve the issues faced by fog computing.

In their work, Alrawais et al. [4] proposed a key exchange protocol based on the CP-ABE that can be employed to facilitate authentic transmissions between the set of fog nodes and the cloud, while maintaining confidentiality. To accomplish this goal, the researchers integrated digital signature techniques with a CP-ABE protocol, within which they studied the effectiveness of the protocol, in terms of both performance and security. To demonstrate its practicality, the protocol was implemented and contrasted with the certificate-based scheme. The results indicated that the protocol proposed was more practicable, as well as more effective than the certificate-based schema. However, the study did not examine the security requirements between the fog nodes and the end users.

Meanwhile, Vohra and Dave [5] held a similar view to Alrawais et al. [4], and investigated the privacy problems in fog computing, and the efficiency of the ABE schema suggested in [6]. The outcomes of the investigation demonstrated that ABE successfully guarded sensitive information, but that this method alone was not adequate, due to several security problems, such as backward and forward issues. Thus, a number of methods, such as re-encryption, were suggested by previous researchers, and Vohra and Dave [5] recommended a system that used CP-ABE for encryption and re-encryption to provide access control for the communication between the cloud and the fog. According to the findings of this study, the schema suggested demonstrated improved operation and protection. However, there was a need to reduce the number of messages generated to enhance system efficiency.

In contrast, Zhang et al. [7] presented an access control mechanism using a CP-ABE approach for the secure sharing of data that supported the outsourcing of fog computing for complex encryption and decryption operation. In this approach, a number of operations were implemented locally by the data owner and end user, such as implementing the symmetric algorithm on the data, and additionally encrypting or decrypting the symmetric key, partly via the CP-ABE algorithm. However, the study neglected the limited resources of smartphones, as these operations drain their energy and resources.

Meanwhile, Porwal [8] modified the CP-ABE protocol to enhance the secure exchange of content keys between the IoT, fog devices, and the cloud by exploiting the hierarchy in the attribute set of access policy to obtain a single integrated access policy. In order to reduce the cost of storage, to distribute encrypted content, and to reduce the number of decryption operations, a fog device and cloud received solely entitled content keys using one decryption operation. Despite the limited resources within IoT devices, the study assumed that they were capable of implementing advanced encryption standard (AES) to protect data, and that they shared the content keys used for encryption with fog devices and the cloud that might be unreliable and untruthful.

Finally, Fan et al. [9] considered the problem that when the access policy is sent clearly with the ciphertext it may reveal confidential information, although the CP-ABE scheme provides a secure access policy within the ciphertext (10). Thus, they developed an efficient multi-authority access control mechanism for the fog that supported the IoT. The scheme outsourced partial decryption, and transformed user attributes to anonymous aspects to preserve users' privacy. However, the scheme used fog devices for decrypting only while the complex CP-ABE encryption operations were performed by the data owner on end devices, an approach that is not suitable for IoT devices, because of their limited resources.

A. Data System Requirements in MIoT Environments

The concerns regarding data security, such as data scams, theft, forgery, or destruction, take precedence over the numerous advantages of utilizing fog computing in MIoTs. The primary emphasis of the present research was (1) the safeguarding of information when transporting from sensor network to fog, (2) the safekeeping of data during the transfer from fog to cloud, and (3) the security of data buffering in the fog, and the final storage in the cloud. To achieve these goals, a set of requirements were defined that were based on the previous literature in the field, and then the related work that utilized ABE for the fog-cloud architecture was compared, in terms of these requirements. The comparison is presented in Table I, and the requirements were defined as follows:

- **Confidentiality:** Sensitive data available to authorized users only.
- **Access Control:** Only authorized users can access the functionality and data within the device.
- **Availability:** Medical information can be accessed from anywhere within reach of the cloud services when needed.
- **Integrity:** Unauthorized users cannot change or alter the data.
- **Low-latency:** Protect data near its source; the delay in data protection increases the possibility of attacks on the system, and discloses patients' sensitive medical data.
- **Forward Security:** The prevention of nodes/users who have exited the database from retrieving the information exchanged.
- **Energy-Efficiency:** Preserve energy by transferring the encryption/decryption processes from the end devices to the fog nodes.

TABLE I. DATA SYSTEM REQUIREMENTS IN MIOT ENVIRONMENTS FOR DIFFERENT STUDIES

Design Requirements	References				
	[4]	[5]	[7]	[8]	[9]
Confidentiality	✓	✓	✓	✓	✓
Access Control	✓	✓	✓	✓	✓
Availability			✓		✓
Integrity	✓	✓	✓	✓	✓
Low-latency			✓	✓	✓
Forward Security		✓			✓
Energy-Efficiency					

The limited capabilities of MIoT devices, such as low power and dependence on limited-life batteries, impose restrictions on the kind of operations that can be utilized by these devices. As illustrated by the discussion of the previous studies in this field and Table I, authors of previous work relied on implementing cryptography operations in the IoT devices, which engendered the depletion of their limited resources. This was inconsistent with the energy efficiency requirement that this paper sought to attain, alongside the other security conditions.

IV. THE PROPOSED MODEL

Fog computing can offer a viable solution for handling various security issues in MIoT successfully [11], due to the existence of fog nodes on the network edge that gather confidential health-related data. This thereby provides data processing on the edge, decreasing the transfer of confidential information to the cloud, and supporting the protection of confidentiality by protecting the health-related information in

the fog nodes, and between the fog nodes and the cloud. A variety of procedures can be utilized between the fog nodes and the cloud to preserve data protection [12]. This paper proposes a secure fog-cloud architecture for MIoT to enhance the safety of user data, without affecting the efficient functioning of MIoT. Accordingly, it proposes that fog nodes are an appropriate platform for protecting medical data, due to their closeness to the end-user.

It was necessary to ensure the security requirements for the manipulation of medical data, including transmission and storage, that were identified and discussed in the previous section. Thus a scenario was examined in which the stream of data captured from MIoT devices was transmitted to cloud storage through fog nodes that implemented CP- ABE security. As an advanced type of encryption technology, and the most common type of ABE, CP-ABE addresses the open challenge for maintaining access control, and the security of sensitive data, especially for IoT applications [13]. In a CP-ABE scheme, the attributes are integrated with the user's secret key, and the encrypted text is integrated with an access policy. Therefore, only users with the correct attributes that meet the access policy can decipher the data. Hence, the main advantage of CP-ABE is that it enables the storage of confidential data on an untrusted server, with no need to implement authentication or access control mechanisms [13].

A. System Components

The proposed scheme contains five types of entities: data owner (DO), data user (DU), multiple fog nodes (FN) at the edge, cloud server (CS), and key authority (KA). The model is shown in Figure 1, and the description of the model entities is as follows:

- **Key Authority (KA)** This is responsible for generating cryptographic parameters, and creating the secret keys of all users, according to their attributes. Additionally, the KA has the duty of publishing a list of users (UL) to the FN, and updating the list when adding or revoking a user. This function is necessary to achieve forward security;
- **Data Owner (DO)** This is also referred to as a 'patient' in the proposed scheme. The DO uses MIoT sensors, which are miniature apparatuses with limited capabilities regarding storage, computing processing, and energy. Medical data is collected from the body of the patient and transferred to the fog node linked to the devices. To safeguard the transmissions between the medical devices and the fog, a secure sockets layer (SSL) is used to establish an encrypted channel between the MIoT and the fog node. In the case where the medical device and the FN fail to communicate, the MIoT devices seek the closest alternative FN to contact and transmit the medical data. Each MIoT device has a unique identifying attribute, called an internet protocol (IP) address. The MIoT device can transmit/collect information over the network using its IP address. The framework of the internet affords universal connectivity to medical devices in a heterogeneous network. In MIoT devices, the communication and data transmission is conducted via wireless technology. There are many aspects to consider when using wireless technology. Specifically, for MIoT devices, factors such as energy efficiency, cost-effectiveness, physical dimensions, and user-friendliness must be considered. Hence the suitable wireless specifications for MIoT include Bluetooth, RF4CE, IrDA, Wi-Fi, ZigBee, RFID, NFC, and ANT;
- **Fog Nodes (FN)** These are positioned at the edge of the network, and provide an array of amenities, such as minimal inactivity occasions and real-time functionality. They also control the performance of the encryption processes. Each FN has a unique identity that is connected to the cloud server by an IP network. It is important to note that the structural design of fog computing can facilitate MIoT devices in supplying efficient storage services, processing raw data near its source, and reducing network traffic to prevent congestion, all of which are critical characteristics of MIoT healthcare applications;
- **Cloud Server (CS)** This can store huge amounts of data, and has formidable computing power. It is responsible for storing encrypted data, as well as for processing requests for access to medical data from authorized healthcare providers.
- **Data Users (DU)** These are also referred to as 'healthcare providers', and are the parties who monitor a patient's condition, and request their data from the CS, in order to recommend the appropriate treatment.

B. Security Assumption

In order to simulate the security environment in this study's scheme, the following security assumptions were considered:

- It was assumed that MIoT sensors, FN, and the KA were within the same local network, and that the connection link between them was secure;
- The KA was trusted fully in the scheme, which meant that it would not leak data, or collude with any users;
- Each FN was trusted, but could be vulnerable to attack;
- The CS was a semi-trusted service provider. It was honest and ensured data security, but was curious, and therefore performed analysis to collect private information;
- The DU may collude to gain unauthorized data. It was assumed the revoked DU could not perform data decryption from the FN.

C. CP-ABE Functions

The CP-ABE functions adopted in the scheme were based on [10], where a full description of these algorithms can be found. The main functions included in the proposed solution were as follows:

- **Setup( )$\Longrightarrow$*MK,PK*:** The setup algorithm takes no input, and outputs the master-key (*MK*) and public-key (*PK*);

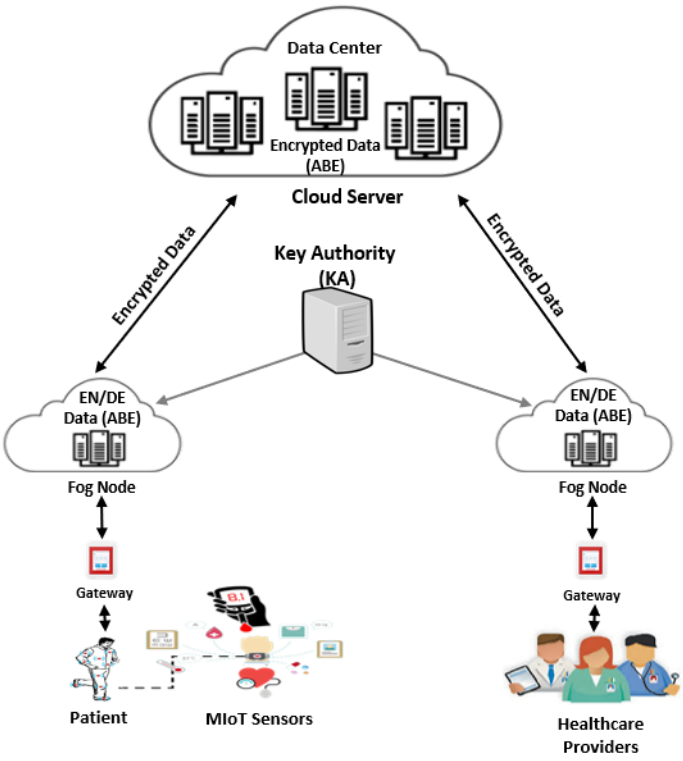


Fig. 1. The Proposed Architectural Design.

- **Encrypt(PK, $Data$, $\mathbb{A}$)$\Longrightarrow CT$:** The encrypt algorithm takes as input the PK, data, and the access-policy ($\mathbb{A}$) of the user. It enciphers the data and outputs a ciphertext (CT) in which only a user who owns a key with a set of attributes that fulfill the $\mathbb{A}$ can decipher the data;
- **KeyGen(MK, S)$\Longrightarrow SK$:** The key-generation algorithm takes as input the MK, and the set of attributes (S) of a user to produce the secret-key (SK) of this user;
- **Decrypt(PK, SK, CT) $\Longrightarrow Data$:** The decrypt algorithm takes as input the PK, SK, and the CT, which contains an $\mathbb{A}$. If the user attributes satisfy the $\mathbb{A}$, then the algorithm deciphers the CT and returns the original data ($Data$).

D. System Workflow

As shown in Fig. 2 the proposed system works in the following steps:

- **Step 1:** The system is initialized, with the KA executing the setup algorithm to generate the MK and the PK. The PK is broadcast to all FN that are connected to the local domain, while maintaining the MK. In addition, the UL is distributed to the FN, which carry the users' identities in the system, SK, S, and $\mathbb{A}$ that are used in the encryption/decryption operations;
- **Step 2:** When receiving the data from the MIoT sensors, the FN performs the encrypt algorithm, after authenticating the identity of the DO via checking the UL to ensure the presence of the user in the system, and to obtain the $\mathbb{A}$ that was used for the encryption. The CT is sent for storage in the CS;
- **Step 3:** If a DU requests data from the CS, the request passes through an FN that first verifies the user's identity from within the UL. This is to prevent users who have left the system from accessing the data, since the data request is sent only to the CS if the user is currently on the UL. Otherwise, the request is rejected;
- **Step 4:** The FN executes the Decrypt algorithm after receiving the CT based on the user's SK that was verified in the UL, then sends the original data to the data user.
- **Step 4:** The FN executes the decrypt algorithm after receiving the CT, based on the user's SK that was verified in the UL, then sends the original data to the DU.

V. Evaluation

In order to represent the proposed model, iFogSim [14], [15], [16] was used to create the physical components and to design the application model as a group of modules, namely the client module, security module, and storage module, that constituted the data processing elements. The client module received the raw ECG signals. It then performed the abstraction process on the signals, and ignored any inconsistent readings.

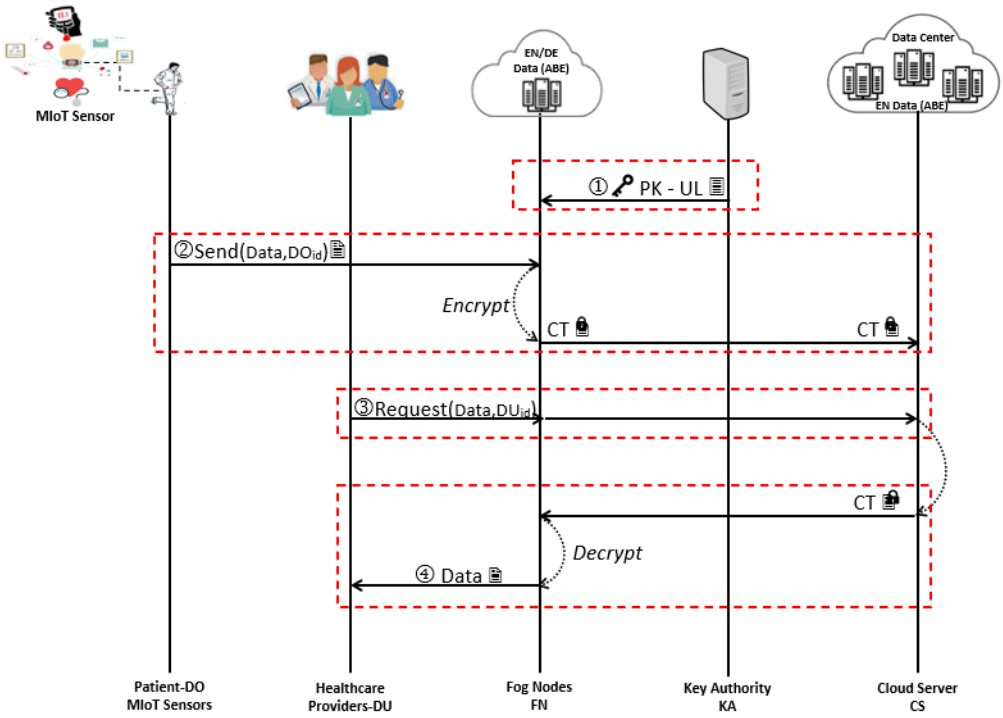


Fig. 2. System Workflow.

If the value of the signal sensed was consistent, it was sent to the security module. In the case of receiving information from the security module, it sent it to the DISPLAY actuator to display. The security module was responsible for protecting the medical data, performing encryption and decryption operations on the data received. The storage module represented the database in the cloud where the encrypted medical data was stored for patients, and recovered when needed.

In addition, this study used CP-ABE, an open-source Java implementation of the ABE scheme found in the paper by Bethencourt et al. [10], and developed by Wang et al. [17], to evaluate the efficacy of the model. Three different placement strategies were tested by implementing the security module, the main module of the present study, in the cloud, fog, or end device, as shown in Fig. 3. The feasibility of the proposed architecture was assessed, in terms of performance, network use, and energy consumption. Furthermore, the capability of the proposed system to deal with the related security vulnerabilities within the MIoT, the FN, and when information was transferred between the FN and the cloud, was examined.

The test was conducted on a local PC with a Windows 10 Home 64bit operating system, a 2.2 GHz Intel Core i7, and an 8GB Memory by Eclipse IDE for Java Developers. In order to test the performance on different sizes of topology, the number of FN was changed, and the number of end devices connected to each FN fixed was retained. Different sizes of network topology were simulated, starting with one FN with four connected end devices, and advancing to 20 FN with 80 connected end devices.

A. Performance Analysis

The various metrics that iFogSim showed were collected, and the results illustrated how different placement strategies for implementing a security module affected the system performance.

1) Time to Security Module:: The time it took to transfer the data from its source to the security module to implement the encryption was the most important factor in this research, as a delay in the transmission of data increased the possibility of the data being attacked. Fig. 4 shows the time taken to transfer the medical data to the security module, illustrating that it decreased significantly when applying the security module to fog devices or end devices. This was more apparent when the number of devices was increased.

The results of the implementation of a security module in the cloud were excluded from the study, which focused on performing a comparison between implementing the security module in the fog and in the end devices, as most of the previous research reviewed focused on implementing the security mechanism in the end devices, because such devices are close to the data source (end users), without considering the limited resources of these devices. In contrast, this study sought to move the security operations to the fog devices to preserve the end devices' resource consumption. As shown in Fig. 5, there was a very small difference in time of not more than six milliseconds between implementing the security module in the fog and in the end devices. This slight delay in data transfer would not affect the data security, especially since the

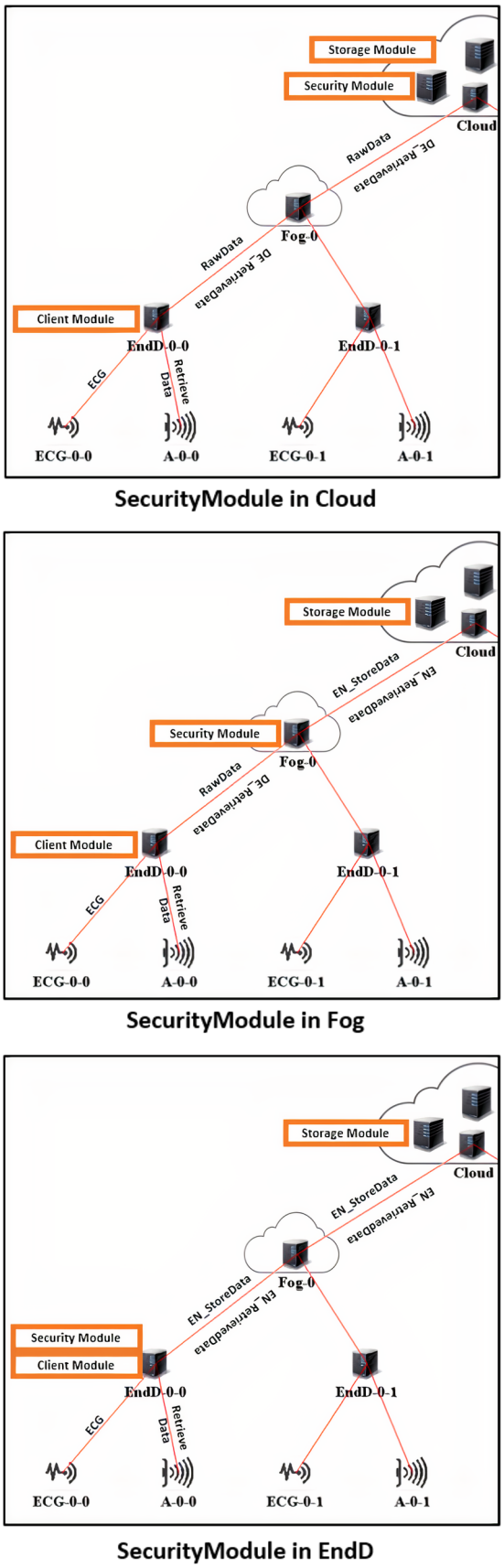


Fig. 3. The Different Placement Strategies.

FN and the end devices were in the same local network, and the connection between them was secure.

2) Network Use:: As shown in Fig. 6, the increase in the number of the connected devices greatly increased the load on the network, especially when implementing the security module in the cloud or end devices. When considering the security module's implementation strategy on the FN, the network use decreased significantly.

This result demonstrated that network congestion can be avoided when adopting a fog-based implementation, as the data is processed near the data source, thereby reducing the amount of information sent to data centers in the cloud significantly. Other placement strategies engender degradation and congestion in the network.

3) Energy Consumption:: MIoTs and end devices, or gateways, have limited processing capability, minimal storage space, and inadequate power, because they are often battery operated, and therefore lack sufficient power to secure medical data. This study sought to preserve the resource consumption, such as the energy resources of the MIoT layer, and consequently recorded various measurements of energy consumption in the end devices, according to different placement strategies for applying the security module.

Fig. 7 shows the average energy resource consumption in the end devices when implementing the security module in the cloud, end device, or FN. A significant reduction in energy consumption was observed in the case of the fog-based implementation that contributed to the preservation of the limited resources in the end devices.

B. Security Analysis

The security of the proposed model was analysed from the perspective of data confidentiality, fine-grained access control, collusion attack resistance, and forward secrecy.

1) Fine-grained Access Control:: This is a mechanism whereby legitimate users are given different access privileges to the data. Fine-grained access control was also attained in the model as it was a feature of the CP-ABE scheme that gave each legitimate user different access to the medical data concerned, according to the user's attributes and role.

2) Data Confidentiality and Privacy:: This is a basic and important requirement when using the cloud for data storage. In this model, it was accomplished using CP-ABE, whereby the medical data was encrypted near its source in the local network, before it was sent for storage in the remote data centers or the cloud. Therefore, no unauthorized user could access the content of the medical information stored there.

3) Collusion Attack Resistance:: In this type of attack, users combine their attributes to obtain unauthorized data illegally. These attacks can be conducted by system users seeking higher access rights, therefore, a system should prevent users from undertaking such attacks. To avoid this kind of attack, and to prevent users from combining their attributes in order to decipher the medical data, the proposed data protection scheme employed CP-ABE; since each attribute was associated with a polynomial or a random number, different users could not collude to obtain higher data access rights.

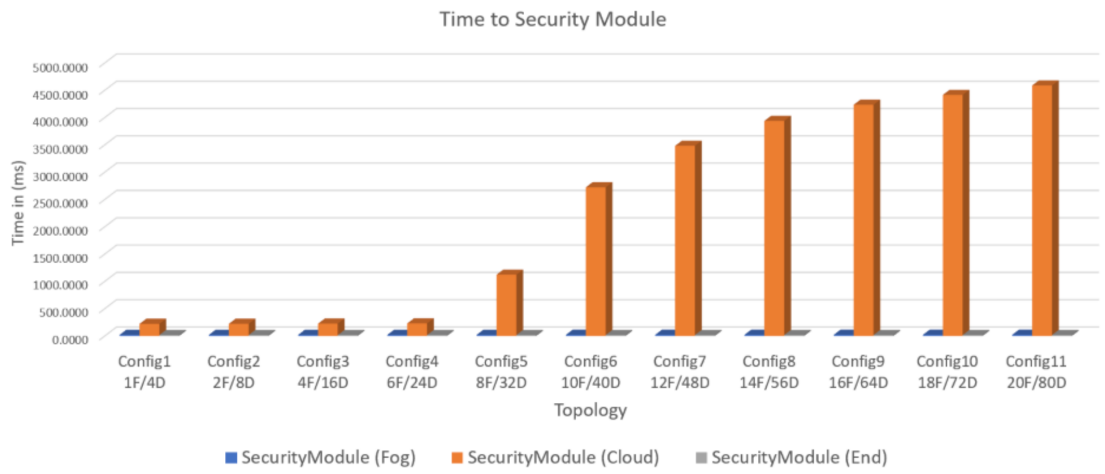


Fig. 4. Time to Security Module.

