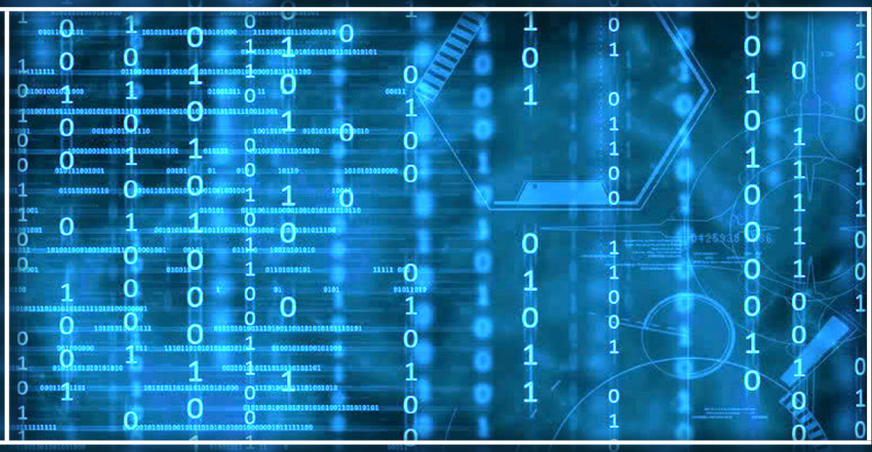
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 10 Issue 11

November 2019



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!

Managing Editor
IJACSA
Volume 10 Issue 11 November 2019
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study

Mahdi Rezapour
University of Wyoming
United States

***Abstract*—The usage of credit card has increased dramatically due to a rapid development of credit cards. Consequently, credit card fraud and the loss to the credit card owners and credit cards companies have been increased dramatically. Credit card Supervised learning has been widely used to detect anomaly in credit card transaction records based on the assumption that the pattern of a fraud would depend on the past transaction. However, unsupervised learning does not ignore the fact that the fraudsters could change their approaches based on customers' behaviors and patterns. In this study, three unsupervised methods were presented including autoencoder, one-class support vector machine, and robust Mahalanobis outlier detection. The dataset used in this study is based on real-life data of credit card transaction. Due to the availability of the response, fraud labels, after training the models the performance of each model was evaluated. The performance of these three methods is discussed extensively in the manuscript. For one-class SVM and auto encoder, the normal transaction labels were used for training. However, the advantages of robust Mahalanobis method over these methods is that it does not need any label for its training.**

***Keywords*—Credit card fraud; anomaly detection; SVM; Mahalanobis distance; autoencoder; unsupervised techniques**

I. Introduction

Fraud detection in large scale is one of the biggest challenges in fraud investigation. Annually credit fraud resulted in a loss of billions of dollars [1]. Fraud could be defined as any wrongful or criminal activity that could result in financial loss to a card holder, and personal gain of the fraudsters [2]. The two main approaches of avoiding fraud is through fraud prevention and fraud detection [3]. An application of fraud detection becomes practical when the fraudsters exceed the fraud prevention systems and start a fraudulent transaction activity. Thus, a responsibility of a fraud detection would be defined as checking any transaction with the objective of preventing a fraudster from a fraudulent activity. Credit card fraudulent activity is a most notorious activity in a financial system.

Fraudulent transaction or Outliers could be divided into two main groups: global and local outliers. A global outlier is a measured observation which has a very high or very low value relative to other observation in the dataset. On the other hand, a local outlier is a sample point with a value within a normal range of the whole dataset but compared with surrounding points, it is usually high or low. An efficient fraud detection system should be able to detect the frauds accurately and also to adjust its performance based on the changes in the behaviors of fraudsters.

Machine learning techniques are primarily methods in identifications of frauds. These techniques could be divided into two groups: supervised and unsupervised methods. In supervised machine learning techniques, a model would be trained on a past sample of fraudulent and legitimate transactions in order to classify new transactions as fraudulent or legitimate. In other words, the supervised learning uses the whole labeled dataset for training. The labels are known since card holders did identify the mismatch of a transaction, or an unusual transaction being identified by credit card agency and confirmed by a credit card holder. The supervised methods have this disadvantage that if fraudsters change their patterns, these models might not be able to detect them based on the old observations.

On the other hand, unsupervised techniques acquire information from new transactions and the anomalies would be based on updated transactions. In unsupervised fraud detection anomalies or unusual transactions would be identified as possible cases of fraudulent transactions. The advantage of unsupervised learning is that a machine does not need the knowledge of the fraud labels to train itself on, and a decision on identifying a transaction as an outlier would be made based on the distribution of the transaction. However, normal transaction labels are needed for most of the unsupervised methods so machine learning techniques would be trained on normal transaction so it can differentiate between normal and fraud for the upcoming transactions.

This study is based on real-life data of transaction from an international credit card corporation. Frauds happen barely compared with the total number of transactions so due to having an imbalanced dataset, under-sampling method was conducted to have balanced categories for comparison. Also, although unsupervised machine learning techniques do not need labeled data for the whole dataset, the data labels were used in this study for performance evaluation of different models.

II. Literature Review

Machine learning techniques for fraud/outlier detection could be divided into two main approaches: supervised and unsupervised approaches. The supervised method needs the whole data to be labeled for fraud identification meaning that it should be clarified in the dataset whether a transaction is fraud or legitimate. In many cases it is not clear whether a transaction is fraudulent or not as that transaction was not completed by the system. For these cases the analysis would favor unsupervised method. This study will go over few studies

conducted on fraud detection using various supervised and unsupervised learning and, then, it moves to the application of the unsupervised learning techniques implemented in this study.