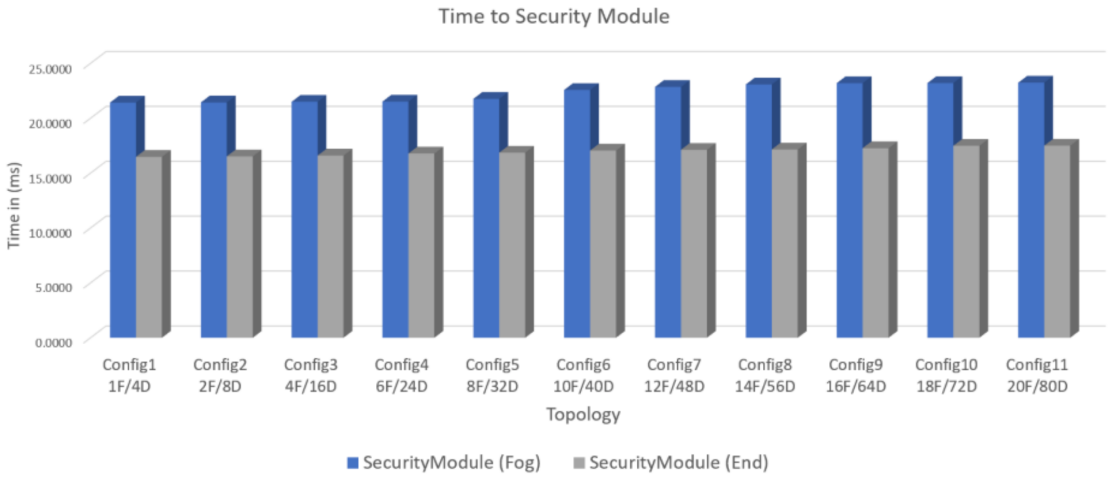


Fig. 5. Time to Security Module - The Comparison between Fog and End Device.

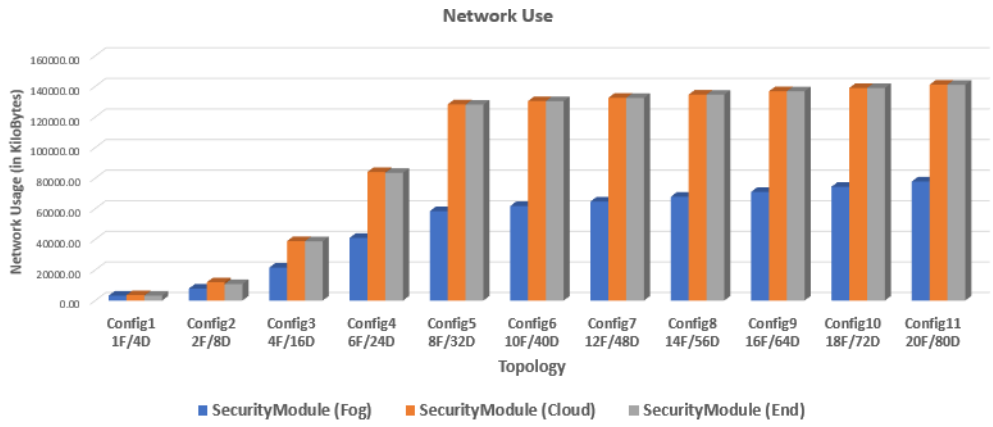


Fig. 6. Network Use.

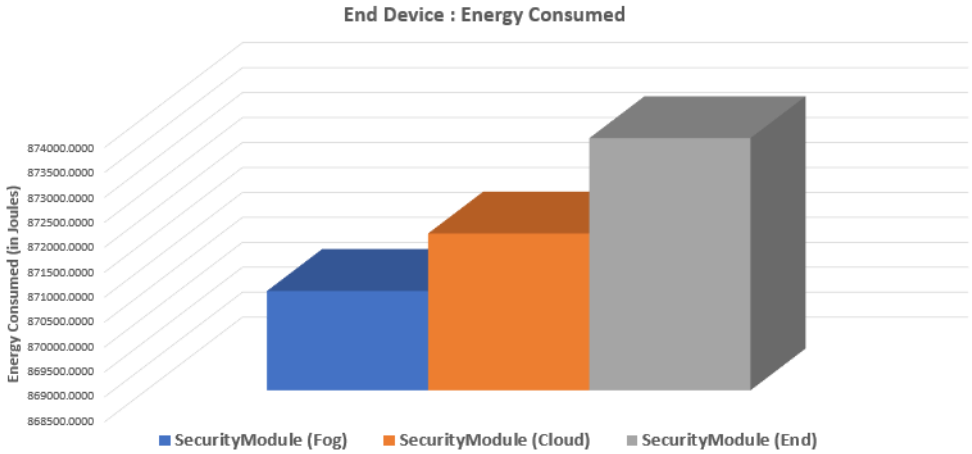


Fig. 7. Energy Consumption.

4) Forward Security:: This involves the prevention of any user who has been revoked from accessing and deciphering the medical data concerned. It was achieved in the proposed model by publishing the UL in the FN that contained the legitimate users with their set of attributes. Once a user left, the list was updated by the KA, thereby preventing the user who had left the system from accessing the medical data.

C. Comparison Analysis

The proposed model was compared with the previous works discussed in Section III in terms of the requirements presented in Section V-C. As previously mentioned in Section V-C, the related works as in [9] relied on MIoT devices to preserve data security and perform complex computations. This engenders the rapid depletion of the end device's energy and occupies it with other non-medical processes that may affect the efficiency of the sensors and the monitoring of vital parameters. In the proposed model, we employed FN to prevent costly computations, and thus, preserve the energy consumption of the MIoT devices. Concisely, the proposed model achieves all the predefined requirements including energy efficiency.

VI. Conclusion

This paper presented a fog-cloud system for the MIoT, in which the resource-limited end devices employed FN to prevent costly computations. It examined the feasibility of the proposed architecture, and the capability of the schema to manage security threats. The results demonstrated the benefits of adopting the fog-based implementation for protecting medical data and conserving MIoT resources, together with the ability to optimize network usage, and to avoid congestion, while reducing the amount of data sent to the data centers in the cloud significantly. It is hoped that the proposed model will encourage the adoption of MIoT devices and services, and thus encourage the healthcare industry to improve patient care services, and provide a better healthcare experience. For future work, we want to consider a technique with a wider range coverage to allow reducing the number of gateways. We also plan to conduct a real-time experiment to validate the proposed model.

References

[1] T. Khubone, B. Tlou, and T. P. Mashamba-Thompson, "Electronic Health Information Systems to Improve Disease Diagnosis and Management at Point-of-Care in Low and Middle Income Countries: A Narrative Review," Diagnostics, vol. 10,5, pp. 327, Multidisciplinary Digital Publishing Institute, 2020.

[2] K. H. Nam, D. H. Kim, B. K. Choi, and I. H. Han, "Internet of Things, Digital Biomarker, and Artificial Intelligence in Spine: Current and Future Perspectives," Neurospine, 16(4), 705–711, 2019.

[3] F. A. Kraemer, A. E. Braten, N. Tamkittikhun, and D. Palma, "Fog computing in healthcare–a review and discussion," IEEE Access, vol. 5, pp. 9206–9222, IEEE, 2017.

[4] A. Alrawais, A. Alhothaily, C. Hu, X. Xing, and X. Cheng, "An attribute-based encryption scheme to secure fog communications," IEEE access, vol. 5, pp. 9131–9138, IEEE, 2017.

[5] K. Vohra and M. Dave, "Securing fog and cloud communication using attribute based access control and re-encryption," In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 307–312, IEEE, 2018.

[6] A. Sahai and B. Waters, "Fuzzy identity-based encryption," In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 457–473, Springer, Berlin, Heidelberg, 2005.

[7] P. Zhang, Z. Chen, J. K. Liu, K. Liang, and H. Liu, "An efficient access control scheme with outsourcing capability and attribute update for fog computing," Future Generation Computer Systems, vol. 78, pp. 753–762, Elsevier, 2018.

[8] S. Porwal and S. Mittal, "HE3: A hierarchical attribute based secure and efficient things-to-fog content sharing protocol," Journal of King Saud University-Computer and Information Sciences, Elsevier, 2019.

[9] K. Fan, H. Xu, L. Gao, H. Li, and Y. Yang, "Efficient and privacy preserving access control scheme for fog-enabled IoT," Future Generation Computer Systems, vol. 99, pp. 134–142, Elsevier, 2019.

[10] J. Bethencourt, A. Sahai and B. Waters, "Ciphertext-Policy Attribute-Based Encryption," In: 2007 IEEE Symposium on Security and Privacy (SP '07), Berkeley, CA, USA, 2007, pp. 321-334.

[11] S. Yi, Z. Qin, and Q. Li, "Security and Privacy Issues of Fog Computing: A Survey," In: International conference on wireless algorithms, systems, and applications, WASA 2015. Lecture Notes in Computer Science, vol 9204. Springer, Cham.

[12] A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog Computing for the Internet of Things: Security and Privacy Issues," IEEE Internet Computing, vol. 21, no. 2, pp. 34–42, IEEE, 2017.

[13] R. R. Al-Dahhan, Q. Shi, G. M. Lee, and K. Kifayat, "Survey on Revocation in Ciphertext-Policy Attribute-Based Encryption," Sensors, vol. 19, no. 7, p. 1695, Multidisciplinary Digital Publishing Institute, 2019.

[14] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, and R. Buyya, "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments," Software: Practice and Experience, vol. 47, no. 9, pp. 1275–1296, Wiley Online Library, 2017.

[15] I. Lera, C. Guerrero, and C. Juiz, "YAFS: A Simulator for IoT Scenarios in Fog Computing," IEEE Access, vol. 7, pp. 91745–91758, IEEE, 2019.

[16] R. Buyya; S. N. Srirama, "Modeling and Simulation of Fog and Edge Computing Environments Using iFogSim Toolkit," Fog and Edge Computing: Principles and Paradigms , Wiley, 2019, pp.433-465.

[17] J. Wang, "Java Realization for Ciphertext-Policy Attribute-Based Encryption," Computer Science College of Shandong University, 2012.

[18] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, Kumar R., and R. A. Khan, "Healthcare Data Breaches: Insights and Implications," Healthcare (Basel, Switzerland), 8(2), 133, 2020.

[19] Q. Huang, and L. Wang, and Y. Yang, "Secure and Privacy-Preserving Data Sharing and Collaboration in Mobile Healthcare Social Networks of Smart Cities," Security and Communication Networks, Hindawi,Volume 2017, Article ID 6426495.

[20] C. Lee, P. Chung, and M. Hwang, "A Survey on Attribute-based Encryption Schemes of Access Control in Cloud Environments," International Journal of Network Security, Vol.15, pp. 231-240, 2013.

[21] R. N. Lakshmi, R. Laavanya, M. Meenakshi, and C. S. G. Dhas, "Analysis of Attribute Based Encryption Schemes," International Journal of Computer Science and Engineering Communications, Vol.3, Issue 3, pp. 1076-1081, 2015.

[22] S. Moffat, M. Hammoudeh, and R. Hegarty, "A Survey on Ciphertext-Policy Attribute-based Encryption (CP-ABE) Approaches to Data Security on Mobile Devices and its Application to IoT," In Proceedings of the International Conference on Future Networks and Distributed Systems (ICFNDS '17). Association for Computing Machinery, New York, NY, USA, Article 34.

Efficient Weighted Edit Distance and N-gram Language Models to Improve Spelling Correction of Segmentation Errors

Hicham GUEDDAH
Intelligent Processing and Security of Systems Team- F.S.R,
E.N.S, Mohammed V University in Rabat,
B.P:8007, Avenue des Nations Unies, Agdal, Rabat, Morocco.
https://www.researchgate.net/profile/Hicham-Gueddah

***Abstract*—In most research that has dealt with the correction of spelling errors, the errors are caused by the misuse of space (deletion or insertion of space) are not tackled. Forgetting to deal with this type of errors in the texts poses a problem of understanding and ambiguity of the meaning of the sentence containing these errors. In this article, we propose a new approach to correct errors due to the insertion of space in a word, and at the same time correct other types of editing errors. This approach is based on the edit distance and uses bi-grams language models to correct words in context. The test conducted on hundreds of erroneous words (by insertion of space and/or by simple editing errors) made it possible to assess the relevance and validity of the methods developed to correct this type of error. The approaches proposed in this article provide a very important clarification and reminder by comparing them to those of other existing approaches.**

***Keywords*—Spelling correction; error; natural language; insertion; space; distance; language models; probability**

I. Introduction

For several years now, the language industry and new information and communication technologies have been evolving. Alongside this progress, thousands of electronic documents such as newspapers, emails, blogs, dissertations and theses are produced on a daily basis. Therefore, the existence and need for spelling correction systems in NLP applications is of paramount importance to improve and help with sound and unambiguous writing.

Automatic spelling error correction is currently ubiquitous and integrated into all computer tools such as word processing, email, social media, search engines, which are frequented every day by millions of people around the world.

For a long time, by analyzing the strategy of correction systems integrated into large word processing software such as Microsoft's WinWord, OpenOffice Writer, or those embedded in web textures (email, search engine), we have pointed out that these remain ineffective for correcting certain types of spelling errors. They're committed when typing Arabic text, for example, which we cite as errors resulting from insertion and/or untimely deletion of the space character in a lexical form.

Their strategies consist only in proposing solutions separately to segments divided by the insertion of space.

Automatic correction of spelling errors has been the topic of much research since the 1960's [1]. Despite the monopolization of this axis by the major computer production industries,it remains a promising domain of research [2][3]. The principle is to propose the most similar solutions to a word detected out of vocabulary based on the lexical similarity inter words.

Among the work in the subject area of spelling correction, we mainly include:

Damerau's analysis of typographical errors [4]. He indicated that about 80-95% of mistakes in English texts are unique errors that are induced by poor insertion, deletion, permutation of a single character or the transposition of two adjacent character. This analysis has been the cornerstone of the concept of error as a simple or multiple combinations of operations, called elementary editing operations (insertion, deletion, transposition, and permutation).

Based on Damerau's work, Levenshtein [5] considered only three editing operations (insertion, deletion, permutation), and subsequently defined a metric that allows us to compare two words while calculating the number of editing operations undergone on one word to turn it into some other word. This distance is also called edit distance which remains, despite the technique, the most widely used in spelling correction and which has also been the themes of several adaptations and weighting [6]. Then a series of similar works were carried out. We can put them into different categories:

- Metric-based correction approaches such as Jaro distance [7], Jaro-Winkler distance [8], Jaccard distance [9], distance of Stoilos [10].
- Probabilistic correction approaches such as n-grams decomposition [11], the correction method based on the noisy channel model [12], Alpha-code methods [13], or those based on probabilistic automatons [14], [15]
- Since 2012 Gueddah, Yousfi and Nejja have carried out a series of work on spelling correction for the Arabic language, with the aim of improving the scheduling rate of solutions returned by the classical edit distance [16] [17] [18], or integrating the morphological analysis into the spelling correction phase [19] [20], or integrating context into spelling correction [21] [22].

II. Related Work

Based on our bibliographical research on spelling correction, we noticed that the bulk of this work did not adequately address the errors due to insertion or deletion of space in the seized texts. The number of works dealing with this category of error is very negligible compared to the number of works in the field of spelling correction.

The work that dealt with errors due to the insertion or deletion of space is of two types:

- Census-type studies: Mitton [3], and Kukich [1] noted that more than 14% of spelling mistakes in seized text are mainly related to the omission of the space between words. Conforming to a statistical study of errors made in Urdu typed text, Tahira [23] found out that space errors are a fairly significant percentage of errors, more than 75% of errors are neighborhood errors, 32% of which are linked to the insertion of space in words.
- Research work that provides hints for dealing with errors due to the insertion/deletion of space in words, either at the level of detection of this type of error or in the actual correction. This work includes, for example, the work that relies on the generation of all possible partitions of the erroneous word and testing whether or not the segments exist in the dictionary [24][25][26].
- The study presented by Alkanhal and al [27], according to the latter, the correction of space insertion errors uses two procedures, the initiative is to merge the different neighboring words from the wrong word. The outcome of this procedure is a list of the various possible combinations of this merger. This list may contain valid fragments and other erroneous fragments that will subsequently be handled on to the second correcting procedure.

In this article, we suggest a new metric approach that uses bi-gram language models to correct spelling errors due to the insertion of spaces (also called segmentation errors), into a correct or incorrect word taking into account the context in which this type of error was induced.

III. Errors Due to Space Insertion and Deletion

A. Defining Space Errors

Although the emergence of new generations of near-present correctors in most word processing editor, emails, blogs, social media, smartphones, these remain ineffective and unsuccessful [28] with regard to correcting errors due to the insertion or deletion of space in or between words. These forms of mistakes can be induced in several situations:

- Because one writes by accelerating without regard to who has been seized, the editor may unintentionally insert one or more spaces within a word in the belief that he inserted it to separate between two words. Therefore, this type of error leads to the appearance of two or more segments of words that may be lexicon or wrong words.
- Moreover, this kind of error can be due to poor optical text recognition (OCR), which can additionally insert the space character inside the word, or as a result of a file type conversion, such as converting Word documents to Pdf or vice versa.

However, the problem of actual word error is more complex. Generally, such an error disrupts the syntax and then rectify it.

Example:

Instead of typing the word "misspelled", you add a space in that word and you get both sequences "missp" and "elled". Instead of typing the two words "to get vaccinated", you remove the space between these two words, and you get the only word "toget vaccinated".

There are cases where this type of error is combined with other types of editing errors, i.e. in the same word w, we have errors due to editing operations more than a space insertion(after inserting the space we have both words w_1 and w_2).

In this case, there are four cases:

- Both segments w_1 and w_2 are not changed. In this case the solution of the correction is simply w_1-w_2-w, for example "vacci nating instead of vaccinating".
- The first segment w_1 has been modified, and w_2 not, for example: "vaxi nating instead of vaccinating".
- w_2 has been modified and w_1 not, example of "vacci nathing instead of vaccinating".
- Both segments w_1 and w_2 are modified, for example: "vaxin ating instead of vaccinating".

In the last three cases we must stick the two segments w_1 and w_2 (w_1-w_2) and then correct the new merged word.

In the recently published work Yousfi and al.[30], the authors used and adapted the Levenshtein's algorithm to detect and correct errors due to space deletion between words in the case of Arabic texts.

In this paper, we will propose a new approach to correct space insertion errors, and to integrate it with other approach that which corrects deletion errors and other editing errors, in a single one that corrects these three types of error at the same time.

B. Introducing the Approach to Correcting Deletion Space Errors

Either w_{err} a erroneous word of length n, and w_i a word of the lexicon of p_i length. Among the errors in the wrong word w_{err} perhaps the space that is removed between several words more other types of editing errors (insertion, deletion, and permutation).

The approach is done in two stages:

- The detection of the position in the word w_{err} where space will be inserted.
- The correction of the two sequences obtained after the insertion of the space.

We note by: $D_L(w_{err}, w_i) = D_L(n, p_i)$, Levenshtein's distance between two words w_{err} and w_i. For the detection of the position where space will be inserted (noted $pos.space_i$), the authors gave the following rule:

$$pos.space_i = \underset{j=1,\ldots,n}{\arg\min}\ DL(j, p_i) \tag{1}$$

After the insertion of space, we get the two words w_{1space_i}, w_{2space_i}, and we move on to the second phase.

The second phase verifies whether the two words exist in the lexicon or not, otherwise we move on to the correction. To correct errors due to the deletion of spaces between words and other types of editing errors at the same time, the authors defined a new distance (noted D_{LS} based on the Levenshtein distance (edit distance):

$$\begin{aligned} D_{LS}(w_{err}, w_i) = Min[D_L(w_{err}, w_i); D_L(w_{err}, w_{1space_i}) \\ +space + w_{2space_i}] \\ where\, D_L\, is\, the\, edit\, distance \end{aligned} \tag{2}$$

The scheduling of solutions is based on this new D_{LS} distance.

In the sequel, we will present the new approach we propose in this paper to correct errors in the insertion of space inside word, and then we will show how to integrate it with the previous approach in a single one. Then all types of spelling errors can be corrected (editing errors, errors due to deletion and insertion of space).

IV. The Approach to Correcting Space Insertion Errors

This approach is based on the edit distance and on the bi-grams language models. In the rest of this paragraph, we will give a little reminder on these two concepts.

A. The Edit Distance

The metric method introduced by Levenshtein [5] measures the similarity between two words by calculating an edit distance. The edit distance is defined as the minimum number of basic editing operations required to turn an erroneous word into another word in the dictionary. Thus, to correct a wrong word, we retain a set of solutions requiring as few editing operations as possible.

The procedure for calculating the distance of Levenshtein between two strings $X = x_1x_2\ldots x_m$ of length m and $Y = y_1y_2\ldots y_n$ of length n, consists of calculating from near to near in a matrix of order $(m*n)$ the edit distance between the different sub-chains of X and Y.

The calculation of the case (i,j), which corresponds to the editing distance between the sub-chains $X_1^i = x_1x_2\ldots x_i$ and $Y_1^j = y_1y_2\ldots y_j$, is given by the following recurring relationship:

$$D(i,j) = Minimum \begin{cases} D(i-1,j)+1, \\ D(i,j-1)+1, \\ D(i-1,j-1)+cost \end{cases} \tag{3}$$

with

$$cost = \begin{cases} 0 & \text{if} \quad x_{i-1} = y_{j-1} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

The limitation of such a spelling correction system using the edit distance is that it does not allow a good scheduling of suggested solutions for a set of candidates with the same edit distance.

B. N-gram Language Models

The importance of language models is quite clear. They are used in several areas of NLP, such as continuous speech recognition, machine translation, etc. The main goal in these different applications is to have some solutions weighed against others.

A n-gram language model shows the fact that the probability of a word appearing after a sequence of words can be given only on the basis of the last $n-1$ words [29].
This model verifies:

$$Pr(w_i/w_1, w_2..., w_{i-1}) = Pr(w_i/w_{i-n+1}, .., w_{i-1}) \tag{5}$$

In practice, the value of n does not exceed order 3.

- If $n = 1$, the model is called a uni-gram model. This type of model does not take into account any history of the word.
- If $n = 2$, the model is called a bi-gram model. This type of model only takes into account the previous word: $Pr(w_i/w_1, w_2..., w_{i-1}) = Pr(w_i/w_{i-1})$
- If $n = 3$, the model is called a tri-gram model. This type of model takes into account only the previous two words: $Pr(w_i/w_1, w_2..., w_{i-1}) = Pr(w_i/w_{i-1}, w_{i-2})$

For the construction of n-gram language models, learning is done on a corpus of texts that must encompass all possible successions of words belonging to the vocabulary of the language used. This construction consists of estimating all the probabilities already mentioned.

V. Processing Space Insertion Errors

The processing of this type of error normally goes through the following two steps: the detection that one has an error due to the insertion of space into a correct or erroneous word, and the phase of correction.

A. Detection of Errors Due to the Space Insertion

Here we cite methods that are not 100% correct to detect errors in inserting space into a correct or erroneous word. We have cases where the probability of having a space insertion error is very high.

Among these cases we cite:

- If two successive words w_1 and w_2 are erroneous, then the probability is very high to have inserted a space that gave us these two wrong words. In this case, we concatenate the two words, and we treat $w_1 - w_2$ as a single word that must be submitted to

the correction procedure. However, this is not always the case as we can have two successive erroneous words without having a space insertion error ("The wolrd vaxinated against Covid" instead of "The world vaccinated against Covid"

- If we have a word consisting of a single character, then this may be due to a bad insertion of space into a word ("v accination" instead of "vaccination".
- If we have some kind of erroneous word and the following word is correct, it may be due to a space insertion into a word ("vacci nation" instead of " vaccination").

So from what we have presented, it is very difficult to fix cases and say that they are the only ones that exist for space insertion errors.

There are even cases where you insert a space into a word and you get two sequences that are both correct ("foot ball" instead of "football").

For this, we will treat the problem of space insertion according to two methods:

- **Method 1**: The space insertion error is only addressed if we have two successive words that are wrong, working with an H1-rated hypothesis that will be defined afterwards. If we have a single erroneous word, we correct it in a simple way.
- **Method 2**: Once we have a wrong word and regardless of the next word, plus the simple correction of that wrong word, we add the corrections of the error due to the wrong insertion of space between that wrong word and the two neighboring words.

B. Correcting Space Insertion Errors using Method 1

This method uses the following hypothesis:

If we have a correct word, then we should not change it.

Be a text $T = w'_1 w'_2 \dots w'_n$ consists of a set of words typed in that order, and suppose that we have two successive erroneous words w'_i and w'_{i+1}. To also take into account space insertion errors, and use the H1 hypothesis, all the corrections proposed for these errors consist of two sets:

- All simple corrections of the two words w'_i and w'_{i-1} (assuming we don't have errors in inserting or deleting space). This set is given by the following rule:

$$S_i^1 = \underset{w_k \in Lexique}{\arg\min} \frac{D_L(w'_i, w_k)}{Pr(w_k / w'_{i-1})} \tag{6}$$

and

$$S_{i+1}^1 = \underset{w_k \in Lexique, w_i \in S_i^1}{\arg\min} \frac{D_L(w'_{i+1}, w_k)}{Pr(w_k / w_i)} \tag{7}$$

- we Concatenate w'_i and w'_{i+1} (w'_i+w'_{i+1}), i.e. assume that we have inserted a space in the word $w'_i + w'_{i+1}$ which produced the two wrong words w'_i and w'_{i+1}. For correction, we check if $w'_i + w'_{i+1}$ is in the lexicon of the system. If so, we keep the word $w'_i + w'_{i+1}$ as an ideal solution with edit distance equal to zero; otherwise we make the correction with the edit distance weighted by the bi-gram language model:

$$S_i^2 = \underset{w_k \in Lexique}{\arg\min} \frac{D_L(w'_i + w'_{i+1}, w_k)}{Pr(w_k / w'_{i-2})} \tag{8}$$

To take into account space insertion errors during the correction operation, we propose the distance rated D_{AL} which is defined as a follow-up:

$$D_{AL}(w'_i, w_k) = Min[\frac{D_L(w'_i, w_k)}{Pr(w_k / w'_{i-1})} + \frac{D_L(w'_{i+1}, w_k)}{Pr(w_k / w'_i)}, \frac{D_L(w'_i + w'_{i+1}, w_k)}{Pr(w_k / w'_{i-2})}] \tag{9}$$

The best corrections are those that check :

$$\min_{w_k \in Lexique} D_{AL}(w'_i, w_k) \tag{10}$$

C. Correcting Space Insertion Errors using Method 2

For this method, we do not use the H1 hypothesis; and in this case, we can modify words that were correct. The list of solutions or corrections for the erroneous word w'_i proposed is of three types:

- The set of all simple corrections of the word w'_i (assuming we have no errors in the insertion or deletion of space). This set is given by the rule given in equation number (6)
- we concatenate w'_{i-1}, correct word, and w'_i ($w'_{i-1} + w_i$) and check if $w'_{i-1} + w'_i$ in the lexicon of the system. If so, we keep the word $w'_{i-1} + w'_i$ as an ideal solution with edit distance equal to zero, otherwise the correction is made with the edit distance weighted by the bi-gram language model :

$$S_i^2 = \underset{w_k \in Lexique}{\arg\min} \frac{D_L(w'_{i-1} + w'_i, w_k)}{Pr(w_k / w'_{i-2})} \tag{11}$$

- We concatenate w'_i and w'_{i+1} (w'_i+w'_{i+1}) and we check if w'_i+w'_{i+1} is in the lexicon of the system. If so, we keep the word w'_i+w'_{i+1} as an ideal solution with distance equal to zero; otherwise the correction is made with the edit distance weighted by the bi-gram language model:

$$S_i^3 = \underset{w_k \in Lexique}{\arg\min} \frac{D_L(w'_i + w'_{i+1}, w_k)}{Pr(w_k / w'_{i-1})} \tag{12}$$

To take into account space insertion errors during the correction operation, we apply the D_{AL} distance, which this time is defined as follows:

$$D_{AL}(w'_i, w_k) = Min[\frac{D_L(w'_i, w_k)}{Pr(w_k / w'_{i-1})}, \frac{D_L(w'_{i-1} + w'_i, w_k)}{Pr(w_k / w'_{i-2})}, \frac{D_L(w'_i + w'_{i+1}, w_k)}{Pr(w_k / w'_{i-1})}] \tag{13}$$

The best corrections of the wrong word w'_i are given by the following formula:

$$\underset{w_k \in Lexique}{\arg\min}\ D_{AL}(w'_i, w_k) = \underset{w_k \in S_i^1 \cup S_i^2 \cup S_i^3}{\arg\min} [\frac{D_L(w'_i, w_k)}{Pr(w_k/w'_{i-1})}, \frac{D_L(w'_{i-1} + w'_i, w_k)}{Pr(w_k/w'_{i-2})}, \frac{D_L(w'_i + w'_{i+1}, w_k)}{Pr(w_k/w'_{i-1})}] \quad (14)$$

Its solutions belong to one of the sets of S_i^1, S_i^2 and S_i^3 :

1) If a solution belongs to S_i^1, we correct w'_i by this solution, and we go to w'_{i+1} to check whether or not this word exists in the lexicon; otherwise we repeat our correction approach quoted here on w'_{i+1}
2) If a solution belongs to S_i^2, we removew'_{i-1}, we correct w'_i by this solution, and switchto w'_{i+1}. We check w'_{i+1} whether we exist in the lexicon or not, or we apply our correction approach to w'_{i+1}.
3) If the solution belongs to S_i^3, we correct w'_i with this solution, we delete w'_{i+1}, and we move to w'_{i+2}. We check whether w'_{i+2} exists in the lexicon or not, otherwise we apply our correction approach to w'_{i+2}.

When the correction is completed, the new sentences are scheduled with the original sentence, $w'_1 w'_2 \dots w_n$ by the edit distance taking the space as a character: $D_L(w'_1 w'_2 \dots w'_n, w'_1 w'_2 \dots, w'_{i-2}, w_{i-1} \dots, w_k) \quad k \leq n$

Example:
"The world vaccinated against Covid"
We have two successive erroneous words, so we do the processing according to the two methods.

Method 1:

- The corrections of the two erroneous sequences "vacc" and "inated" are "vacc" ={ vaccine, vice, Vicki, vaccines..}, and "inated"={noted, anted, named, mated, hated...}. The best corrections are distance 1, so the sum 1+1=2.

- The corrections of the word "vaccinated", after deleting space between sequences "vacc"and "inated" we obtain this word which is a lexicon entry.
- Applying the distance D_{AL} the min between 0 and 2 is 0 and like that, the best correction is "vaccinated" ("The world vaccinated against Covid")

Method 2:

- The first wrong word is "vacc", so the three sets without taking into account the language models in the formulas, S1, S2, and S3 are:
  - S1- corrections of the wrong word "vacc" ={ vaccine, vice, Vicki, vaccines..}
  - S2- corrections of the wrong word "worldvacc"={$\varnothing$}, in reality no suggestion
  - S3- the corrections of the word "vaccinated", this word is a lexicon entry.

TABLE I. RECALL AND ACCURACY OF DIFFERENT METHODS

	Recall	Accuracy
Method 1	82%	91%
Method 2	76%	88%

Words in S1 have a distance greater than or equal to 1, and words in S2 have a distance greater than or equal to 2. So the words that check the rule in equation (14) is "vaccinated" that is, the solution belongs to S3, so we correct "vaccinated" and we delete "inated", So the new sentence after the correction is: "The world vaccinated against Covid".

VI. TESTS AND RESULTS

The corpus on which we conducted our test is composed of 100 paragraphs taken from the Wikipedia site, and in each paragraph, we randomly created space insertion errors in words in those paragraphs in addition with other types of editing errors.

In total, we have 1000 errors due to the insertion of space into words that are correct or erroneous. These errors are created according to the four types of space insertion errors cited in subsection (3.1.)

In order to evaluate the different methods used for spelling correction, we use the classic evaluation measurements by calculating recall and accuracy. The results obtained are cited according to these four types of error, and are as mentioned the Table I. In order to test the robustness of our approach with these two methods, we compared it to a widely recognized commercial spell checker.

The first results approved that our approach achieves a very high correction rate compared to this corrector: 89.5 % of the suggested correction for errors due to the insertion against only 19.12 % for the commercial corrector. this can be interpreted by the fact that this corrector does not take into account the existence of this type of error due to the insertion.

CONCLUSION

According to the results obtained, we can say that our new proposed approach is an effective method to correct errors due to space insertion comparing with other commercial spell checker. In other hand, our approach present an acceptable and better complexity level compared with other approach based only on n-gram language models.

REFERENCES

[1] Kukich K.," Techniques for automatically correcting words in text ", ACM Computing Surveys (CSUR),Volume 24, Issue 4, pp: 377-439, 1992.

[2] Mitton R., "Spelling Checkers, Spelling Correctors, and the Misspellings of Poor Spellers ",in Information Processing & Management, Volum23, issue 5, pp: 495-505, 1987.

[3] Mitton R., " Spellchecking by computer ", in Journal of the Simplified Spelling Society, Volum 20, Issue 1, pp :4–11, 1996.

[4] Damerau F.J.," A Technique for computer detection and correction of spelling errors", Communications of the ACM, Volum 7, Issue 3, pp: 171-176, March 1964.

[5] Levenshtein V.I," Binary codes capable of correcting deletions, insertions, and reversals ", Soviet Physics Doklady, pp: 707-710, 1966.

[6] HORST B., "A Fast Algorithm for Finding the Nearest Neighbor of a Word in a Dictionary ". IAM-93-025, 1993

[7] Jaro M. A.," Advances in record-linkage methodology as applied to matching the census of Tampa", Journal of the American Statistical Association, pp : 414-420, Florida, 1985.

[8] Winkler W. E.," The State of Record Linkage and Current Research Problems ", Statistical Society of Canada, Proceedings of the Section on Survey Methods, pp: 73-79, 1999.

[9] Jaccard P.," The Distribution of the flora in the alpine zone", New Phytologist, Volume 11, pp: 37-50, 1912.

[10] Stoilos G., Stamou G., Kollias S., "A string Metric for Ontology Alignment ", International Semantic Web Conference, pp: 624-37, 2005

[11] Angell Richard C., Freund George E. et Willett P.," Automatic spelling correction using a trigram similarity measure", Information Processing and Management, Volume 19, Num. 4, pp: 255-261, 1983.

[12] Kernighan M.D, Church K.W. and Gale W.A.," A Spelling correction program based on a noisy channel model ", In Proceeding of the 13th Conference on Computational Linguistics, pp : 205-210, 1990.

[13] Pollock J. J. and Zamora A.," Automatic spelling correction in scientific and scholarly text", Communications of the ACM, Volume 27, Num.4, pp: 358-68, 1984.

[14] Oflazer K.," Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction", Computational Linguistics, Volume 22, Num. 1, pp: 73-98, 1996.

[15] Savary A.," Recensement et description des mots composés - méthodes et applications", Thèse de doctorat en Informatique Fondamentale, Université de Paris 7, pp : 149-158, 2000.

[16] Gueddah H., Yousfi A. and Belkasmi M.," Introduction of the weight edition errors in the Levenshtein distance", International Journal of Advanced Research in Artificial Intelligence, Volum 1, Issue 5, pp : 30-32, 2012.

[17] Gueddah H. et Yousfi A.," Impact de la proximité et de la similarité inter-caractère Arabe sur la correction orthographique ", Proceeding of the 8th International Conference on Intelligent Systems : Theories and Applications- SITA 13, pp : 244-246, EMI-Rabat 2013.

[18] Nejja M., Yousfi A., "A lightweight system for correction of Arabic derived words", in Mediterranean Conference on Information & Communication Technologies, Volum 1, pp: 131-138, Saïdia 2015.

[19] Bakkali H., Gueddah H., Yousfi A. and Belkasmi M.," For an Independent SpellChecking System from the Arabic Language Vocabulary ", Proceeding of International Journal of Advanced Computer Science and Applications, Volume 5 Issue 1, pp : 114-116, Janvier 2014.

[20] Nejja M., Yousfi A., " Contexts impact on the automatic spelling correction", in International Journal of Artificial Intelligence and Soft Computing archive, Volum 6, Issue 1, pp: 56-74, 2017.

[21] Gueddah H., Aouragh L., et Yousfi A.," Adaptation de la distance de Levenshtein Pour la correction orthographique contextuelle ", Proceeding du 9éme Conférence Internationale sur l'Intelligence Artificielle : Théories et Applications, SITA 14, pp : 242-245, INPT- Rabat 2014.

[22] Nejja M., Yousfi A., " Contexts impact on the automatic spelling correction", in International Journal of Artificial Intelligence and Soft Computing archive, Volum 6, Issue 1, pp: 56-74, 2017.

[23] Naseem T., Hussain S., "A novel approach for ranking spelling error corrections in Urdu ", Language Resources and Evaluation, Volum 41, Issue 2, pp: 117–28. DOI 10.1007/s10579-007-9028-6, Springer Science+Business Media B.V, 2007.

[24] Naseem T., Hussain S., "A Hybrid Approach for Urdu Spell Checking ", MSc thesis, National University of Computer & Emerging Sciences, Pakistan, 2004

[25] Attia M., Pecina P., Samih Y., Shaalan K., et Genabith J.V., "Arabic Spelling Error Detection and Correction ", in: Natural Language Engineering, Volum 22, Issue 5, pp: 751-773 Cambridge University Press, 2016.

[26] Maha M. A., William J. T., "Automatic Correction of Arabic Dyslexic Text". Computers 2019, Volum 8, Issue 1, 2019.

[27] Alkanhal M. I., Al-Badrashiny M. A., Alghamdi M. M., and Al-Qabbany A. O., "Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions ", in IEEE Transactions on Audio, Speech, and Language Processing Volum 20, Issue 7, pp: 2111–2122, 2012.

[28] Alexis Amid Neme, " Why Microsoft Arabic Spell checker is ineffective ", Linguistica Communication, http://www.al-erfan.com/, 2014, Arabic Language in Information Technology, 16, pp.55, hal01081965.

[29] G. Hirst, "An Evaluation of the Contextual Spelling Checker of Microsoft Office Word 2007 ", Department of Computer Science university of Toronto Toronto, Canada 2008.

[30] Yousfi A., Aouragh L., Gueddah H and Nejja M, " Spelling correction for the Arabic language: space deletion errors ", in Procedia Computer Science, pages: 568-574, Volum 177, 2020,ISSN 1877-0509.

New SARIMA Approach Model to Forecast COVID-19 Propagation: Case of Morocco

Ibtissam CHOUJA[1]
Engineering Sciences Laboratory
Faculty of Sciences
Ibn Zohr University
Agadir, Morocco

Sahar SAOUD[2]
Technical Research Laboratory
Faculty of Applied Sciences
Ibn Zohr University
Agadir, Morocco

Mohamed SADIK[3]
Engineering Sciences Laboratory
Faculty of Sciences
Ibn Zohr University
Agadir, Morocco

***Abstract*—The aim of this paper is to avoid any future health crises by analysing COVID-19 data of Morocco using Time Series to get more information about how the pandemic is spreading. For this reason, we used a statistical model called Seasonal Autoregressive Integrated Moving Average (SARIMA) to forecast the new confirmed cases, new deaths, cumulative cases and deaths. Besides predicting the spreading of COVID-19, this study will also help decision makers to better take the right decisions at the right time. Finally, we evaluated the performance of our model by measuring metrics such as Mean Squared Error (MSE). We have applied our SARIMA model for a forward forecasting in a period of 50 days, the MSE reported was 62196.46 for cumulative cases forecasting, and 621.14 for cumulative deaths forecasting.**

Keywords*—COVID-19, machine learning, seasonal autoregressive integrated moving average, SARIMA, statistical modeling, time series forecasting

I. Introduction

Since the announcement of sars-cov2 outbreak as a public health emergency by the world health organization (WHO) in January 2020, The world didn't stop counting its damage due to the spreading of this virus. The incessant stream of new unpredictable cases has caused crises in many countries. For instance, in Italy, the death toll caused by COVID-19 raised dramatically in a short period of time making it the biggest crisis since the Second World War. To control the crisis, many decisions have been made all over the world. However, all these efforts lacked efficiency since we didn't yet predict what the world needs in terms of tools and information to make better decisions, to end the pandemic as fast as possible, or at least prevent some of its dire consequences.

In Morocco, the first case was detected on March 2020, and by July 2021 the total cases reached 552635 confirmed cases and 9427 deaths according to the Moroccan health ministry [1]. The situation has become critical especially when COVID-19 reaches its peak at the critical phase. To manage this, the Moroccan government has taken many decisions regarding the timeline development of the disease, but as the pandemic hasn't been stopped yet, we can expect other peaks that should be predicted beforehand to minimize the consequences.

Many studies have been done to decrypt the spreading mechanism of the mysterious virus that has caused vast damage in different sectors. It was important for mathematicians, epidemiologists, and data analysts to analyze and model the phenomenon using various models and methods such as compartmental models [2, 3, 4, 5], machine learning and artificial intelligence models[6, 7], or statistical methods [8, 9, 10, 11].

In this paper, we present a new SARIMA model to forecast the new cases and deaths in Morocco to help decision-makers to prevent any critical phases. We have chosen to forecast COVID-19 using time series analysis. In Section II, we will present the related work. Then Section III will present the time series approach for forecasting, and the process of this study. Afterward in Section IV, we will describe the materials and methods we used in this work. Section V will be devoted to presenting the numerical results that will be discussed in Section VI. Finally, we will end this article with a general conclusion.

II. Related Work

A. Compartmental Models

Compartmental models are mathematical modelling of infectious diseases. One of the most important and commonly used of these models is the "Susceptible-Infectious-Recovered or Resistant" model known also as the SIR model, a simplest compartmental model based on the research of Anderson Gray McKendrick and William Ogilvy Kermack in 1927[12]. This model divides the population into three compartments or categories and calculates the transition rate between compartments. Other models had been extended from the SIR model like the SEIR[13] model dividing the population into four compartments instead of three, by adding a new one named "Exposed" representing people highly exposed to the infection.

Several studies have adapted the SIR and SEIR models to investigate the new coronavirus. Chen et al. [2] created the Bats-Hosts-Reservoir-People (BHRP), a transmission network model describing the transmission chain of the virus, from bats to people passing by the hosts and the seafood market, using the SEIR model. While Euloge Mouvoh et al. [3]described the impact of different intervention strategies on the spread of COVID-19 in Morocco using a contact-structured and age-structured-SEIR model considering the interventions dynamic between different age groups. On the other hand, Ben Hassen et al. [4]combined the Poisson Markov process with a SIR model to build a hybrid model called SIR-Poisson model to estimate deaths number at a certain date t. However, the SIR model failed to forecast COVID-19 in Isfahan, because the

assumptions that the SIR model is based on are not seemingly true for COVID-19 [5].

B. Artificial Intelligence and Machine Learning Models

In addition to compartmental modelization, Artificial Intelligence and machine learning models are highly recommended in such problems it can be used in various aspects. Ahmed et al. [6] suggested some problems related to the pandemic where Artificial Intelligence can act efficiently and expected that studies using those models will increase significantly when more COVID-19 data will be available. In this direction, Mbilong et al. [7] proved the efficiency of Deep Learning models by using six models of machine learning and deep learning to forecast COVID-19 in Morocco during 7 days.

C. Statistical and Time Series Methods

statistical methods have been used to forecast, identify and measure the damage due to the COVID-19 pandemic. Shang et al. [10]used time series analysis to identify any change points using excess in the counts of COVID-19 caused excess death in Belgium. Salgotra et al. [9] have chosen to predict COVID-19 in India using models based on genetic programming and evaluated by statistical methods and metrics. Alsulami et al.[11] presented a statistical method to study the effects of various factors on the deaths due to COVID-19. ArunKumar et al.[8]have forecasted COVID-19 cases using the statistical models SARIMA and ARIMA for 60 days in 16 countries.

In this paper, we have chosen to use the statistical model SARIMA to forecast cases in Morocco. In the next section, we will present the model and the process of this work.

III. COVID-19 Propagation by using Time Series Approach

Time series analysis is the process of analyzing a sequence of data recorded at a specific interval of time. As its name indicates time is a crucial variable in this process. It shows us how the problem studied is adjusted over time, and it provides an additional source of information extracted from the dependencies established between data and the time variable. Time series analysis is usually utilized for problems that are highly affected by time, it can be used to analyze and forecast trading data, weather data, electricity data, etc. With a representative sample and cleaned data, Time Series analysis can provide a lot of information about the phenomenon in question, or even forecast the future with high accuracy.
The Time series process requires 5 principal steps as shown in Fig. 1:

- **Data collection**: It's the first thing to do when using time series analysis, data can be collected by observation, experimentation , simulation or simply derived.
- **Data-preprocessing**: In this step, we prepare data collected to be used for forecasting, the preprocessing consists of cleaning data and handling the missing values.
- **Selection of Time series model**: Many time series models are available, but their efficiency depends on the data used for the study. In this step, we analyze data to extract its characteristics, to choose the best model for this study.
- **Model tuning**: It's a process that creates a combination of a manually specified subset of the hyperparameter space and evaluates each model to choose the best combination of those parameters. It's called also grid search.
- **Forecasting and interpretation**: After specifying the best parameters of our model, we can finally fit it and forecast our variables.

After explaining briefly the process of a time series analysis, we propose in the next section a comparison between two time series models known as Autoregressive Integrated Moving Average(ARIMA) and seasonal Autoregressive Integrated Moving Average (SARIMA).

A. Time Series Generalities (ARIMA VS. SARIMA)

The Autoregressive Integrated Moving Average model (ARIMA), is one of the models that are most widely used to forecast a time series. It describes the autocorrelations in data by combining the autoregressive model (AR); a linear regression between past and future values; and moving average model (MA) that considers the past forecast error and future values of data. The added value of ARIMA is the Integrated component (I) that describes the order of differentiation used to make data stationary [14].

ARIMA model depends on 3 parameters:
p: trend autoregressive order, **q**: trend difference order, and **d**: trend moving average order.

When time-series data present a seasonality pattern, we use the Seasonal Autoregressive Integrated Moving Average model (SARIMA) instead of ARIMA. SARIMA is an extension of ARIMA that can deal with seasonality to forecast seasonal Time series. It is composed of four parts. In addition to the three components of ARIMA described before, SARIMA is composed of seasonal components describing the seasonality pattern in data [14].

SARIMA depends on 4 other hyper-parameters (P D Q s) in addition to the ARIMA's parameters (p d q):

- **P**: Seasonal autoregressive order
- **D**: Seasonal difference order
- **Q**: Seasonal moving average order
- **s**: The number of time steps for a single seasonal period.

SARIMA model is described mathematically as :

$$\boxed{\varphi_p(B)\Phi_p(B^s)\nabla^d\ \nabla_s^D y_t = \ \theta_q(B)\ \Theta_{\mathcal{Q}}(B^s)\varepsilon_t} \qquad (1)$$

where:
y_t: is the forecast variable
$\varphi_p\,(B)$: is the regular AR polynomial of order p
$\theta_q\,(B)$: is the regular MA polynomial of order q
$\Phi_p\,(B^s)$: is the seasonal AR polynomial of order P

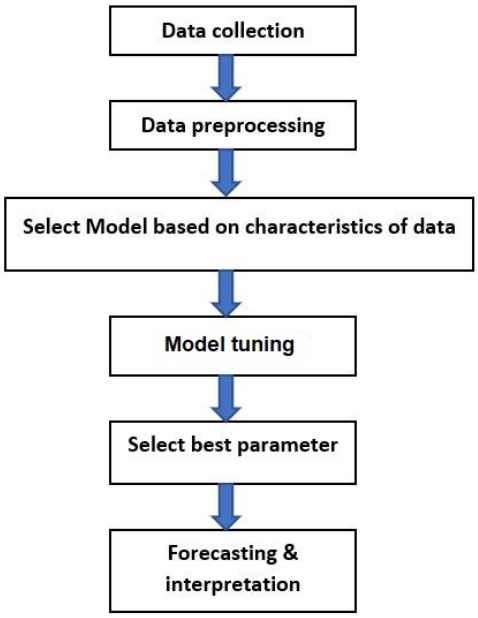


Fig. 1. SARIMA Process.

$\Theta_Q\,(B^s)$: is the seasonal MA polynomial of order Q
∇^d is the differentiation operator
∇_s^D is the differentiation operator
ε_t white noise process B back shift of y_t

To optimize our model and make sure that the parameters chosen are optimum, we used a process that creates a combination of a manually specified subset of the hyper-parameters space and evaluates each model to choose the best combination of those parameters. This process is known as grid search or model tuning.

IV. Materials and Methods

A. Methods and Data Preprocessing

We collected data from [15], it is a recording of corona cases in the whole world, for the period going from 24/02/2020 to 18/07/2021 with 103348 rows and 59 columns.

The data preprocessing in our case consists of creating new data for Morocco by using the global data collected, then cleaning this one by detecting the aberrant values in order to correct them, and finally dealing with the missing values by deleting some useless rows or filling them by fitting values. The specificity of time series data preprocessing is that we should also handle the trend and stationarity by de-trending data and making it stationary especially when using models like ARIMA or SARIMA.

The final dataset is new data recorded from 02/03/2020 (the date of the first case in Morocco) to 18/07/2021 with 504 rows (observations) and 29 columns. The new dataset is shown in Fig. 2

The new data was divided into two categories:

- Training data: is used to train and fit the model to find the best parameters.
- Testing data: is used for forecasting and assessing the performance of our model.

B. Time Series Analysis

In this section, we analyse the components of our data to extract the different characteristics of our time series data.

1) Trend: The first thing to check on time series data is the presence of a long-term movement called a trend in a cyclical context [16].

By plotting the data (Fig. 3), we observe the presence of a trend, but this should be proved by some tests. To test the presence of a trend in this time series, we used "pyMannkendal" a Python implementation of nonparametric Mann-Kendall trend analysis .

The mann-kendall test confirmed the existence of an upward trend on the data of new cases and new deaths with a significant p-value of p=5.329070518200751e-15 for new cases and 1.4821224705308111e-05 for new deaths. Before applying any model of forecasting, we should de-trend our data. There are many ways to do this, the most common method is differentiation.

2) Stationarity: Stationarity is one of the most important things to check before dealing with a time series. Stationary data means that the variance and the mean are constant over time. In other words, the data is stable [17]. However A non-stationary data is unpredictable data unless we make it stationary which is possible using some transformations. To test stationarity, we used two tests ADF and KPSS from the statsmodels package:

date	total_cases	new_cases	total_deaths	new_deaths	reproduction_rate	new_tests	total_tests	tests_per_case	total_vaccinations	people_vaccinated	...
2020-03-02	1	1	NaN	NaN	NaN	NaN	29.0	NaN	NaN	NaN	...
2020-03-03	1	0	NaN	NaN	NaN	4.0	33.0	NaN	NaN	NaN	...
2020-03-04	1	0	NaN	NaN	NaN	2.0	35.0	NaN	NaN	NaN	...
2020-03-05	2	1	NaN	NaN	NaN	7.0	42.0	NaN	NaN	NaN	...
2020-03-06	2	0	NaN	NaN	NaN	10.0	52.0	NaN	NaN	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
2021-07-14	547273	2257	9404.0	9.0	1.53	23516.0	6547425.0	12.6	20343869.0	10854278.0	...
2021-07-15	549844	2571	9418.0	14.0	NaN	23535.0	6570960.0	11.7	20794848.0	11185961.0	...
2021-07-16	552635	2791	9427.0	9.0	NaN	NaN	NaN	NaN	NaN	NaN	...
2021-07-17	555488	2853	9434.0	7.0	NaN	NaN	NaN	NaN	20833568.0	11213841.0	...
2021-07-18	557632	2144	9450.0	16.0	NaN	NaN	NaN	NaN	20840340.0	11219675.0	...

504 rows × 29 columns

Fig. 2. Moroccan Data.

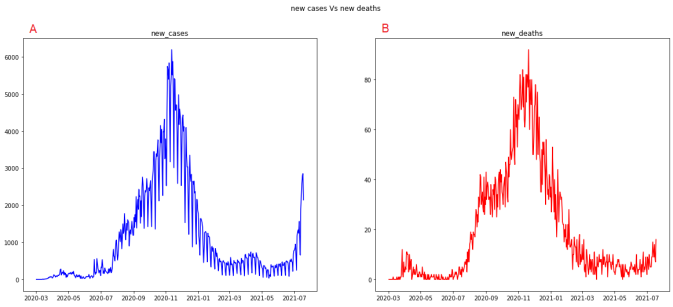


Fig. 3. A- Daily New Cases of COVID-19 in Morocco. B- Daily New Deaths due to COVID-19 in Morocco.

Fig. 4. A- Differencing Results of Original Series of New Cases. B- Differencing Results of Original Series of New Deaths.

- Augmented Dickey-Fuller test or ADF test: in which the null hypothesis admits the existence of unit root which means that the data is non-stationary [18].
- Kwiatkowski Phillips Schmidt Shin or KPSS test: unlike ADF the null hypothesis that we are testing in the KPSS is that our data is stationary [19].

The results of these two tests are shown in Table I.

Both ADF and KPSS tests confirm the non-stationarity of our data for the reasons below:

- For ADF:
  - The p-value is higher than the significance level of 0.05.
  - The ADF statistical test is higher than all the critical values.
- For KPSS:
  - The p-value is less than the significance level of 0.05.
  - The KPSS statistical test is higher than all the critical values.

It's indispensable to make data stationary before using it on a forecasting operation to make it predictable and to get relevant results. In our case, we have chosen the differentiation method to make it stationary. Fig. 4A and Fig. 4B show the data after 1st and 2nd order differencing, we can observe that our data had become stationary with the first differencing order and there is no need for a second one. To prove the stationarity of the new data, we have retested it with KPSS and ADF.

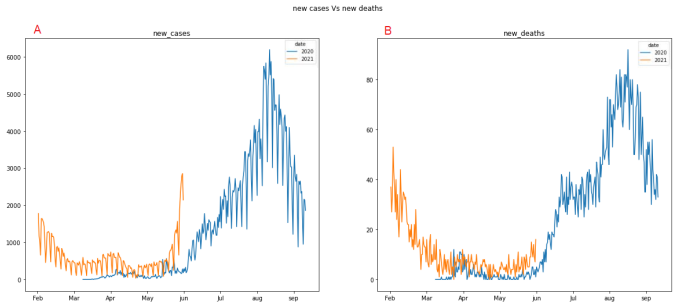


Fig. 5. A- Daily New Cases per Year in Morocco. B- Daily New Deaths per Year in Morocco.

3) Seasonality: The last component that we should verify is seasonality which means the presence of a pattern. In the Fig. 5A and Fig. 5B, we observe the existence of some pattern between 2020 (year of the appearance of COVID-19 In Morocco) and 2021. For example, the number of new cases increases between June and August, probably because of the school and summer vacations in which most people travel a lot and don't respect restriction policies. In contrast, the number of new cases decreases between November and February, and it is somehow stationary between February and January.

TABLE I. KPSS VS. ADF TEST RESULTS

		KPSS				ADF		
New cases	**Stastical test**	0.449026				-1.813390		
	p-value	0.01				0.373834		
	critical values	10 %	5 %	2.5 %	1 %	10 %	5 %	1 %
		0.119	0.146	0.176	0.216	-2.570297	-2.868161	-3.445368
New deaths	**Stastical test**	0.449026				-1.813390		
	p-value	0.01				0.373834		
	critical values	10 %	5 %	2.5 %	1 %	10 %	5 %	1 %
		0.119	0.146	0.176	0.216	-2.569954	-2.867518	-3.443905

V. NUMERICAL RESULTS

In this section, we present the results of COVID-19 daily new cases, new deaths, cumulative cases, and cumulative deaths forecasting using our proposed models.

After the model tuning, we have chosen the best parameters for each model. So we selected SARIMA (5,1,0)(1,0,1,7) to forecast new cases, and SARIMA (0, 1, 1) (2, 0, 2, 7) for the new deaths forecasting. The Fig. 6A shows the prediction of daily new cases, while figure Fig. 6C shows the prediction of daily new deaths for 104 days.

To measure the performance of our models, we calculated the Mean Equared Error(MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) for 104 days in Table II, and we have compared our results with other persistent models.

In Fig. 6B and Fig. 6D we noted that the model fits very well at the first 50 days of forecasting and start deviating after, so we recalculated the errors for 50 days. Table III shows the metrics reported for the 50 days forecasting.

VI. DISCUSSION

In the present study, we developed a SARIMA model to fit the reported data of COVID-19 in Morocco. The proposed model was prepared to select the best parameters in order to predict cumulative cases and deaths, new daily cases, and deaths.

The results show that the proposed models fit very well at the first 50 days with a reported MSE of 316960.2267400527 for new daily cases forecasting, 8.28 for new daily deaths, 62196.46 for cumulative cases, and 621.14 for cumulative deaths. The model starts to deflect after 50 days and the error starts to become higher. We tested our model for a forecasting period of 104 days and have reported an MSE of 4881.384615384615 for cumulative deaths forecasting, and an MSE of 108689697.75 for cumulative cases.

The 104 days predictions results are good if we compare them to most of the reported results in [8] for 60 days forecasting in 16 countries, where the MSE variates between 6.63E+04 for cumulative cases in Bangladesh and 2.69E+09 for Brazil, while the MSE of cumulative deaths forecasting for the same number of days variates between 3.10E+00 in Spain and 2.24E+06 in India.

VII. CONCLUSION

Since the outbreak of COVID-19 in Morocco, the government has made many decisions regarding the real-time situation, therefore forecasting the COVID-19 impact could help in preventing any critical situation by making a well-informed decision.

In this paper, we used time series analysis, one of the most important aspects of data analytic to analyze Moroccan data by understanding components like trend and seasonality. The ultimate objective of this work was to predict cases (new confirmed cases, new deaths, cumulative cases, and cumulative deaths) for 104 days using the SARIMA model. The results of our proposed models matched with test data especially for the first 50 days of forecasting for both cumulative cases and deaths. Decision-makers should consider forecasting models like SARIMA in preventing any future critical situation.

REFERENCES

[1] "Statistique of covid-19 in morroco," http://covidmaroc.ma/Pages/AccueilAR.aspxhttp://covidmaroc.ma/Pages/AccueilAR.aspx, accessed: 2021-07-16.

[2] T.-M. Chen, J. Rui, Q.-P. Wang, Z.-Y. Zhao, J.-A. Cui, and L. Yin, "A mathematical model for simulating the phase-based transmissibility of a novel coronavirus," *Infectious Diseases of Poverty*, vol. 9, no. 1, 2020.

[3] A. C. Euloge Mouvoh, A. Bouchnita, and A. Jebrane, "A contact-structured SEIR model to assess the impact of lockdown measures on the spread of COVID-19 in Morocco's population," in *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. Kenitra, Morocco: IEEE, Dec. 2020, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9314462/

[4] H. Ben Hassen, A. Elaoud, N. Ben Salah, and A. Masmoudi, "A SIR-Poisson Model for COVID-19: Evolution and Transmission Inference in the Maghreb Central Regions," *Arabian Journal for Science and Engineering*, vol. 46, no. 1, pp. 93–102, Jan. 2021. [Online]. Available: http://link.springer.com/10.1007/s13369-020-04792-0

[5] S. Moein, N. Nickaeen, A. Roointan, N. Borhani, Z. Heidary, S. H. Javanmard, J. Ghaisari, and Y. Gheisari, "Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan," *Scientific Reports*, vol. 11, no. 1, p. 4725, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41598-021-84055-6

[6] A. Ahmed, P. Boopathy, and Sudhagara Rajan S., "Artificial Intelligence for the Novel Corona Virus (COVID-19) Pandemic: Opportunities, Challenges, and Future Directions," *International Journal of E-Health and Medical Communications*, vol. 13, no. 2, pp. 1–21, Jul. 2022. [Online]. Available: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJEHMC.20220701.oa5

[7] P. M. Mbilong, A. Berhich, I. Jebli, A. El Kassiri, and F.-Z. Belouadha, "Artificial Intelligence-Enabled and Period-Aware Forecasting COVID-19 Spread," *Ingénierie des systèmes d information*, vol. 26, no. 1, pp. 47–57, Feb. 2021. [Online]. Available: http://www.iieta.org/journals/isi/paper/10.18280/isi.260105

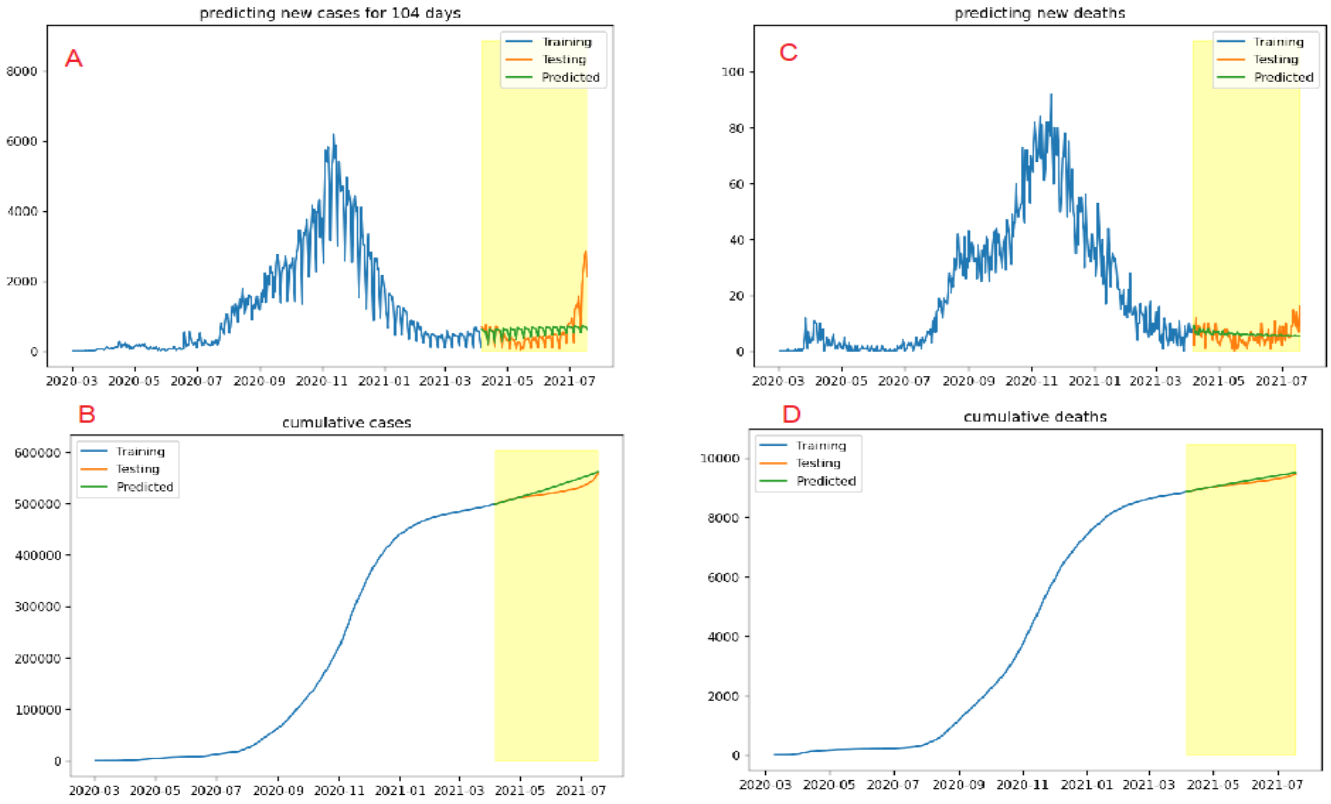


Fig. 6. 104 Days Forcasting. A-Forecasting of Daily New Cases. B- Forecasting of Cumulative Cases.C-Forecasting of Daily New Deaths. D- Forecasting of Cumulative Deaths.

TABLE II. 104 DAYS FORECASTING ERRORS

	MSE	MAE	RMSE
New cases	254695.1826923077	326.72115384615387	504.67334256160956
New deaths	10.192307692307692	2.4615384615384617	3.1925393799149435
Cumulative cases	108689697.75	8267.288461538461	10425.435134803727
Cumulative deaths	4881.384615384615	56.73076923076923	69.86690643920493

TABLE III. 50 DAYS FORECASTING ERRORS

	MSE	MAE	RMSE
New cases	316960.2267400527	549.3631462639955	562.9922084186003
New deaths	8.28	2.16	2.8774989139876315
Cumulative cases	62196.46	203.38	249.39218111239975
Cumulative deaths	621.14	17.78	24.922680433693323

[8] K. ArunKumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)," *Applied Soft Computing*, vol. 103, p. 107161, May 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7869631/

[9] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming," *Chaos, Solitons & Fractals*, vol. 138, p. 109945, Sep. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960077920303441

[10] H. L. Shang and R. Xu, "Change point detection for COVID-19 excess deaths in Belgium," *Journal of Population Research*, Mar. 2021. [Online]. Available: http://link.springer.com/10.1007/s12546-021-09256-2

[11] M. D. Alsulami, H. Abu-Zinadah, and A. H. Ibrahim, "Machine Learning Model and Statistical Methods for COVID-19 Evolution Prediction," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–6, Dec. 2021. [Online]. Available: https://www.hindawi.com/journals/wcmc/2021/4840488/

[12] G. McKendrick, "A contribution to the mathematical theory of epidemics," p. 22.

[13] M. Y. Li and J. S. Muldowney, "Global stability for the SEIR model in epidemiology," *Mathematical Biosciences*, vol. 125, no. 2, pp. 155–164, Feb. 1995. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0025556495927565

[14] A. Atangana and S. İğret Araz, "Modeling and forecasting the spread of COVID-19 with stochastic and deterministic approaches: Africa and Europe," *Advances in Difference Equations*, vol. 2021, no. 1, p. 57, Dec. 2021. [Online]. Available: https://advancesindifferenceequations.springeropen.com/articles/10.1186/s13662-021-03213-2

[15] "World data of covid-19," https://ourworldindata.org/coronavirus/country/moroccohttps://ourworldindata.org/coronavirus/country/morocco, accessed: 2021-07-18.

[16] M. Hussain and I. Mahmud, "pyMannKendall: a python package for non parametric Mann Kendall family of trend tests." *Journal of Open Source Software*, vol. 4, no. 39, p. 1556, Jul. 2019. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01556

[17] Badi H. Baltagi, "Time-Series Analysis," in *Econometrics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 355–377. [Online]. Available: https://doi.org/10.1007/978-3-540-76516-5_14

[18] R. Harris, "Testing for unit roots using the augmented Dickey-Fuller test," *Economics Letters*, vol. 38, no. 4, pp. 381–386, Apr. 1992. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/016517659290022Q

[19] D. Lee and P. Schmidt, "On the power of the KPSS test of stationarity against fractionally-integrated alternatives," *Journal of Econometrics*, vol. 73, no. 1, pp. 285–302, Jul. 1996. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0304407695017410

Text to Image GANs with RoBERTa and Fine-grained Attention Networks

Siddharth M, R Aarthi
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

***Abstract*—Synthesizing new images from textual descriptions requires understanding the context of the text. It is a very challenging problem in Natural Language Processing and Computer vision. Existing systems use Generative Adversarial Network (GAN) to generate images using a simple text encoder from their captions. This paper consist synthesizing images from textual descriptions using Caltech-UCSD birds datasets by baselining the generative model using Attentional Generative Adversarial Networks (AttnGAN) and using RoBERTa pre-trained neural language model for word embeddings. The results obtained are compared with the baseline AttnGAN model and conduct various analyses on incorporating RoBERTa text encoder concerning simple encoder in the existing system. Various performance improvements were noted compared to baseline Attention Generative networks. The FID score has decreased from 23.98 in AttnGAN to 20.77 with incorporation of RoBERTa model with AttnGAN.**

***Keywords*—Natural language processing; computer vision; GANs; AttnGAN; RoBERTa**

I. Introduction

Text to Image generation is an application of Generative Networks. The underlying problem comprises recognising the context of the text description and generating a realistic image that matches the caption. It is a multi-modal problem that challenges natural language Processing for context understanding and Computer Vision for Images. There are numerous applications in arts and design and have advanced considerably in recent years. Text to image generation can assist game developers to generate more distinct characters or skins with ease. Artists could use them to create starter comics from descriptions of the scene.

GANs are primarily used as generative networks for Image Synthesis from text and use Deep Convolutional GANs [1]. Recently, enormous progress has been made in synthesising images from texts for a single class of datasets like the Caltech-UCSD Birds-200-2011 dataset [2] or Oxford 102 Flower dataset. This paper uses the baseline AttnGAN model [3] that make use of Long Short term Memory for processing the text description. The latest language models developed lately has proved to be very efficient for text related problems. Thus it is required to utilize the latest transformer models so that better attention can be obtained within the natural language and hence increase the overall performance of the problem.

This problem can be divided into different parts and approached individually as a module. Text description is taken with details as input. The description includes features of its appearance, like the colour of particular body parts and its length. Text like, "this small bird has a short, pointy orange beak and white belly", are provided as input. The RoBERTa language neural model [4] to capture attention and understand the context of the description.

Fig. 1. Result of a Generated Bird from the Generative Model using RoBERTa Text Encoder and the Attention Captured during the Generation of Image at Epoch 600.

RoBERTa language model is used for embeddings words into a feature representation. Transform models use an attention mechanism to capture critical details associated with the word, and they can link them using the attention heads [5]. AttnGAN are used to train the generative networks. With each stage, higher resolution images are synthesized. Another component used in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). The attention mechanism and the DAMSM are used to find the similarity between the image generated by the GANs and the sentence using both the global sentence level and the fine-grained word-level information. The DAMSM component provides a fine-grained image and text comparison loss that can be used to train the generator [3].

The goal is to explore the current state-of-the-art model that can generate images based on the description from text using the Attention mechanism. RoBERTa language model is incorporated and experiments are performed on how the

existing system is affected. The same is analyzed and compare each result using the Fréchet inception distance (FID) score obtained from the base model. The focus is on the CUB Bird dataset for this paper, and the dataset also provides us with boundary box segmentation of bird images. Segmentation [6,7,8] can help train the model for generating a specific object in the boundary of images used from training . Analysis also performed on how RoBERTa embeddings have an effect in an interval of epochs on generating the images. The models are implemented using deep convolutional neural networks because they enhance the processed image [9,10]. The final output of the generative network is a high-resolution image matching the text description. Various attentions mapped are recorded with the experiments conducted and we obtain results as in Fig. 1.

The major benefits on exploring research on this problem lay in understanding the use of transformers with basic attention based generative system. This will help us explore how latest language models that uses attention heads can help understand the text association with the image better. A larger intuition could be developed in natural language association with text generation and scene prediction. This will help artists, game developers, animation industries to develop characters based on the textual description provided by the artists.

II. Related Work

Generative Adversarial networks are generative models that can synthesize images and are used for generative learning. In this network, a generator generates images based on the input from the embeddings and the noise. The discriminator help discriminate the picture as real or fake. Both generator and discriminator improve over time. The generator aims to generate images to fool the discriminator into thinking and classifying the images as real. These models were the first approach in generative networks [11] by Ian Goodfellow. There have been a variety of works on generative networks, and GANs have been able to generate photorealistic images with very high resolutions lately. Lately, generating images from text descriptions has been an area of research, and have few novel approaches to this problem.

The first approach to this problem was synthesizing low-resolution images from captions using Deep Convolutional GANs [1]. This system however could not completely produce image that looked realistic enough. Many images that were synthesised didn't exactly match the description either. This lead the author [12] to introduce Generative Adversarial What-Where Network (GAWWN). It exposed the control with the object's bounding box in the image and focused on particular parts. It modelled the distribution on various components like the tail and beak to obtain efficient results by focusing on that area. This proved to help identify key objects for generation but couldn't exactly focus on key details of objects like its poses or structure. A conditional Pixel Convolutional Neural Network (CNN) was used to synthesize images from text and used a multi-scale model structure. An image closer to the text description could be generated and a starting point to research the text to image using GANs. The image quality generated was an issue motivating Stack GANs, which used a stacked approach to improve the image resolution in different stages and generated 256 x 256 sized photorealistic images. The initial model was able to generate 64 x 64 resolution images. This approach generated an initial 64 x 64 images and was trained with GANs in two stages to get 128 x 128 images in Stage-I and 256 x 256 images in Stage-II. Each stage had an aim to improve the image quality to gain a photorealistic effect. StackGANs could generate photorealistic birds and flower images [13].

While this generated photorealistic images, proper context extraction was lacking. It paved the way to AttnGAN, which took a new approach by using an attention mechanism from the text description. Attention capturing helped the network to a closer understanding of context and better generation of images from text. AttnGAN used word embeddings and could capture important words from descriptions of birds [2] in the Caltech-UCSD Birds-200-2011 dataset. The attentional generative network could synthesize fine-grained details at various image subregions by providing detailed attention to the relevant words provided in the text description. This paper also introduced a deep attentional multimodal similarity model (DAMSM), used as the loss function and matched with text description and the generated image features. Word level condition selection was introduced to synthesize image details [3]. AttnGAN made use of bidirectional LSTM for the natural language processing. Similar to this work is the Controllable Text-to-Image Generation, which is used to synthesize high-quality images effectively by controlling parts of the image generation concerning natural language descriptions. In addition to the attention mechanism followed in AttnGAN, this paper used a channel-wise attention module and a word-level discriminator. It adopted a perceptual loss [14] in the text-to-image synthesis. Experimental researches have been performed by updating the architecture within the GANs by connecting generated image with the input description. The method of redescription was performed in MirrorGANs and Cycle GANs [15,16] using the BERT language model, where the authors obtained a great performance enhancement on complex datasets. Lately transformers have enhanced the neural language processing and is widely used in most of the latest intelligent systems. While the works performed till now has shown great result, it is very important to understand how these models could perform with the latest transformer models. This lead us to using latest neural language model RoBERTa as a pretrained model and incorporate it with the AttnGAN network instead of using basic LSTM system at language processing end. The aim of this paper is to analyze how well the AttnGAN model improves its performance using this system.

III. Dataset

Table I shows the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [2] image dataset that contains 200 categories of birds were used for experiments. There are a total of 11,788 images with annotations. The images and annotations together size up to 1.1 gigabytes. This dataset is used as a benchmark dataset for all the text to image synthesis research works. The images along with them have boundary boxes provided and are of various sizes each.

This dataset [2] contains images of North American birds from 200 different species of ranges. The dataset (CUB-200) was created in 2010 and contains approximately 6000 photos of each of the 200 bird types. Additional label data, such as

TABLE I. STATISTICS OF DATASETS

Dataset	Train	Test
No: of Samples	8,855	2,933
No: of Captions	10	10

bounding boxes, crude segmentations, and additional features, accompanied this. The dataset was updated in 2011 (CUB-200–2011) to include new photos, bringing the total number of images in the dataset to around 12,000 (CUB-200–2011). 15 component locations, 312 binary attributes, and a bounding box per image were added to the accessible attributes . The photos and class labels will be used to create and train networks for predicting bird class for the majority of this series.

IV. METHODS

A. Attention GANs

Fig. 2 shows the architecture of the AttnGAN networks with RoBERTa neural language model. AttnGAN make use of an attention mechanism that embeds the generated caption from the Birds dataset and run through the RoBERTa model to generate word and sentence vectors. The text encoder takes the caption, which is a T words sentence. The sentence features contribute to the global vector, which is passed on to the noise vector. The sentence feature is the final hidden state with a dimension D.

$$\bar{e} \in R^D \tag{1}$$

Similarly, word features are extracted separately. It produces a hidden state from all timesteps for the T-word sentence.

$$e \in R^{DXT} \tag{2}$$

The conditioning augmentation has the randomly sample latent variables from the Gaussian distribution. So $\bar{e}$, which is received as an input to caption feature, is split into μ and σ with a fully connected linear layer. This is the mean and variance from the sentence embedding. The mean and variance generated are used to parameterize the normal distribution from which a sentence embedding sample gets generated to be passed on to the generative network.

It is combined with a noise vector so that the generated images show higher variation for a single caption. The c vector is concatenated with the Z noise vector, and this is used in further stages for the generation of various features of birds in the network. Similarly, word features are extracted separately. It produces a hidden state from all timesteps for the T-word sentence.

$$\begin{aligned} &\bar{e} \longrightarrow \mu, \sigma \\ &c = \mu + \sigma * \varepsilon, \varepsilon \sim N(0, I) \end{aligned} \tag{3}$$

The first generative network is mainly responsible for upsampling. The nearest neighbour interpolation is used to upsample with a scaling factor of 2. The output is generated with a 64 X 64 image. This stage does not use any word-level features that are extracted using the RoBERTa model. It utilizes the sentence level features, which is taken as input from the noise vector space.

$$\begin{aligned} &h \in R^{\hat{D} \times N} \\ &h_0 = F_0\left(z, F^{ca}(\bar{e})\right) \end{aligned} \tag{4}$$

The first attention model combines word features e, with the previous stage context h_{i-1}. The word features before combination are brought into a common space. This is represented using e' and obtained by adding a new perceptron layer. $e' = Ue$, where $U \in R^{\hat{D} \times D}$. Each column of h is a feature vector of a sub-region of the image.

$$s'_{j,i} = h_j^T e'_i \tag{5}$$

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \text{ where } \beta_{j,i} = \frac{\exp\left(s'_{j,i}\right)}{\sum_{k=0}^{T-1} \exp\left(s'_{j,k}\right)} \tag{6}$$

Combining them with the context, it generates a score for a particular sub-region j, and a word i. So a combination is brought out with a particular word with a sub-region and it's used for the word-context vector for that region. This process is repeated for each region. This provides us with the output of the attention network.

$$F^{attn}(e, h) = (c_0, c_1, \ldots, c_{N-1}) \in R^{\hat{D} \times N} \tag{7}$$

The second generator also is used for upsampling of the image and it obtains an image of 128 X 128. Here, along with the previous output as input from the first generator which carries the context vector, the word embeddings through the attention networks are also added which carries the word context vectors. The residual blocks here, make the network deeper and train them without degradation. Similarly, one more generator was used to upscale the image up to 256 X 256 and it takes input similar to that of the second generator.

In the end, 256 X 256 image is passed to an image encoder. In the image encoder, local image features can be extracted and this is converted to a common space to match the text encoder feature. These two are combined to make the Deep Attentional Multimodal Similarity Model (DAMSM) and this is trained with attention loss. The DAMSM model is pre-trained for stability in the system.

There are three discriminators each attached with its respective generators. The sentence level features is taken without noise vector as input to each discriminator. Two forms used in the network is the unconditional form that tells if the image is real or fake and the conditional form that tells if the image and caption are of the same pair. In unconditional pair, a result close to 1 is obtained if both the pair are matching.

1) Text Encoder: RoBERTa makes use of transformers that has attention mechanism which learns the contextual relationship between the words within the sentences.

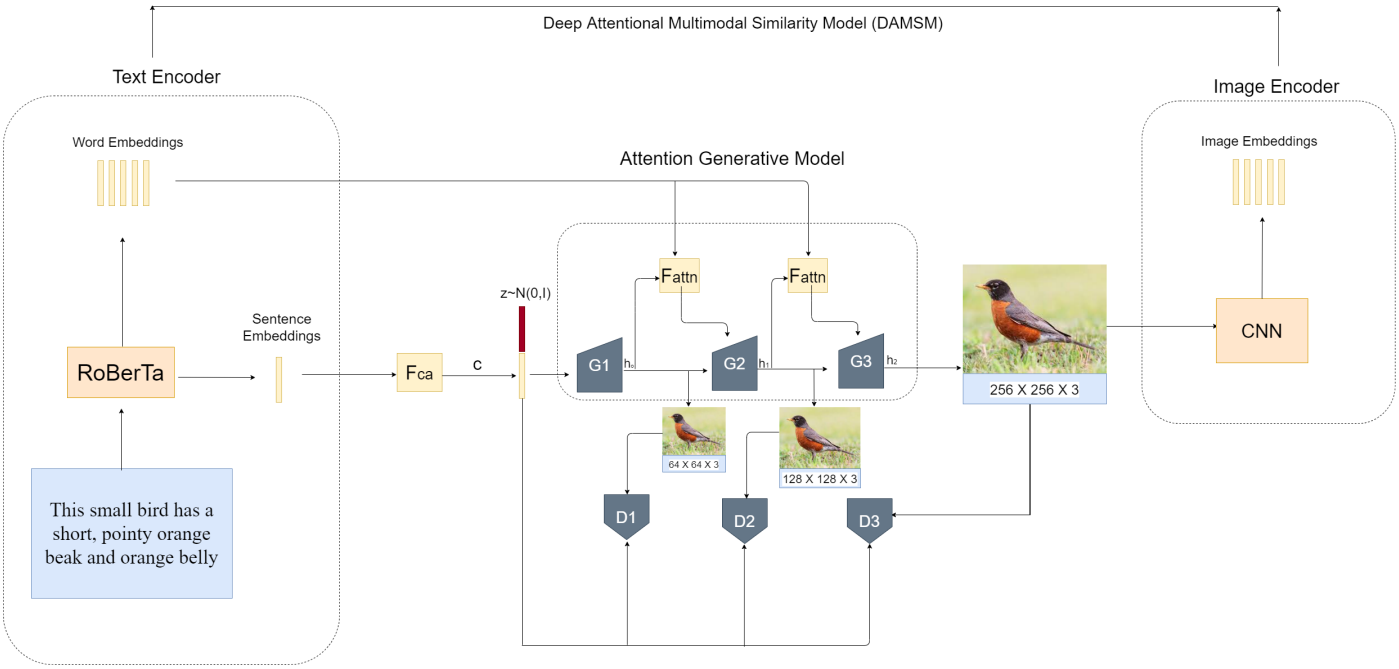


Fig. 2. Architecture of the Proposed System. RoBERTa Text Encoder is for Word Embeddings passed to Generative Network with Fine Grained Attention Networks; Text-Image Matching Loss Generated with DAMSM for the Generative Networks.

2) Image Encoder: The Image encoder is used with DAMSM as a convolutional neural network to extract the features out so that it can map to a common space. With CNN, the intermediate layers can learn various features associated with the different sub-regions of the image and the latter learns about the global features associated with them. A pre-trained Inception-v3 model on ImageNet was used as the image encoder. 768 is the local features dimension and it resizes the image to 299 X 299 pixels, to get 289 sub-regions in the image. In the end, these features are converted to similar space to that of the text encoder by adding perceptron layers.

3) Loss: For every generation, G_i a discriminator D_i and the loss is a combination of both conditional and unconditional at each stage. The embeddings of sentences is being conditioned on. The unconditional loss brings the generated images sampled from the generator of the particular distribution and is passed to the discriminator. The loss is minimized here so that the discriminator is fooled to think the image coming is real. For the conditional loss, passed $\bar{e}$ along with the generated image to the discriminator.

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(D_i\left(\hat{x}_i\right)\right]\right.}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(D_i\left(\hat{x}_i, \bar{e}\right)\right]\right.}_{\text{conditional loss}} \tag{8}$$

The discriminator uses cross-entropy loss and has data from the original distribution and the generated distribution. The discriminator will try to bring the original distribution close to 1 and the generated images output close to 0 to minimize the discriminator loss.

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}E_{x_i \sim p_{\text{data}_i}}\left[\log D_i\left(x_i\right)\right] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{x}_i\right)\right)\right] +}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2}E_{x_i \sim p_{\text{data}_i}}\left[\log D_i\left(x_i, \bar{e}\right)\right] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{x}_i, \bar{e}\right)\right)\right]}_{\text{conditional loss}} \tag{9}$$

B. RoBERTa

Attention GANs use basic RNN, which is a bidirectional LSTM. LSTM is used on the text description to extract the semantic vectors. With bi-directional LSTM, each word corresponds to two hidden states representing one for each direction [3]. RoBERTa: A Robustly Optimized BERT Pre-training Approach is the latest language model introduced by Facebook that optimizes the existing BERT architecture. It introduces the dynamic masking, hence the masked token changes during the training epochs. RoBERTa uses 160 GB of text for pre-training, including large Books Corpus and English Wikipedia are used in BERT. The additional data included CommonCrawl News dataset, Web text corpus and Stories from Common Crawl. The RoBERTa makes use of similar architecture as BERT Model but uses the byte-level BPE as the tokenizer [4]. The 'roberta-base' is the model used for prediction. The model was trained with an embedding dimension of 768. The pre-trained RoBERTa model is used to obtain the word and sentence embeddings and pass them with a fully connected layer before remaining in the Attention GANs architecture. The pre-trained model is 12 layered with

12 heads for the attention mechanism of the transformer and has around 125M parameters.

C. Deep Attentional Multimodal Similarity Model

Deep Attentional Multimodal Similarity Model (DAMSM) verifies if the generated image follows the description. It accompanies various steps to check this and update the network. The image features are brought f and $\bar{f}$ into a common space by adding a perceptron layer. The dimmension D is similar to the text encoders dimmension.

$$v = Wf, \quad \bar{v} = \bar{W}\bar{f} \tag{10}$$

$$v \in, R^{DX289}\bar{v} \in R^{D} \tag{11}$$

The matching score is driven by attention for the text and image features and calculate them as a pair The similarity matrix is calculated first for every pair in the sub-region of the image using

$$s = e^{T}v \tag{12}$$

In this equation, $s \in, R^{TX289}$ and i^{th} word in that sentence with the j^{th} sub-region available in that image. The normalized matrix for better result and stability,

$$\bar{s}_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_{k=0}^{T-1}\exp\left(s_{k,j}\right)} \tag{13}$$

Region-context vector c_i, is calculated. Earlier the interest was in generating the image, so it went through all the words and found the sub-region at each time. But, here it can be found if that particular word has any significance in the generation of that particular image. So for all the sub-region, it is needed to be checked one word at a time. This is summed here for all 289 sub-regions. $gamma_1$ is the attention scaling factor used in the equation and a score is generated for that and multiplied with the image features.

$$c_i = \sum_{j=0}^{288}\alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp\left(\gamma_1 \bar{s}i, j\right)}{\sum k = 0^{288}\exp\left(\gamma_1 \bar{s}_{i,k}\right)} \tag{14}$$

Word level relevance of i^{th} word is calculated with cosine similarity. It uses the current words, region-context vector and words vector. $R\left(c_i, e_i\right)$ tells us the score of each of those words on how important they are in generating the actual image.

$$R\left(c_i, e_i\right) = \left(c_i^{T}e_i\right) / \left(\|c_i\| \, \|e_i\|\right) \tag{15}$$

The word-level features are used to calculate the final global level scores. This image description score is calculated using word-level features with the hyperparameter γ_2, which signifies word to region context pair importance.

$$R(Q, D) = \log\left(\sum_{i=1}^{T-1}\exp\left(\gamma_2 R\left(c_i, e_i\right)\right)\right)^{\frac{1}{\gamma_2}} \tag{16}$$

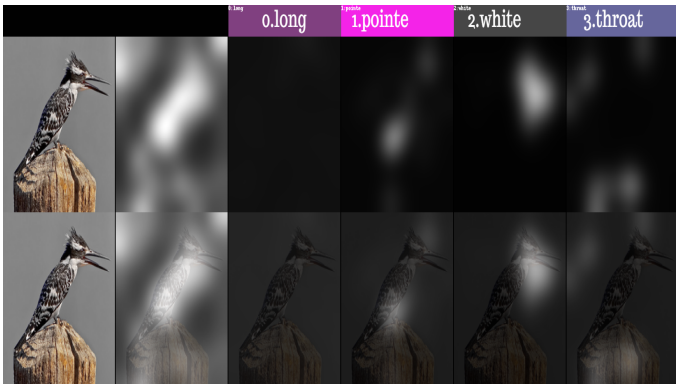


Fig. 3. Attention Map Generated by DAMSM on using RoBERTa Text Encoder. The Figure shows one Part of the Entire Caption Captured.

Similarly, using sentence-level features, using cosine similarity between global sentence and image features.

$$R(Q, D) = \left(\bar{v}^{T}\bar{e}\right) / \left(\|\bar{v}\|\|\bar{e}\|\right) \tag{17}$$

In the training process, calculation is done for the DAMSM loss for all the pairs. Thus with multiple descriptions and multiple images. The posterior probability is calculated of D_i matching with Q_i. So this gives a probability of how likely is that a description will be selected out of all the descriptions available. γ_3 is a hyperparameter for smoothing and stability in training the DAMSM. Similarly, it is also found the posterior probability when there is description and the images needs to be found.

$$P\left(D_i \mid Q_i\right) = \frac{\exp\left(\gamma_3 R\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R\left(Q_i, D_j\right)\right)} \tag{18}$$

$$P\left(Q_i \mid D_i\right) = \frac{\exp\left(\gamma_i R\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R\left(Q_j, D_i\right)\right)} \tag{19}$$

D. Total Loss

Total DAMSM loss in the network is calculated by

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^{w} + \mathcal{L}_2^{w} + \mathcal{L}_1^{s} + \mathcal{L}_2^{s} \tag{20}$$

L_1^{w} provides the word-level loss with respect to the description given the image and is the negative summation of log value of $P\left(D_i \mid Q_i\right)$. L_2^{w} provides the word-level loss with respect to the image given the description and is the negative summation of log value of $P\left(Q_i \mid D_i\right)$. L_1^{s} provides the sentence-level loss with respect to the description given the image. L_2^{s} provides the sentence-level loss with respect to the image given the description. Both the sentence level loss is same as word level loss except it use $\bar{e}$ instead of e.

Total loss in the entire network

$$\mathcal{L} = \mathcal{L}_G + \lambda\mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=1}^{3}\mathcal{L}_{G_i} \tag{21}$$

Here, L_G is the generator loss summed with L_{DAMSM} multiplied by a hyperparameter λ for smooth training.

Fig. 4. Image of an Indigo Bunting Generated at Every 50 Epochs by the Model.

Fig. 5. Comparison of Image Generated by the Model and the Ground Truth Image.

Fig. 6. Bird Generated by the Model for the Caption "This Bird has Wings that are Black and has a Red Belly".

E. Frechet Inception Distance (FID)

To analyze the model generated images, the Frechet Inception Distance (FID) score [17] is used as the metric. FID score is the best metric that can be used for the evaluation of the system as it measures the distance between the feature vectors of both real and generated images. The baseline model's paper has used the inception score as a metric for evaluating the GANs. The problem associated with the inception score being taken as a metric is that it does not find how the generated images compare with the actual images. With FID, it evaluate the generated images based on the generation distribution with the actual image in that particular target domain. For the FID score, the lower, the better. A lower score indicates that the generated images are closer to authentic images and are higher quality images and features with real one's match.

V. Experiments and Result

Multiple experiments were performed with the proposed Generative network using a pre-trained RoBERTa language model. The comprehensive network was initially pre-trained for DAMSM up to 200 epochs. Tesla V100 GPU with 16GB VRAM and 24GB CPU RAM for training with worker set as 4 were used for experimentation. The batch size for training was kept at 48 with a learning rate at encoder at 0.00005, and gradient clipping of 0.25 was kept to make sure training was stable. Various smoothing parameters were set during training, which helps train the various losses in multiple steps followed at DAMSM. $\gamma 1 = 4$, $\gamma 2 = 5$, $\gamma 3 = 10$, respectively. The parameters choosen for experimentation are taken from AttnGAN model and used for direct comparison. (GF_DIM) is the number of conv filters in the first layer of the generator and (DF_DIM) is the number of conv filters in the first layer

of the discriminator.

Ten captions were taken per image for training with the number of dimensions for the latent representation of the text embedding as 768. In previous experiments that were conducted in AttnGAN used bidirectional LSTM, which only required 256 embeddings. the base size was set as 299 which captured the attention map generated while pretraining the DAMSM with the pre-trained RoBERTa text encoder. The text encoder model was adopted from the hugging face library, providing the tokenizer for RoBERTa. It was observed that word vectors like colours get clustered together in vector space. The training of the transformer started from the pre-trained 'roberta-base' model with a RoBerTa tokenizer. The training of CNN started from the ImageNet pre-trained Inception-v3 model. Fig. 3 shows an attention map captured from image of a bird. At each frame a part of bird is being captured based on the text assosiated with it. The language model finds the relationship linking the body part and the colour designated for it. Attention mechanism explicates how each word corresponds to synthesizing a selective part of the bird image. Once the pre-training is completed, the DAMSM model generates a text encoder and an image encoder. This is used in the training of the AttnGAN architecture. AttnGAN network was trained using RoBERTa text encoder for 600 epochs. Due to limited resource allocation, kept the number of convolutional filters in the first layer of the generator (GF_DIM) and the number of convolutional filters in the first layer of the discriminator (DF_DIM) as 32.

The batch size was restricted to 8. The discriminator and generator learning rate was set to 0.0002. The generator and discriminator models were saved at every 50 epochs for analysis. For performance comparison with the baseline AttnGAN model, which trained the model with λ set as 5. The dimension of the RoBERTa text encoder is 768, amidst ten captions per image. The number of dimensions of the Noise vector was kept as 100 throughout the training process. Fig. 4 shows an Image generated at every 50 epoch by the AttnGAN with RoBERTa language model network. It is observable that around 200 epochs, the generator learns to generate an image close to a real-life bird. Images generated after 400 epochs looks realistic. By 600 epochs, it concluded the training and the models were saved. Fig. 5 explicates how the image generated by the model resembles the ground truth image. The model has learnt well to capture the essential details of birds like the body parts like wings, beaks, eyes, and feathers and understand its colour. It has also captured the pose of a bird to a reasonable extent. With more GPU power, it can use more generative networks to convert the image to higher quality. Fig. 6 was generated by the model around 600 epochs. The text as "this bird has black wings and a red belly", were provided to the generator synthesized an image matching the text description. 'roberta-base' was used as the model for capturing the context from the text description. The natural language model's main idea is to find attention heads and associate words in a bidirectional way. Fig. 7 envision the attention head for essential words in the sentence and how the RoBERTa model builds the attention mechanism. With the hugging face xbert tool, visualizing how the roberta-base model works in associating each word within the sentence with each other is simpler. The 'roberta-base' model uses 12 attention heads for generating a semantic relationship between each word in any direction. The word "this" is associated with itself and many other words within the sentence. The RoBERTa model learns that bird is the best associated and predictable word as the model is trained. There can be seen a strong connection between "this" and "bird". The word "bird" with the other words in the sentence is associated with particular words like "wings", "black", and "belly".These are the semantic relationship found by RoBERTa, and these get correlated with each other.

This assists in image generation and particularly in developing the text encoder in DAMSM loss. The word "black" here is a colour, and "wings" and "belly" are the two strong words that "belly" correlates. Colours like "red" get strongly related to "belly" in the sentence. Fig. 6 shows us the generated image by the model. The belly is red, and have black wings. This shows us how the RoBERTa model associates each word and what level of attention is provided for each word which nurses in the generation of the image with that particular features like body parts or colour.

TABLE II. FID SCORE GENERATED FOR EVERY 50 EPOCHS

Epoch	FID Score
0	275.833490
50	45.298608
100	32.540591
150	29.566392
200	25.616694
250	27.216560
300	24.648624
350	26.616447
400	22.377503
450	23.709113
500	22.760225
550	**20.773709**
600	21.468151

The Frechet Inception Distance(FID) score was calculated as part of the validation to compare the results with previous models as shown in Table II. Multiple epochs were ran and the FID score for λ = 5 were recorded. The number of generated images used for FID calculation was 2928, and the number of real images to be used in FID calculation was 11788. For epoch 0, had initially got a score of 275.833490. As each epoch is being trained, it can be seen the FID score keep reducing (Fig. 9).

A lower FID score indicates the model can generate more realistic images and various distributions of images. Around 550 epoch, a score of 20.773709 was obtained, the lowest and best for this model. This shows a good improvement from the baseline AttnGAN module [3] that used bidirectional LSTM as it got an FID score of 23.98 around the same number of epochs. Experiments were conducted on different values for λ at 100 epochs. Table III shows how the model performed without DAMSM and, by altering values of DAMSM, how the FID score or generation of different results is affected. A score of 35.44 was obtained. While tuning in too much attention value, the stability is lost in training, and it ends up with a score of 54.77 for λ set to 100. λ value of 1 seems to be stable for training with the AttnGAN and generating the bird images. A λ value set to 1 seems to work well for the model with RoBERTa embeddings and AttnGAN network. The λ value may fit differently for different datasets and should be

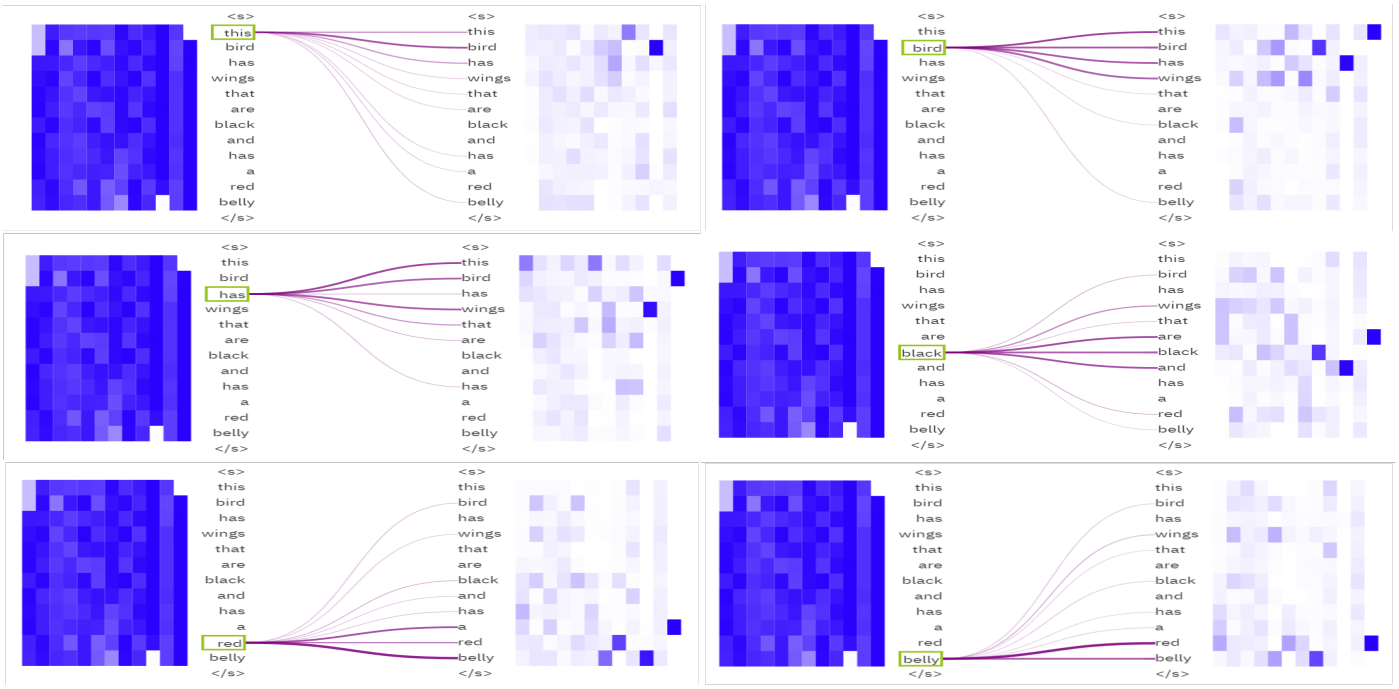


Fig. 7. Attention Head Generated by "Roberta-base" Pre-trained Model on the Text "This Bird has Black Wings and a Red Belly".

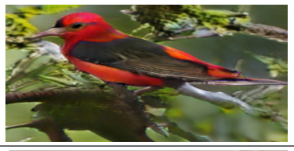
this bird has a red body and black wings with a small beak	
a small bird with a yellow belly and a small bill that curves down	
this bird is brown with white belly and long pointy beak	
bird with an all white body and a long black beak	
this is a yellow bird with a white head and a pointy beak	

Fig. 8. Images of Birds Generated by the Model for Various Provided Captions.

Fig. 9. Frechet Inception Distance(FID) Score for λ = 5.

explicitly experimented with that dataset. Table IV shows we have got a FID score of 20.77 with λ value set as 5 comparing to various other models. Fig. 8 shows some examples of text description and the images generated by the model.

TABLE III. FID SCORE GENERATED FOR 100 EPOCHS FOR DIFFERENT VALUES OF λ

Epoch	FID Score
0	35.440683
0.1	30.596095
1	28.663923
5	32.540591
10	34.538827
50	46.275016
100	54.775857

TABLE IV. COMPARISON OF FID SCORE WITH VARIOUS MODELS WHEN $\lambda = 5$

Model	GAWWN [12]	StackGANs [13]	AttnGAN [3]	RoBERTa GAN
FID Score	67.22	51.89	23.98	**20.77**

VI. CONCLUSION

This paper used the baseline AttnGAN model with the latest pre-trained language model RoBERTa. It used transformers with the generative network to analyze the fine-grained text to image generation. The generative network takes in captions in word and sentence embeddings level and uses the latent space of noise vector to synthesize birds images matching the text description. With the help of a deep attentional multimodal similarity model, and found the fine-grained image-text matching loss. This loss was further used to train the generator. Pre-trained Inception V3 model was used for the Image encoder along with pre-trained RoBERTa for Text Encoder. The baseline AttnGAN model had achieved a Frechet Inception Distance (FID) score of 23.98. The model with RoBERTa text encoder improved this performance and obtained a score of 20.77 on the CUB dataset. Various experiments were performed and recorded the results for the proposed architecture of generative networks.

REFERENCES

[1] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June). Generative adversarial text to image synthesis. In International Conference on Machine Learning (pp. 1060-1069). PMLR.

[2] Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. "Caltech-UCSD Birds 200". California Institute of Technology. CNS-TR-2010-001. 2010.

[3] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[6] Sikha, O. K., Kumar, S. S., & Soman, K. P. (2018). Salient region detection and object segmentation in color images using dynamic mode decomposition. Journal of Computational Science, 25, 351-366.

[7] Subbiah, U., Kumar, D. K., Thangavel, S. K., & Parameswaran, L. (2020, September). An Extensive Study and Comparison of the Various Approaches to Object Detection using Deep Learning. In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 183-194). IEEE.

[8] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks", in 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2017

[9] Aarthi, R., & Harini, S. (2018). "A Survey of Deep Convolutional Neural Network Applications in Image Processing". International Journal of Pure and Applied Mathematics, Vol. 118 No. 7, pp. 185-190.

[10] Brunda, R. & Divyashree, B. & Rani, N Shobha. (2018). Image segmentation technique- A comparative study. International Journal of Engineering and Technology(UAE). 7. 3131-3134. 10.14419/ijet.v7i4.18445.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[12] Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. Advances in neural information processing systems, 29, 217-225.

[13] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

[14] Li, B., Qi, X., Lukasiewicz, T., & Torr, P. H. (2019). Controllable text-to-image generation. arXiv preprint arXiv:1909.07083.

[15] Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1505-1514).

[16] Tsue, T., Sen, S., & Li, J. (2020). Cycle Text-To-Image GAN with BERT. arXiv preprint arXiv:2003.12137.

[17] Chong, M. J., & Forsyth, D. (2020). Effectively unbiased fid and inception score and where to find them. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6070-6079).

Performance Evaluation of BDAG Aided Blockchain Technology in Clustered Mobile Ad-Hoc Network for Secure Data Transmission

B. Harikrishnan[1]*, T. Balasubaramanian[2]*
PG & Research Department of Computer Science & Applications
Sri Vidya Mandir Arts & Science College (Autonomous), Katteri – 636 902, Uthangarai, Tamil Nadu, India[1, 2]
Department of Computer Science, Periyar University, Salem – 636 011, Tamil Nadu, India[1]

***Abstract*—In mobile ad-hoc network (MANET) environment, routing of data packets is a challenging task due to rapid changes in mobility and network topology. In addition, the security aspect of routing is disturbed by attacks caused by malicious nodes. These attacks greatly affect the Quality of Service. To overcome the challenges faced in routing message packets, the Bayesian Directed Acyclic Graph (B DAG) Aided Blockchain model is proposed for clustered MANET environment. The proposed model encompasses the following processes: (i) Multi factor authentication of users by using BLISS algorithm. This step involves acquisition of user credentials and generates hash values for those credentials using Cube Hash algorithm. These hash values are further used to generate public and private keys by BLISS algorithm, (ii) Weighted sum computation for clustering to reduce complexity in the MANET environment. Cluster head (CH) and cluster member (CM) are classified based on energy status, geometric distance, link quality and direction, (iii) A secure AODV based routing protocol using Dolphin Swarm Optimization (DSO) algorithm. This step involves selection of reputed node based on link stability, relative velocity, available bandwidth, energy, queue length, and trust. The packet forwarding is based on the reputation value of the node, by which the trust provided by malicious nodes are eliminated to improve security and (iv) Bayesian DAG aided blockchain, in which the user authenticity, data integrity of packets and signature are verified to mitigate the routing attacks created by nodes in the MANET environment. The proposed model is experimented in NS 3.26 network simulator tool and its performance in terms of multiple QoS metrics is evaluated.**

***Keywords*—Mobile ad-hoc network; node authentication; BLISS algorithm; blockchain; Bayesian direct acyclic graph**

I. INTRODUCTION

The current mobile nodes highly necessitate the secure communication. The key goal of this secure communication is to facilitate Quality of Service (QoS) and reliable data transmission in mobile ad-hoc network (MANET) [1,2]. The secure communication in MANET can be established by the combination of trusted mobile nodes and trusted third party. In secure environment, trusted third party plays a crucial role in several disaster applications. The security far-sightedness requirement in MANET is essential because of the succeeding issues, such as vulnerable entities in MANET to the routing attacks (black hole, gray hole, timing and intruder attacks), false warnings in the network and false registered credentials with the transmitted message [4-6] [9]. These issues exemplify the importance of the security in the MANET security. Further, the QoS parameters are affected due to lack of security, stability, and scalability [3]. When the network meets these three factors, each mobile node can be obtained with high QoS by means of high packet delivery ratio, throughput and low packet loss, routing overhead, route acquisition delay, end-to-end delay, and energy consumption.

Authentication is a main process in MANET security that verifies the provided credentials during registration process [6]. Few works [1,2,7] have performed authentication with an aid of blockchain technology to overcome failure during centralized administration when group of attackers combined to target it. Here, the transactions/blocks signature process is performed to provide security in MANET. The public key infrastructure (PKI) based authentication process has performed in blockchain where the Elliptic Curve Digital Signature Algorithm is used to generate key for the data transmission [7,8,15]. In blockchain, elliptic cryptography based authentication is contributed in the MANET. It performs authentication by considering the identity based signature procedures. Routing plays a vital role in the mobile communication and hence, providing security is significant while selecting best next hop. To ensure this, trust based node selection procedures are emerged in the MANET network [10]. In general, source node considers the two trusts viz. direct and indirect trusts for next hop selection [11]. Two different algorithms have utilized for estimation of direct and indirect trusts [12]. In direct trust, recommendation trust is estimated by considering the dropped and forwarded counts of the nodes. The trust is estimated based on two factors that are reputation information collection and trust value estimation procedures. A secure and efficient routing protocol (AOMDV) is designed to provide secure routing in the mobile ad-hoc network [13,14]. The MANET is largely utilized in several sensitive and non-sensitive applications. However, proving security in the MANET is difficult because of the issues, such as dynamic topology, communication latency, network scalability, and high processing security algorithms [16]. These issues induce difficulties in proving security to the MANET. Hence, the main aim and scope of this work is to provide high level security to the MANET network with better data transmission.

*Corresponding Author.

Paper outline: The remaining part of the paper is structured as follows: Section II describes the related works in the field of MANET, Section III presents the major problem statements, Section IV briefly explains the proposed system design and architecture in a well-organized manner and section V presents the experimental settings for the proposed system design and also evaluates the comparison between the proposed as well as previous approaches. Finally, the conclusion is work prospects are presented and summarized in Section VI.

II. Related Work

In this section, the existing works of secure cluster based routing in MANET are discussed. In [17], topological change adaptive ad-hoc on demand multipath distance vector routing protocol (TA-AOMDV) is proposed. This protocol is highly adaptable for high speed node movement in support of QoS. In this protocol, shortest path selection is implemented which takes four parameters as inputs for routing i.e. residual energy, available bandwidth, queue length, and link stability. In general, the path selection is not always optimum due to nodes movement and thus stable path is selected for routing. It considers both path stability and node density, but does not adapt for high speed scenario under large scale environment. In [18], routing attacks (black hole and gray hole) are detected by proposing an intrusion detection system on mobile ad-hoc environment. In attack scenario, the malicious node does not cooperate with other mobile nodes and intentionally disturb data communication by sending false route request & response forwarding and also false data transmission. These behaviors are timely detected in this work through deployment of IDS in most legitimate node. This legitimate node is determined by the connected dominating set model, which considers node energy and blacklist status for deploying IDS entity into it. The proposed method connected dominating set can be applied for small scale region, which is not efficient in large scale region. Further, the IDS deployment is a crucial process that must be running in decentralized mode, since malicious nodes can change the blacklist and nodes energy status. So, the routing attackers cannot be eliminated.

In [19], authors presented graph structure for mobile nodes communication over the network environment. To consume less energy and ensure QoS while transmitting packets from the source to destination node, the graph kernel structure based clustering algorithm is proposed in this work. In graph kernel, all mobile nodes are connected and CH is elected by means of certain significant parameters for data transmission. In particular, there are two processes are executed including cluster head election phase and cluster head maintenance phase. Based on node stability, CH is selected and CH maintenance is initiated by solving K-hop problem and node's current situation (stability and connectivity to other nodes). Among the set of CHs, shortest path is selected. Graph theory consumes more energy where clustering alone can able to perform. Since more computations are required to select K-hop shortest path selection. In [20], a hybrid optimization is proposed for energy related parameters adjustment. As a result of topology changes in network, energy consumption and packet losses during packet transmission is more. To address this issue, chronological and earth worm optimization algorithms are combined as a hybrid algorithm. Clustering is built by graph model (Gabriel graph) where node's power, connectivity, mobility, link lifetime and distance are considered for CH elected. In the graph structured cluster model, data packets having high power and energy are transmitted through nodes. Due to lack of infrastructure, secure communication is the way to minimize errors and data loss for data transmission. This further optimizes power and energy parameters when topology changes exponentially. In [21], network topology is controlled by hybrid artificial swarm intelligence algorithms i.e., robotic Darwin particle swarm optimization with graph based algorithm (RDPSO-GBA). Obviously, transmission delay happens when number of hop counts increases. In graph constructed framework, node's mobility and link connectivity between nodes are predicted. With these criterions, optimal route is selected and routing problem is solved via optimization solution. For route selection, ant colony optimization (ACO) is proposed. However, packet drop rate and energy consumption rate are high. In [22], expected transmission count (ETX) metric is used as one of the metrics to evaluate routing overhead. Through this metric, light ETX, light reverse ETX, and power light reverse ETX that minimizes routing overhead and also improves other ad-hoc network performances. All three metrics are computed for AODV routing protocol and implemented using NS3 network simulator. In experiments, Throughput, Packet Delivery Ratio, Useful Traffic Ratio, Jitter and End-To-End Delay are evaluated. Eventually, ETX does not support for optimum routing because AODV protocol works well in other networks i.e., static network, but does not suit for dynamic networks like MANET.

In [23], author discusses the trusted environment for IoT assisted MANET. The presented trust scheme is a combination of two individual metric i.e., direct and indirect trusts. The sum of direct and indirect trust values is used for final trust value. To compute trust values, beta probabilistic distribution is applied for combining different trust evidences and compute direct trust is calculated. A recommendation trust is computed using ARMA/GARCH, which is predicted for ensuring reliable and secure end-to-end forwarding of packets. Trustworthy nodes are selected in the routing. In [24], author proposes QASEC security routing protocol for secure data communication in MANET. This is a simple and lightweight model for best link selection from the set of available links for packets transmission between nodes. This link must be of optimal transmission link and thus produces minimized end to end delay. To ensure node authenticity, a simple authentication scheme is proposed, which relies on symmetric encryption for each mobile device shared secret keys generation. In particular, device identity, unique session key, and authentication token are considered for legitimate nodes identification. The symmetric encryption requires high energy consumption and also large key size that led to high processing time.

In [25], authors proposed a model for smart packet forwarding based on game theory. This model was proposed to identify and eliminate self-centered nodes which tend to drop the packets thereby increase the retransmission rate and reduce the performance. This model identifies the malicious

nodes by considering two factors namely possibility of packet drop and reward factor. Once the node is identified as malicious, the model stimulates the nodes to cooperate in packet forwarding paradigm. The formulated game theory model controls the energy consumption during modulation and data transmission. Even though the proposed method reduces retransmission rate, the network burden and complexity of the method is increased when the number of nodes are increased. In [26], authors proposed a method to protect the packet drop by two major attacks namely black hole attack and grey hole attack. The proposed method uses Artificial Bee Colony optimization algorithm (ABC) based on intelligent swarm algorithm in Artificial Neural Network (ANN) as a deep learning algorithm. The ABC optimization segregates the normal node and attacker node based on the fitness value. The ANN is trained with the output of optimization algorithm and is used to identify the malicious nodes thereby reducing the energy consumption and increasing the security. The ANN used in this method has no determined proper structure; the appropriate structure of the network is achieved through experience and trial and error. The duration of ANN is unknown and this will affect the performance of the method. In [27], authors proposed a model to evaluate the credibility of the mobile nodes by using trust reasoning model based on cloud model and Fuzzy Petri Net (FPN). The routes with minimal trust are selected and added to the routing table. Finally, a routing algorithm based on trust entropy routing algorithm and optimized link state routing protocol is presented. Based on the output of the proposed algorithm, the trust value gets assigned to the nodes and the route selection is carried out on the trustworthy nodes. Fuzzy logic is not always accurate hence results are obtained based on assumptions and the number of rules in the fuzzy logic makes it time consuming. In [28], authors proposed a routing mechanism named Energy Efficient Cloud-Assisted Routing (EECRM), which consists of three phases that cloud assist routing mechanism for efficient employment of route discovery, energy consumption phase for efficient utilization of energy and cloud service update phase. Suppose if any packet drop occurs in data transmission the adjacent route for better transmission is selected and the routing is performed to reduce the energy consumption. The parameters considered for selecting the backup nodes are not sufficient for optimal data transmission. The proposed method does not address problems caused by network attacks but practically these attacks increase latency and reduces performance during data transmission.

III. Problem Statement

This section summarizes the specific problems on blockchain aids and secures routing protocol for MANET. A framework for secure data transmission was proposed in [29]. The blockchain for security data collection in mobile ad-hoc network (B4SDC) framework exterminates of two types of attacks, such as collusion and spoofing. With the use of cooperative receipt report, the collusion attack is detected via control information forwarding for selected routes nodes only and mitigated whereas, the secure digital signature is used for spoofing attacks detection via signing sent messages from source to the destination. Further, excessive message forwarding attack is mitigated through the spoofing attacker's detection. The problem existing in this work, such as blockchain structure is not scalable. As the mobile nodes are always moving dynamically throughout the network, the use of conventional blockchain chain structure is not suitable.

Transaction confirmation time is higher since ECDSA, and $(SHA\text{-}2)^2$) consumes more processing time. Furthermore, creating blocks for dense MANET consumes high energy compared to clustered environment. Authentication is implemented and security credentials are transmitted through public channel and secret keys are received by the same that may facilitate the receiving of key by the compromised and malicious nodes. The mining complexity is higher for verification of block transactions during message transmission. Hence, a secure routing protocol was proposed to defend against the popular network attack named black hole attack [30]. This work is called as Blackhole Protected AODV routing protocol (BP-AODV) for malicious nodes detection. In routing process, blackhole attackers are detected through nodes past behaviors. Here, the history of node is collected and stored in routing table. To ensure security in the routing process, Chaotic Map features (Ergodicity, Randomness, and Sensitivity) are added which control the conditions and control parameters. However, the above work contains the following significant problems: Chaotic Map is not efficient enough because it needs successive monitoring neighbors' nodes. Besides, it also consumes more bandwidth during data transmission. Further, challenge at source node and secret responses at destination node are not be optimal when nodes move at high mobility. The routing is performed through BP-AODV protocol, which is not efficient to find optimum route since trustworthiness of nodes are computed by historical behaviors. This requires more packet retransmission due to poor computations of trust. Thus, it causes high packet losses and requires optimized solution. To evaluate the trust of the node, a support vector regression based corrective linear program classification model was proposed [31]. The main aim of this work is to reduce minimum end to end delay and maximize packet delivery ratio. In the first step, Tanimoto kernelled support vector regression is proposed to predict node features, such as node history and current energy status. Based on the node examined information, nodes are classified into two classes as "Trusted & Non-Trusted Nodes".

Subsequently, linear program boost classifier is used for trusted node selection for secure data transmission. Here, routing is performed based on node history, current energy status, and cooperativeness. In MANET environment, mobility of the node is more important, since lack of mobility information leads to high data loss. Further, trust values can be easily modified by malicious users. Tanimoto kernel in SVM provides less accuracy in trusted node prediction so that packet losses by malicious nodes involvement and also it leads to poor accuracy when mobile nodes mobility at high rate. To improve the energy efficiency and reduce complexity clustering based protocol was proposed [32].

The fuzzy logic based clustering is presented for cluster formation and cluster head selection. The cluster heads are selected by energy status, node degree, distance, trust level, and node mobility. A standby CH is also presented in case of

CH dies, moving out of coverage, or CH comprises. After the CH selection, trusted nodes are determined for secure route determination. Nevertheless, 243 fuzzy rules are generated for CH election, which increases energy consumption in cluster formation and also CH election. The strong computation mechanisms are required to estimate the trust level of mobile node. Here, the trust value is estimated based on the public historical behaviors. Thus, it induces malicious node participation in data transmission and also introduces high packet loss. Message may be modified or false data packets can be injected by the malicious nodes in intermediate hops.

The problem statements of these researches insist the need of a secure data transmission protocol which routes data in the optimal way through the trustworthy nodes whose trustworthiness is evaluated with strong parameters. The problems stated are illustrated in Table I. In this way, our proposed system is carried out towards the solutions of these problems by improving the security and scalability of data transmission in MANET.

TABLE I. EXISTING PROBLEMS

Method/ Technique	Concept	Drawback
B4SDC [29]	Detects collusion and spoofing attacks	The existing blockchain structure is not scalable furthermore; creating blocks for dense MANET consumes high energy
BP-AODV [30]	Blackhole attackers are detected through nodes past behaviors	The routing is not efficient since trustworthiness of the nodes is computed by considering only the historical behavior
Tanimoto SVM based linear program [31]	Classification of nodes based on node history and current energy status	Lack of mobility information leads to high data loss
Fuzzy logic based Clustering [32]	The trust nodes are determined by the cluster head	The number of fuzzy rules generated for CH election will make the process time consuming

IV. PROPOSED METHODOLOGY

The proposed system addresses challenging issues present in current MANET security works. The mobile ad-hoc network has become more popular in wide variety of applications that eventually increases intruders. This section is categorized into sub-sections depending on hierarchy of process involved to provide highly confidential authentication and data security in MANET that avoids severe attacks in network. In the present study, we have employed BDAG method to detect the malicious attacks mainly because of its relatively minimum resource utilization, fast detection of attacks and high delivery ratio.

A. B-DAG CHAIN Architecture

The proposed system comprises of four processes. The processing handled in each process is user validation, clustering, optimal route selection and data packet verification. The entities present in the proposed MANET environment are Security Node, Bayesian Network Directed Acyclic Graph aided Blockchain Nodes, and Mobile Nodes. The malicious traffic is provoked with the objective of demolishing the performance of the network. Initially the network consists of both the legitimate users and malicious users. Let the users be $\{U_1, U_2, U_3 \ldots U_n\}$ in which the data transmission is to be done between the legitimate users so the authentication of the legitimate users is important in secure data transmission. The users who possess original security credentials are termed to be legitimate users. The security nodes present in the environment is denoted as $\{S_1, S_2 \ldots S_n\}$. The security credentials of the legitimate users are managed by the security nodes. Once the security node receives the credentials of the users it starts to compute the key for message signing process. The number of users in the MANET environment is huge and it is not possible to manage the users individually hence clustering of mobile nodes is introduced, this process significantly reduces the complexity of managing the mobile nodes. Fig. 1 depicts the system architecture of the proposed work. The cluster head is elected and it is managed by security node. In order to reduce the packet drops and retransmission of data packet the data packet transmission is to be done through trusted nodes. The reputation value of each node is estimated and the source node selects the forwarder based on the reputation. In the destination side, both the sender's authentication and integrity of data packets are verified and the timestamp is also checked for secure transmission of data. On designing the BDAG chain architecture, the malicious nodes are detected and several attacks caused by these nodes are defended. The user authentication is carried out by using CubeHash algorithm and bliss algorithm. Then clustering is done using weighted sum computing for clustering algorithm and finally the secure routing is carried out by using Dolphin Swarm Optimization. Fig. 1 shows the proposed BDAG chain architecture and the processing carried out in each process. All these processes are completely focused on mitigation of the attacks caused by the malicious nodes.

B. Multi-Factor Authentication

The U_n users of MANET environment are validated when the user requests with a service. User validation includes processing of two phases as (I) Registration phase and (II) Authentication phase. The steps in registration phase are given as follows.

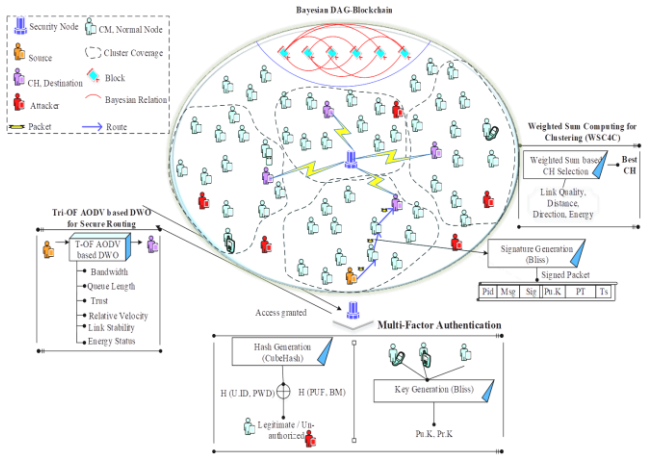


Fig. 1. System Architecture.

STEP 1: First assume U_1 be the user who requests the security node (SN) for registration with his / her unique Binary format of user finger vein, Physically Unclonable Function (PUF), Password and User ID. These credentials are submitted to SN for registration.

U_1 (BM||PUF||PWD||ID) ⟶ SN (1)

STEP 2: Then, receiving user credentials from U_1, the SN generates hash value for all user credentials by using CubeHash algorithm. The parameters involved in CubeHash algorithm are listed below:

- RD - The number of iterations in {1to 16}.
- Bpb – It is represented in bits per packet block {1 to 128}.
- O – The output bits range from {2 to 512}.
- Pkt - the actual message as a string of bits in size {0 to 2^{128-1}}.

The algorithm involves the stages listed below:

- The state S is initialized based on (O, Bpb, RD)
- The Pkt should be divisible by Bpb for that padding may be performed.
- The RD rounds of the ordering on state S is performed for every Bpb-byte block in the padded O xored with first Bpb- bytes of state.
- The state S is then finalized.
- The final result is delivered as the first O bits of state S.
- Each state is built of 32 bits. The first three values are set to O/8, Bpb, RD in the initialization stage. Further the state is ordered to 10RD rounds. Until the value of Pkt is divisible by Bpb the padding is carried out by appending the pkt by 1. The preserved symmetry is broken by performing the XOR operation to the last bit with the integer value1. Finally, the state is reordered through 10RD rounds once again. CubeHash seems to easily configure the input parameters such as BM, PUF, PWD and ID. The maximum security is obtained by deploying CubeHash 8/1-512 and the energy efficiency is obtained by deploying CubeHash 16/32. The hash values H is determined which is used to generate the unique signature.

Procedure for Key Pair Generation

1: Begin
2: Initialize nodes
3: $\boldsymbol{U_1}\rightarrow$ request key pairs to the $\boldsymbol{SN}$
4: SN (For U_1 generate key pairs)
hash value acquisition
Key pair generation(Pr.k, Pu.k)
5: End for
6: $\boldsymbol{SN}\rightarrow$ reply key pairs to U_1

Procedure Description: The U_1 requests the key pairs to the security node which will be generated from the hash values determined for the user credentials. Once hash values are obtained the two different keys Pr.k and Pu.k are generated. The security node assigns the Pr.k and Pu.k to the user here it is U_1. Further the keys from SN are stored in user's device.

The BLISS algorithm is used to generate the digital signature. First, the two different keys Pr.k and Pu.k are generated. The Pr.k is a (short) matrix $S \in Z_{2q}^{mxn}$ and Pu.k is given by the matrix $A \in Z_{2q}^{n\times m}$ such that AS = q**I**n (mod 2q). The security node assigns the Pr.k and Pu.k to the user here it is U_1.

SN ⟶ U_1 (Pr.k, Pu.k) (2)

STEP 3: Further the keys from SN is stored in user's device which is required during the authentication process. The registration phase for the device is completed.

Procedure for Message packet transmission

1: Initialization
2: $U_1 \rightarrow$ requests for connection with G_1
3: Timestamp verification by G_1
4: $G_1\rightarrow$ reply ack to U_1
5: U_1 (for message transmission)
Message $\boldsymbol{\mu}$ is generated
Signing of message with generated keys (Pr.k, Pu.k)
6: End for
7: $U_1\rightarrow$ Message transmission to G_1

Procedure description: U_1 requests for the connection establishment with G_1. The timestamp for the request message is verifies by G1 and the ack for connection is sent to U_1. For message transmission the message μ is generated and the digital signature is signed with the key. The signed message is then transmitted to G_1 in several hops. The steps followed in authentication phase are described below:

STEP 1: During authentication the request from U_1 is submitted to G_1 for data transmission. On receiving request, the timestamp T_1 freshness is verified. Then the G_1 replies the U_1 with the acknowledgment (ack). Once the ack is received by the U_1, the message μ is generated and the private key is used to sign the message. The signed message is denoted as δ which is sent to the destination G_1

$$U_1(\mu \oplus Pr.k) \longrightarrow U_1\ (\delta\ ) \quad (3)$$

Procedure for Message packet reception

1: Reception
2: G_1 (For key generation)
verification of sender identity (U_1)
verification of data integrity (δ)
Pr.k is generated for the destination node
3: End for
4: The message packet is decrypted

Procedure Description: After the reception of encrypted message packet the G_1 verifies the sender identity of the message and the data integrity of the packet. After the verification process, the private key is generated to decrypt the received message.

- STEP 2: The signed message packet is transmitted to the next reputed node and thus the message packet reaches the destination in several hops.

$U\_1\ (\delta) \longrightarrow G\_1$ (4)

STEP 3: Once the destination node receives the packet, the private key for the respective message packet is generated to the destination. This way of data transmission provides security to the transmitted data effectively.

$$G_1(\delta) \longrightarrow G_1(\mu \oplus Pr.k) \quad (5)$$

The authentication of users enables to withhold illegitimate user's access into the network. A system model without authentication process is much easier for attackers to initiate attack which degrades network performance. The authentication of user by private and public key ensures to allow access only for the legitimate users. The goal of our work is to defend the attacks and to perform the data transmission securely.

C. Weighted Sum Computation for Clustering

The mobile nodes in the MANET environment are clustered to minimize the computation complexity. The clustering of nodes undergoes two phases they are (I) cluster head election and (II) cluster formation. The nodes are grouped based on the factors such as Energy status, distance, link quality and direction. The resulting cluster will contain single cluster head (CH) and number of cluster members. The cluster head has complete information about the cluster members and link state details. Each and every node in the cluster is connected to the cluster head with a bi-directional link, through this each node in the cluster knows which cluster it belongs to. The proposed Weighted Sum Computation. For Clustering (WSC4C) performs well in optimal cluster formation in highly moving MANET environment. The proposed algorithm computes the weight score for each factor and then elects the optimal node as cluster head.

1) Energy status calculation: The energy status for each node in the MANET environment is obtained by finding the difference of the energy availability in two sub-phases. The energy spent on forming the topology is considered to obtain the correct energy status of the mobile node. For that the energy status before topology formation and the energy spent on topology formation. The energy status can be found by the equation as follows:

$$E_{status} = E_{Bt} - E_{St} \quad (6)$$

2) Geometrical distance calculation: The distance between each node is to be found to decide the cluster size and the optimal cluster head for the respective cluster. The distance calculation is carried out in meters (m). The optimal distance between the node and the security node is calculated. The reason to calculate distance is that when the distance between two nodes is small the energy required in transmitting the packet between the nodes reduces. The geometrical distance between two nodes is calculated as follows:

$$D(t) = \sqrt{A1 + A2 + \theta} \quad (7)$$

3) Link quality estimation: The link quality is one of the significant parameters to be considered in highly moving MANET environment. The estimation of link quality is carried out through number of transmissions expected (NET) and the path count expected (PCE). The number of transmissions expected (NET) between the transmission of packets between sender and receiver is used to calculate the link quality. Let node U(i) be the intermediate node which has the probability of receiving and transmitting the message packets of $\alpha(i) and\ \beta(i)$ respectively. The probability of packet delivery ratio between the nodes is calculated as follows:

$$\alpha(i) = \frac{n_w}{w/_\tau} \quad (8)$$

The probability that a previous node sends packet to the node U(i) is stated as $P_{pre} = 1 - \Pi_{i>s}(1 - \alpha_s\beta_i)$ and the probability that the next node receives a packet from U(i) is stated as $P_{next} = 1 - \Pi_{d>i}(1 - \alpha_i\beta_d)$. Then the number of transmissions expected is determined by

$$NET = \frac{1}{Ppre \times Pnext} = \frac{1}{(1-\Pi_{i>s}(1-\alpha_s\beta_i))(1-\Pi_{d>i}(1-\alpha_i\beta_d))} \quad (9)$$

The PCE value is calculated based on the probability of packet delivery ratio. This metric is calculated to find the optimal pair of the node U(i). The PCE value from sender to receiver is can be calculated as

$$PCE(s,d) = \frac{1+\sum_i PCE\ (c_i^{s,d},d)\ (1-\Pi_{i>s}(1-\alpha_s\beta_i))\Pi_{j-1}^{i-1}(1-\Pi_{d>i}((1-\alpha_s\beta_j))}{1-\Pi_{s>i}(1-\alpha_s\beta_i)} \quad (10)$$

The above determined NET and PCE metrics are used to estimate the link quality between the nodes in the MANET environment.

4) Direction estimation: The direction of nodes in a MANET environment can be estimated from the change in link quality between the nodes. If the link quality increases between the nodes, then they are moving towards one another and if the link quality degrades between the nodes, then they are moving away from one another. Thus, the direction of the nodes can be classified into two classes they are (i) moving toward (TW) and (ii) moving away (AW). The weighted sum computation for clustering is carried out with the above calculated factors and the clustering is performed in an optimal manner. From the nodes present in the cluster, the cluster head (CH) is elected based on the calculated weighted sum. Table II illustrates the formation of cluster and election of cluster head based on WSM score. The node $U3$ is elected as the Cluster Head (CH), as the Weighted Sum Computation (WSC) score for the respective node is greater when compared to other nodes.

TABLE II. CLUSTER FORMATION AND CLUSTER HEAD SELECTION

Nodes	Input parameters for clustering				WSC score	Role
	C_1	C_2	C_3	C_4		
$U1$	5	12	75%	*AW*	0.75	CM
$U2$	8	10	78%	*TW*	0.82	CM
U3	**11**	**3**	**90%**	***TW***	**0.91**	**CH**
$U4$	6	8	83%	*AW*	0.69	CM
$U5$	9	15	67%	*AW*	0.56	CM

Once the cluster is formed, each node in the cluster will broadcast its neighbor table for every periodic time period as a hello message. After receiving hello message from the neighbor node, the respective nodes ID and role (CH or CM) will be registered in the neighbor table. These clusters are managed by the security nodes. If any attack patterns found by the security node, it will immediately isolate the particular malicious node and inform this message throughout the network. This way of forming the cluster and electing the cluster head is efficient and this will greatly minimize the complexity and overhead in packets and control message forwarding.

D. TRI-OF AODV based DSO for Secure Routing

In the MANET environment, communication is done by sharing message packets from one node to another node. In this work, Ad hoc On-demand Distance Vector (AODV) protocol which is a significant routing protocol in MANET is used. The AODV protocol overcomes the issues of mobile network such as high mobility, packet loss, etc. The AODV routes the message packets to the next node based on the trust value of the respective node. The trust value is classified into two types they are (i) direct trust and (ii) indirect trust. The direct trust is estimated by using the past behavior and successful transaction of the node whereas the indirect trust is provided by the security node. In order to achieve more secure message packet routing, we use Dolphin Swarm Optimization (DSO) algorithm. The proposed algorithm computes the reputation value for each node and forwards the message packets to the most reputed node. The algorithm is classified into three stages which are as follows:

- Search stage.
- Call stage.
- Response stage.

Search stage: In the search stage, the mobile node searches its neighbor node from the neighbor table. The algorithm obtains two nodes one is, the node which is selected by the source node and the second is the node which is recommended by another neighboring node. For each node U(i) (i= 1, 2, 3…N), two corresponding possibilities X_i (i=1 to N) and Y_i (i= 1, 2, 3…N). Where, X_i is the node selected by the source node and Y_i is the node recommended by the neighbor nodes. First the node U(i) calculates the fitness for the node selected by itself and then for the node recommended by the neighbor nodes. The comparison of both the fitness value will help the source node to select the best node. The fitness value is calculated based on the factors such as link stability, relative velocity, available bandwidth, energy, queue length, and trust.

For the node X_i that node U(i) gets, its fitness F_{xit} is calculated as follows:

$$Fxit = Fitness\ (Xi) \tag{11}$$

Then the fitness value for node Y_i is calculated as F_{yit} is calculated as follows:

$$Fyit = Fitness\ (Yi) \tag{12}$$

$$\text{If } F_{yit} > F_{xit} \tag{13}$$

Then Yi is replaced by Xi. Otherwise, Xi which is selected by the source node itself does not change. After the updation of the reputed node, the DSO enters into next stage.

Call stage: In the call stage, the source node U(i) informs the neighbor node about the result obtained in the search stage and requests the connection from the selected reputed node for data transmission. The transmission time matrix (TM) gets updated as follows:

$$TM_{U(i)} = \left\lceil \frac{Dist}{speed} \right\rceil \tag{14}$$

After the updation of the transmission time matrix gets updated and the DSO enters the next stage.

Response stage: The reply for the request sent to the reputed node is received by the source node and the message packet is generated and sent to the respective node. The reputed node receives the message packet from the source node within the mentioned transmission time matrix. (TM). When the time becomes.

$$TMU(i) = 0 \tag{15}$$

The reputed node will no longer get the message packet, which means that the message packet will be transmitted to the reputed node within the time. This process gets looped until the destination node receives the packet. In this process the trust values provided by the malicious nodes gets eliminated and the message is transmitted through the reputed nodes and reaches the destination node in the optimal time. The objectives achieved by the Dolphin Swarm Optimization is listed below:

- Low Relative Velocity and High Link Stability.
- High Available Bandwidth and High Energy.
- High Trust and Low Queue Length.

The Procedure 4 presents the overall working of DSO based AODV for multi-hop routing. The above explained three stages along with the start and end phase is involved for fitness computation of each relay node. In the start phase the mobile nodes are initialized with its input parameters. This process is significant for the initialization of parameters for fitness evaluation.

The best fitness function is chosen for packet transmission. The end condition is achieved when the packet reaches the destination. The matrix updation is done in both call stage and response stage which is done to make the routing process within the labeled time period.

Procedure: Route selection by DSO algorithm.
1: initialization
Collects information about neighboring nodes from neighbor table
Nodes= {U1, U2, U3...} in the neighbor table.
2: begin loop
While the end condition is not satisfied **do**
2.1: search stage
Two nodes Xi and Yi are obtained.
Calculate fitness F_{xit} and F_{yit}.
Highest fitness node is selected.
2.2 Call stage
Request for connection
Update time matrix.
2.3 Response stage
Reply received
Packet generation
Update time matrix.
3 **If** packet reached destination then
End loop
Else
Increment loop
End if
End while
End

E. Attacks Mitigation by Bayesian DAG-Blockchain

In this section, the attacks associated with the routing in MANET environment is explained briefly and a model is proposed to defend against these routing attacks. Generally, routing in mobile ad-hoc network is to be performed in an optimal way to reduce the latency and increase the privacy of the mobile nodes. Some of the nodes try to bypass the message packet to steal the sensitive information embedded in it; these nodes are collectively called as malicious nodes. To mitigate the attacks occurring in routing process, the Bayesian Direct Acyclic Graph (B DAG) aided Blockchain model is proposed. The security of the routing protocol is increased by using blockchain technology. The blockchain makes use of asymmetric encryption and authorization processing. Fig. 2(a) illustrates the format of the data block; (b) describes the individual specification of the data blocks.

Version	Block Types	Previous Block Hash	Timestamp	Trust Value	Residual Energy	Bliss Signature	Index Record

(a)

version	Block version number
Block types	Node details
Previous Block Hash	Link for connecting block
Timestamp	Block creation time
Trust value	Node trust
Residual energy	Energy availability
Bliss signature	Public and private key
Index record	Holds index information

(b)

Fig. 2. (a) Format of Data Block, (b) Data Block Description.

The transaction of packets stored on the blockchain is public, but the identity of the nodes is encrypted and the key for encryption is provided to the data owner only thereby ensuring the privacy. Each and every action in blockchain is transparent and is available to every node of the block however the data cannot be deleted or changed. The decentralized management of network reduces a number of risks that emerge with data being managed centrally. This type of management has no central point of failure. The priority of trust is given equally to all the nodes. The verification of message packet is performed to find whether the source of the packet is a legitimate user and the data integrity of the packet. To do so each and every event of all the nodes are stored in a chain of blocks. To eliminate the problem of storage requirement, these events are converted into hash values and these hash values are table verification process. The previous hash of the block is stored in the current block therefore the security of the blockchain increases with the number of blocks. Traditional blockchain technologies arrange the blocks in a single chain. So, in order to verify the packet integrity, the hash values of source to destination are to be verified and this will lead to delay in the verification process. To overcome this delay the blocks are presented in Direct Acyclic Graph (DAG). This enables the verification process to check the needed blocks only. But in order to select the required block, all the blocks associated with the process is to be checked. Therefore, a Bayesian Network is used to estimate the probability of the hash value in the respective block. This significantly reduces the delay associated with the process. The Block Independence value (BI) is presented as a triplet of $J\ (x, y, z)$. Here the value x denotes the key of $J(x, y, z)$. The factors determining the Block independency is listed as follows:

- Symmetry
- Decomposition
- Weak union
- Contraction

The graph is plotted for the number of blocks in a hyper tree construction ordering. The hyper graph is denoted as.

$$\langle B, P\rangle = \{b_1, b_2 \ldots b_n\} \quad (16)$$

Where, $b_i \subseteq P$ is called a *hyper edge* of ⟨B, P⟩, and $P = b_1 \cup b_2 \cup \ldots \cup b_n$. the hyper graph B is determined as acyclic if $, si = bi \cap (b_1 b_2 \ldots b_{i-1}) \subseteq b_j, 1 \leq j \leq i-1,$ where s is referred to as segregator. Thus, an acyclic hypergraph is constructed and thereby removing the dependency between each block in the blockchain.

V. Results and Discussion

This section presents the simulation results with the detailed description of three sub-sections include simulation setup, comparison analysis, security & efficiency analysis for the proposed model in comparison with the existing approaches.

A. Simulation Setup

The Network Simulator version 3 (NS3) is used to evaluate our proposed model. The hardware and software requirements for the proposed model are illustrated in Table III.

The mobile node configuration is deployed to transmit the data by sensing the varied circumstances. The network simulator 3 tool has better features of network and provides all specifications of MANET. The proposed B DAG aided blockchain based secure routing model is experimented in 1000m × 800m simulation environment for evaluation of packet transmission by defending various attacks. The codes implemented for the construction of clusters, route establishment for transmission of data. Secure routing is a significant factor which is considered in many research works and our work also focuses on secure routing in MANET environment. Table IV describes the parameters associated in the experiment of proposed work in simulation tool. Fig.3 illustrates the flow of the proposed work in the NS3 tool.

B. Comparison Analysis

This sub-section presents the formulation of the proposed B-DAG aided blockchain secure routing model with respect to several QoS metrics. The proposed research work is compared with several existing works. Particularly, the following metrics are considered for performance analysis: attack prediction rate, end to end delay, packet delivery ratio, route acquisition delay, routing overhead, security strength and throughput. These metrics are described in the following in detail.

TABLE III. SYSTEM CONFIGURATION

Hardware specifications	Processor	NS3.26
	RAM	2GB
	Hard Disk	60GB
Software Specifications	Network Simulator	Pentium Dual Core and Above
	OS	Ubuntu 14.04 LTS

TABLE IV. SIMULATION PARAMETERS

Parameters	**Description**
Simulation area	1000 m*800 m
Number of nodes	100 with 10% of attackers
Node mobility model	Random waypoint model
Node speed (Max)	5 m/s
Forwarding capacity	2 Mbps
Number of flows	50
Transmission range	250 m
Packet transmission average rate (per flow)	1024 bytes/packet
Node buffer size	64 packets (fixed)
Queue type	Priority queue
Traffic type	TCP, UDP, and ICMP
Nodes distribution	Random
Interface type	Physical wireless
Neighbor nodes waiting time	0.3 s
Duration for packets carrying	1 s
Propagation delay mode	Constant speed
MAC type	Ad Hoc Wi-fi MAC

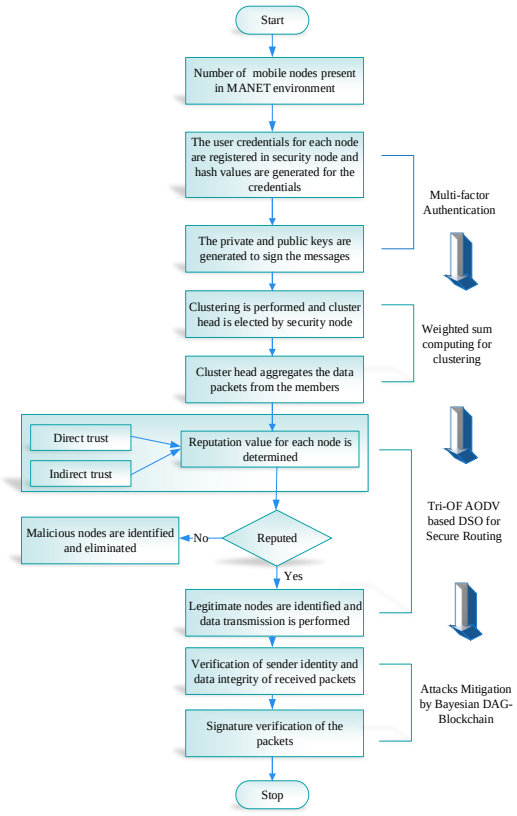


Fig. 3. Flow of B-DAGChain Simulation.

1) Impact of attack prediction rate: Attack prediction rate is defined as the number of attacks predicted for the given unit time. The true positive value is used to compute the attacks. It is defined as the number of packets which are determined correctly as normal for the total number of packets forwarded. It is calculated as.

$$APR = \frac{no.of.detected\ attacks}{no.of.attacks} \times 100\% \quad (17)$$

$$APR = TPV = \frac{TP}{TP+FN} \quad (18)$$

Fig. 4 depicts the attack prediction rate of three works including the proposed work with respect to number of malicious nodes. With the implementation of B DAG aided Blockchain model, attack prediction rate gets improved for the proposed research work which is evaluated for different number of malicious nodes. This work primly focuses on prediction attacks caused by the malicious node and mitigation of those attacks Table V. Firstly, the packets are signed with the digital signature and then packets are forwarded through reputed nodes and reach the destination. Existing works such as B4SDC and BP-AODV considers the trust values of the malicious nodes.

2) Impact of end-to-end delay: The End-to-End delay is defined as the delay of packets between two nodes. The end-to-end delay is calculated with respect to number of nodes.

Fig. 5 depicts the end-to-end delay the proposed method in comparison to the existing works for increasing number of nodes. The proposed work implements clustering of nodes

using Weighted Sum Computing for Clustering (WSC4C) algorithm which will reduce the complexity and delay associated with quantity of nodes Table VI. In the existing works [29], [30] the delay increases exponentially with increase in the number of nodes.

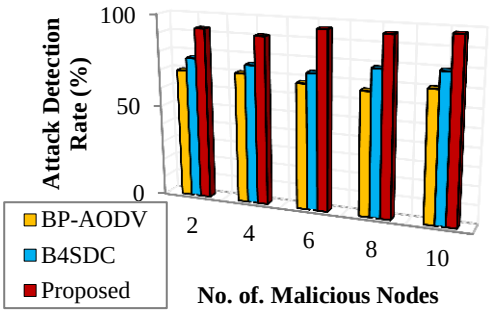


Fig. 4. Attack Prediction Rate vs. no. of. Malicious Nodes.

TABLE V. COMPARISON OF ATTACK DETECTION RATE OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF MALICIOUS NODES

No of Malicious Node	Attack Detection Rate (%)		
	BP-AODV	B4SDC	Proposed
2	69.7	76.7	92.9
4	70.5	75.3	90.8
6	67.8	73.5	96.3
8	66.5	78.3	95.7
10	70.3	79.4	97.7

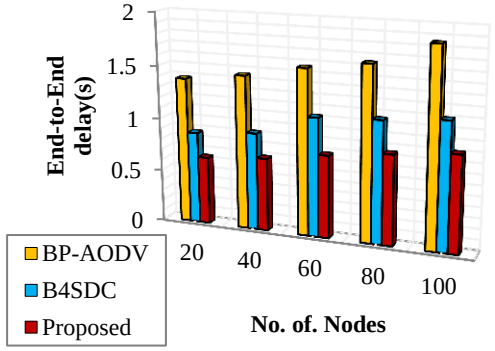


Fig. 5. End-to-End Delay vs. no. of. Nodes.

TABLE VI. COMPARISON OF END-TO-END DELAY OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF NODES

No. of Nodes	End-to-End Delay(s)		
	BP-AODV	B4SDC	Proposed
20	1.38	0.87	0.64
40	1.45	0.92	0.69
60	1.56	1.12	0.78
80	1.64	1.15	0.85
100	1.85	1.2	0.91

3) Impact of packet delivery ratio: Packet Delivery Ratio (PDR) is termed as the ratio between the numbers of packets delivered with respect to the number of malicious nodes present in the MANET environment. Fig 6 depicts the performance of proposed method in terms of packet delivery. With the increased number of malicious nodes, the existing systems [29], [30] fails to achieve better packet delivery ratio. The proposed work uses Dolphin Swarm Optimization (DSO) for reputed node selection and AODV protocol for secure route selection and this eliminates the trust provided by the malicious nodes thereby not affecting the packet delivery ratio Table VII.

4) Impact of security strength ratio: Security Strength is described as the level of security provided to the sensitive information embedded into the message packet. The security strength is improved by authentication of legitimate users and encryption of message packets during transmission of packets. Fig. 7 depicts the comparison of security strength provided by the proposed methods and other existing methods with respect to the packet size. The proposed method proves to be more secure even when the size of the packets is increased. The Bayesian Direct acyclic graph aided blockchain technology implemented in the proposed work improves the security of routing Table VIII. The existing works fails to focus on security when the packets and number of nodes increase which is a significant drawback.

5) Impact of throughput: Throughput is defined as the rate of successful reception of message packets by the destination node. It is one of the important factors in determining the accuracy and safety of the research work. Fig. 8 depicts the throughput achieved by the packets with respect to the increasing number of malicious nodes in three research works. This shows that the proposed work has higher throughput rate when compared with other research works. The proposed work uses DSO in selecting the reputed nodes thereby eliminating the malicious nodes which will reduce the number of packets drop and other adversary activities. The existing works are attracted by the smaller number of hop counts presented by the attackers and considers the trust value provided by them but the proposed work calculates the reputation value based on several features including the indirect trust provided by the security node Table IX. Therefore, it is able to differentiate the malicious nodes from the legitimate nodes. Thus, the proposed work has better throughput.

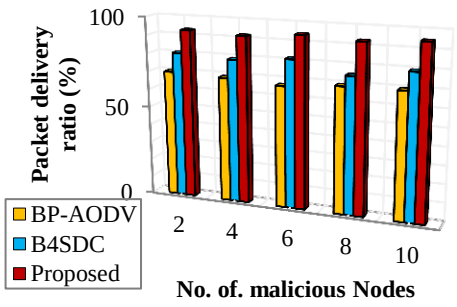


Fig. 6. Packet Delivery Ratio vs. no. of. Malicious Nodes.

TABLE VII. COMPARISON OF PACKET DELIVERY RATIO OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF MALICIOUS NODES

No. of Malicious Nodes	Packet Delivery Ratio(%)		
	BP-AODV	B4SDC	Proposed
2	69.8	80.5	93.2
4	68.7	78.9	91.9
6	66.8	81.3	94.3
8	69.2	74.9	92.8
10	69.5	79.4	94.7

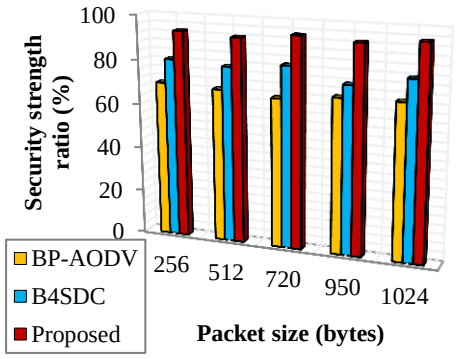


Fig. 7. Security Strength Ratio vs. Packet Size.

TABLE VIII. COMPARISON OF SECURITY STRENGTH RATIO OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF PACKET SIZE IN BYTES

Packet Size (Bytes)	Security Strength ratio (%)		
	BP-AODV	B4SDC	Proposed
256	69.8	80.5	93.2
512	68.7	78.9	91.9
720	66.8	81.3	94.3
950	69.2	74.9	92.8
1024	69.5	79.4	94.7

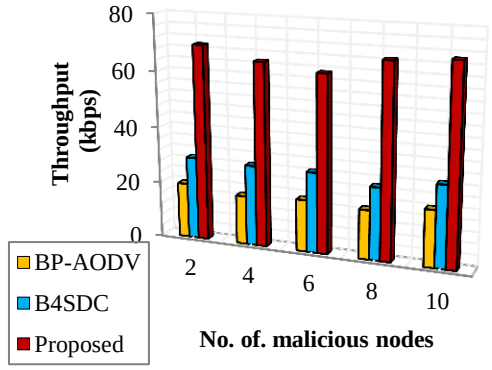


Fig. 8. Throughput vs. no. of. Malicious Nodes.

TABLE IX. COMPARISON OF THROUGHPUT OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF MALICIOUS NODES

No. of Malicious Nodes	Throughput(kbps)		
	BP-AODV	B4SDC	Proposed
2	19.6	29.4	69.7
4	17.4	28.7	65.3
6	18.5	28.6	62.9
8	17.5	25.8	68.7
10	20.2	29.2	70.2

6) Impact of routing overhead: Routing Overhead is defined as the ratio of number of packets generated to the number of packets transmitted in the established route. The routing overhead depends on the link stability and quality. The number of nodes and mobility of nodes also influences the routing overhead. Fig. 9 depicts the performance of proposed work in terms of routing overhead with respect to number of nodes. This shows that the proposed work has less routing overhead when compared to the other existing works, this is because the proposed work implements the clustering process by using Weighted Sum Computation for Clustering (WSC4C) to cluster the nodes in the MANET network thereby increasing the link quality between the nodes which will reduces the routing overhead Table X.

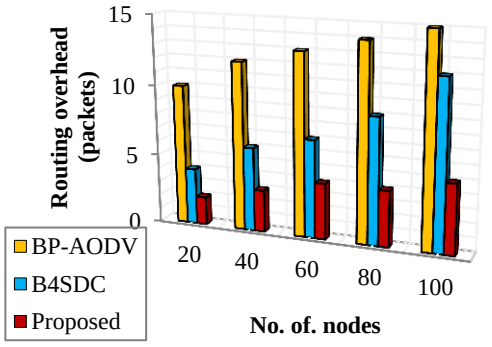


Fig. 9. Routing Overhead vs. no. of. Nodes.

TABLE X. COMPARISON OF ROUTING OVERHEAD OF PROPOSED METHOD WITH EXISTING METHODS FOR DIFFERENT NUMBER OF NODES

No. of Nodes	Routing Overhead (Packets)		
	BP-AODV	B4SDC	Proposed
20	10	4	2
40	12	6	3
60	13	7	4
80	14	9	4
100	15	12	5

7) Impact of routing acquisition delay: Route Acquisition is described as the establishment of route between two nodes for forwarding the data packets from source to destination. The delay associated with the route establishment is termed as route acquisition delay. This delay is occurred mainly due to lack of selection of neighboring nodes. Fig. 10 depicts the comparison of route acquisition delay in three research works including the proposed work with respect to the routing hops. Existing works focuses on only the selection of neighbor nodes based on the trust value but the selected node may not have sufficient energy level to forward the packet and this will lead to route acquisition delay Table XI. The proposed work selects the reputed neighbor node based on link stability, relative velocity, available bandwidth, energy, queue length, and trust. Hence the delay associated with route acquisition is decreased significantly.

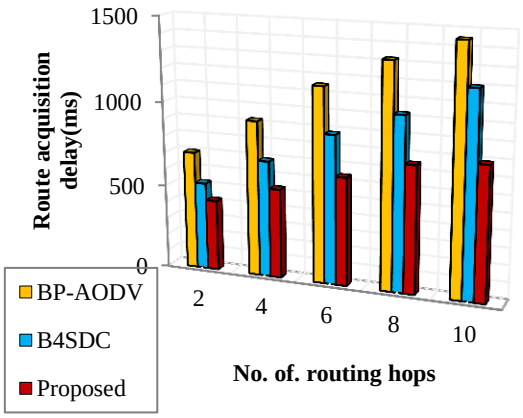


Fig. 10. Routing Acquisition Delay vs. no. of. Routing Hops.

TABLE XI. COMPARISON OF ROUTING ACQUISITION DELAY OF PROPOSED METHOD WITH EXISTING METHODS FOR NUMBER OF ROUTING HOPS

No. of Routing Hops	Routing Acquisition Delay(ms)		
	BP-AODV	B4SDC	Proposed
2	700	520	420
4	920	690	530
6	1150	880	640
8	1320	1024	750
10	1450	1200	790

C. Security and Efficiency Analysis

The packet transmission in the MANET environment carries sensitive data and the security of these message packets should be ensured. The malicious nodes in the network lead to routing issues. The attacks caused by these nodes will affect the performance of routing between the nodes and greatly affects the privacy of the nodes. Some of the important routing attacks and the process followed by the proposed work to overcome these attacks are explained below:

Black Hole attack: In this type of attack, the attacker will attract the source node with low hop routes. Once the sender begins to transmit data packets, the attacker will drop the packets without transmitting it to the next node. This increases the latency among the nodes. The proposed system selects the reputed node for pack forwarding by considering both the direct and indirect trust of each node thereby identifying the false trust given by the malicious nodes.

Grey Hole attack: This is an advanced level of black hole attack in which the malicious node will behave like a legitimate node. Not all the packets send through this node will be dropped, the attacking node drops only the sensitive packets. These types of nodes are difficult to identify. However, the indirect trust provided by the security node will identify the symmetrical pattern of packet drop through which the grey hole attacks can be identified and the proposed method will eliminate these nodes during packet forwarding.

Worm Hole attack: A tunnel like structure is created by two or more nodes thereby decreasing the hop count of the route. Once the source node begins to transmit the packet through this tunnel, the source node gets attacked resulting in denial of service and replay attacks. The proposed work performs clustering of mobile nodes and the cluster head is elected for each cluster. The cluster head will have the link details of each node by doing so the nodes involved in these types of attacks can be identified and are eliminated during packet transmission.

Timing attack: in this type of attack, the attacker alters the time slot of the packet passing through it thereby causing a delay to the packet in reaching the destination. Due to this manipulation the receiver may not receive the packets on time. The proposed work identifies the malicious node even before the attack takes place, hence these types of attacks can be avoided.

Intruder attack: In this type of attack, the attacker is from outside the network. These attackers enter the network by manipulating the user present in the network. Once the malicious nodes enter the network they start to attract the source nodes which will leads to packet drop and security threat to the sensitive information. The proposed method secures the authenticity of the user through multi factor registration and authentication. The credentials such as Binary format of user finger vein, Physically Unclonable Function (PUF), Password and User ID are registered hence the manipulation of malicious nodes is prevented.

The efficiency of the proposed work is expressed in terms of QoS, energy consumption and security which is depicted in Fig. 4 to 10. The proposed work has better performance when compare to existing works which is explained in terms of security strength ratio, routing overhead ratio, packet delivery ratio, attack prevention rate, end-to-end delay, throughput, route acquisition time. The summary of the proposed research work is briefed as follows:

- The user identity ID registered and the hash values for these credentials are generated by using CubeHash algorithm. These hash values are used to create the private and public keys for the encryption of the message packets during transmission.

- The network complexity and the transmission overhead are minimized by forming the cluster of nodes by using Weighted Sum computation for Clustering (WSC4C). Factors such as energy status, geometrical distance, link quality and moving direction are computed and the weighted sum of these factors is used for cluster formation and cluster head election.
- The routing of message packets is done through the reputed nodes which are selected using Dolphin Swarm Optimization (DSO) algorithm. The reputed nodes are estimated based on link stability, relative velocity, available bandwidth, energy, queue length, and trust. Through this algorithm the high link stability, low queue length and high bandwidth is achieved.
- The attack mitigation is carried out by Bayesian Direct Acyclic Graph aided Blockchain. The verification of the received packet is taken place by examining the user authentication and integrity of the data. Once the verification process is over the key to decrypt the received message packet is generated. The blocks in the blockchain technology are distributed in the hyper graph and the block independency of each block is determined thereby the proposed work improves the performance of the process.

VI. Conclusion

In this work, Dolphin swarm optimization is proposed to improve the secure routing in MANET environment. To achieve the highly secure routing of data packets the Blockchain technology is used. Each and every transaction of the nodes in the network gets stored in the blocks and the hash values for these events are generated in the blocks. The user legitimacy is verified by multi-factor authentication of user. The credentials such as Binary format of user finger vein, Physically Unclonable Function (PUF), Password and User ID are collected and hash values are created by using CubeHash algorithm. These hash values are used to generate the keys to encrypt the message packet. Then the mobile nodes are clustered by using Weighted Sum Computation for Clustering to minimize the complexity encountered in routing data packets. The node with the high WSM score is elected as the cluster head. The routing is taken placed through reputed nodes which are selected by using Dolphin Swarm Optimization (DSO) algorithm. In the destination side the packets are verified for user authentication and integrity of data and the decryption of message packets is carried out by using Bayesian Acyclic Graph aided Blockchain technology. The block independence of the blockchain is computed which will improve the security of the routing in MANET environment. In future, the work can be extended by implementing the present method in MANET-IOT and also in view of improving its performance further by integrating diverse DLT algorithms.

References

[1] Narayana, V., & Midhunchakkaravarthy, D. (2020). A Time Interval based Blockchain Model for Detection of Malicious Nodes in MANET Using Network Block Monitoring Node. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 852-857.

[2] Careem, M.A., & Dutta, A. (2020). Reputation based Routing in MANET using Blockchain. 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), 1-6.

[3] Murugan, S., & Jeyakarthic, M. (2020). An Energy Efficient Security Aware Clustering approach using Fuzzy Logic for Mobile Adhoc Networks. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 551-555.

[4] Sharma, V., Renu, & Shree, T. (2020). An adaptive approach for Detecting Blackhole using TCP Analysis in MANETs. 2nd International Conference on Data, Engineering and Applications (IDEA), 1-5.

[5] Pandey, S., & Singh, V. (2020). Blackhole Attack Detection Using Machine Learning Approach on MANET. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 797-802.

[6] Shrestha, S., Baidya, R., Giri, B., & Thapa, A. (2020). Securing Blackhole Attacks in MANETs using Modified Sequence Number in AODV Routing Protocol. 2020 8th International Electrical Engineering Congress (iEECON), 1-4.

[7] Elhoseny, M., & Shankar, K. (2020). Reliable Data Transmission Model for Mobile Ad Hoc Network Using Signcryption Technique. IEEE Transactions on Reliability, 69, 1077-1086.

[8] Chetna, Ramkumar, K., & Jain, S. (2020). Performance Comparison of Spline Curves and Chebyshev Polynomials for Managing Keys in MANETs. 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), 64-67.

[9] Fasunlade, O., Zhou, S., & Sanders, D. (2020). Comprehensive Review of Collaborative Network Attacks in MANET. 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), 1542-1545.

[10] Xia, H., Li, Z., Zheng, Y., Liu, A., Choi, Y., & Sekiya, H. (2020). A Novel Light-Weight Subjective Trust Inference Framework in MANETs. IEEE Transactions on Sustainable Computing, 5, 236-248.

[11] Biswas, A.K., & Dasgupta, M. (2020). A Secure Hybrid Routing Protocol for Mobile Ad-Hoc Networks (MANETs). 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-7.

[12] Olanrewaju, R.F., Khan, B.U., Anwar, F., Mir, R.N., Yaacob, M., & Mehraj, T. (2019). Bayesian Signaling Game Based Efficient Security Model for MANETs.

[13] Veeraiah, N., & Krishna, B.T. (2020). An approach for optimal-secure multi-path routing and intrusion detection in MANET. Evolutionary Intelligence, 1-15.

[14] Krishnan, R.S., Julie, E.G., Robinson, Y.H., Kumar, R., Son, L., Tuan, T.A., & Long, H.V. (2020). Modified zone based intrusion detection system for security enhancement in mobile ad hoc networks. Wireless Networks, 26, 1275-1289.

[15] Alappatt, V., & Joe Prathap, P. M. (2020). Hybrid cryptographic algorithm based key management scheme in MANET. Materials Today: Proceedings. doi:10.1016/j.matpr.2020.09.788.

[16] Singh, N. C., & Sharma, A. (2020). Resilience of mobile ad hoc networks to security attacks and optimization of routing process. Materials Today: Proceedings. doi:10.1016/j.matpr.2020.09.622.

[17] Sharifi, S.A., & Babamir, S.M. (2020). The clustering algorithm for efficient energy management in mobile ad-hoc networks. Comput. Networks, 166.

[18] Chen, Z., Zhou, W., Wu, S., & Cheng, L. (2020). An adaptive on-demand multipath routing protocol with QoS support for high-speed MANET. IEEE Access, 1–1.

[19] Ali Zardari, Z., He, J., Zhu, N., Mohammadani, K., Pathan, M., Hussain, M., & Memon, M. (2019). A Dual Attack Detection Technique to Identify Black and Gray Hole Attacks Using an Intrusion Detection System and a Connected Dominating Set in MANETs. Future Internet, 11(3), 61.

[20] Devika, B., & Sudha, P. N. (2019). Power optimization in MANET using topology management. Engineering Science and Technology, an International Journal, 23(3), 565-575.

[21] Vinoba, R., & Vijayaraj, M. (2020). Novel control topology with obstacle detection using RDPSO - GBA in mobile AD-HOC network. Computer Communications.

[22] Jevtic, Nenad & Malnar, Marija. (2019). Novel ETX-based metrics for overhead reduction in dynamic ad hoc networks. IEEE Access. PP. 1-1.

[23] Alnumay, W., Ghosh, U., & Chatterjee, P. (2019). A Trust-Based Predictive Model for Mobile Ad Hoc Network in Internet of Things. Sensors, 19(6), 1467.

[24] Usman, M., Jan, M.A., He, X., & Nanda, P. (2020). QASEC: A secured data communication scheme for mobile Ad-hoc networks. Future Gener. Comput. Syst., 109, 604-610.

[25] Khan, B.U., Anwar, F., Olanrewaju, R., Pampori, B.R., & Mir, R.N. (2020). A Game Theory-Based Strategic Approach to Ensure Reliable Data Transmission With Optimized Network Operations in Futuristic Mobile Adhoc Networks. IEEE Access, 8, 124097-124109.

[26] Rani, P., Kavita, .., Verma, S., & Nguyen, G. (2020). Mitigation of Black Hole and Gray Hole Attack Using Swarm Inspired Algorithm With Artificial Neural Network. IEEE Access, 8, 121755-121764.

[27] Wang, X., Zhang, P., Du, Y., & Qi, M. (2020). Trust Routing Protocol Based on Cloud-Based Fuzzy Petri Net and Trust Entropy for Mobile Ad hoc Network. IEEE Access, 8, 47675-47693.

[28] Riasudheen, H., Selvamani, K., Mukherjee, S., & Divyasree, I.R. (2020). An efficient energy-aware routing scheme for cloud-assisted MANETs in 5G. Ad Hoc Networks, 97.

[29] Liu, G., Dong, H., Yan, Z., Zhou, X., & Shimizu, S. (2020). B4SDC: A Blockchain System for Security Data Collection in MANETs. IEEE Transactions on Big Data, 1-1.

[30] El-Semary, A.M., & Diab, H. (2019). BP-AODV: Blackhole Protected AODV Routing Protocol for MANETs Based on Chaotic Map. IEEE Access, 7, 95197-95211.

[31] Josephine, J., & SenthilKumar, S. (2020). Tanimoto Support Vector Regressive Linear Program Boost Based Node Trust Evaluation for Secure Communication in MANET. Wireless Personal Communications, 1-21.

[32] NagendranathMVS, S., & Babu, A.R. (2020). An efficient mobility aware stable and secure clustering protocol for mobile ADHOC networks. Peer-to-Peer Networking and Applications, 13, 1185-1192.