A study conducted to compare a performance of different supervised and unsupervised method for studying credit card fraud detection [4]. Four unsupervised anomaly detection methods including one-class support vector machines (SVM), restricted Boltzmann machine, and generalized adversarial network were used as unsupervised methods. The performance across different models was compared using area under the curve (AUC).

Hidden markov model (HMM) was used to detect credit card frauds [5]. The model was initially trained with the normal behavior of a cardholder, and then evaluated on incoming credit card transactions. If an upcoming credit card transaction is not accepted by the trained HMM with high probability, it would be considered as fraud. Supervised and unsupervised machine learning techniques were combined to detect credit frauds [6]. The results showed that the hybrid technique is efficient and could improve the accuracy of detection.

A study was concerned with behavioral fraud through the analysis of longitudinal data [7]. This study implemented an unsupervised method which used changes in behavior or unusual transaction. Another study conducted with the help of unsupervised method of improved nearest neighbor method to detect intrusion [8]. The Minkowski's distance was modified and used as a means for intrusion detection.

A discussion was made about the shortcoming of one-class SVM due to its high false positive rate [9]. Thus, a new approach named enhanced SVM was proposed, which combine traditional SVM with one-class SVM to create an unsupervised machine learning. Genetic Algorithm (GA) was used as a feature selection method for extracting optimized information from raw dataset.

The following paragraphs will highlight the studies that focused on one-class SVM, autoencoder and Robust Mahalanobis distance methods respectively.

One-class SVM was used to detect anomaly problems [10]. As this method is sensitive to outliers, ramp loss function was introduced to this paper to address this issue. The objective of this function was to make sparse semi-supervised algorithm. The obtained results showed an improvement in outlier detection. In another study, one-class SVM was used for detecting anomalous windows registry accesses using registry anomaly detection (RAD). The system was compared with probabilistic anomaly detection (PAD), and the results showed that PAD outperformed the SVM model possibly due to hieratical prior incorporated on the PAD algorithm. Different machine learning techniques such as SVM, random forest, and the logistic regression model were compared for detection of credit card fraud [11]. The models performances were compared based on different metrics such as precision, sensitivity, and specificity. Two-class and one-class SVM were used and compared for detection of fraudulent credit card transactions [12]. These models were considered and evaluated using different Kernels. The results showed the superiority of one-class SVM for the anomaly detection problem over two class SVM.

Turning to studies used Autoencoder for anomaly detection, Autoencoder based on ensemble model was used as an anomaly detection method in building energy data [13]. A comparison was made across ensembles of different auto encoder models. The threshold for normal versus anomalous observations was based on the assumption that 5% of the data are anomaly candidates. Credit card fraud detection was proposed using regular autoencoder and variational autoencoder (VAE), defined as a variant of autoencoder that uses a probabilistic graph as a basic for anomaly detection [14]. Reconstruction error was used as an anomaly score for the autoencoder and a reconstruction probability. It was found that a simple regular autoencoder outperform the VAE for detecting credit card fraud. Another study used anomaly detection to identify anomaly related to deviation of practical building operation due to existence of operating faults and improper control strategies [13]. An autoencoder-based ensemble method was developed for anomaly detection in this study, A number of autoencoder, and autoencoder-based ensemble, were stacked with different architecture. Root mean square was used as a metric for the model evaluation.

One of the ways to implement multivariate outlier detection (MVO) is through Mahalanobis, or Cook's distance. Robust Mahalanobis distance has been used extensively for anomaly detection in the literature review. When Mahalanobis is used for MVO, a large (squared) Mahalanobis distance would be considered as Multivariate outliers [15]. However, Mahalanobis distance is sensitive to the presence of outliers [16] due to sensitivity of arithmetic mean and sample covariance matrix to outliers [17]. The solution of this problem could be achieved by estimating the mean and covariance matrix in a robust manner, resistance against the impact of outlying observations [15]. The minimum covariance determinant (MCD) is most commonly used estimator of multivariate location and scatter due to having a computationally fast algorithm [18]. This matrix is calculated by the subset of observation of size h, which minimize the determinant of the covariance matrix.

Multivariate outlier detection was used in exploration of geochemistry [15]. The method was able to distinguish between extreme values of a normal distribution and values obtained from different distribution. In this study Mahalanobis was used by robust estimates, which downgrade the impact of the extreme values in the solution. In order to simplify the visualization of the outliers spatially on a map, a multivariate outlier plot was introduced, which uses different symbols to illustrate the distance measures from the center of the distribution. Different colors were also used to highlight the magnitude of the distance from the center. The same methodology was implemented for interpretation of multivariate outlier for compositional dataset [19]. The isometric logratio(ilr) transformation was implemented on the data before conducting any analysis.

In this study, a robust geographically weighted method was used for multivariate spatial outlier detection [20]. Also a large Mahalanobis distance was used as a means for anomaly

detection, and in order to have a robust estimate of distance, MCD was used.

The rest of this paper is organized as follows. The data description will talk about the data used in this study. The method will go over the three unsupervised methods being implemented for detection of credit card fraud. The remaining sections will talk about results and discussion.

III. Data Description

The input data consists of numerical values resulted from principal component transformation to preserve confidentially. The response is binary classes with 1 in case of fraud and 0 otherwise. Time was in seconds indicating amounts passed between each transaction, and the first transaction in the dataset. Amount refers to transaction amount. There were 29 predictors in the dataset, including the response variables. Before including the predictors in the model, all the distribution for fraudulent and normal transaction were plotted and compared to see if it is necessary to include all the predictors in the analyses. For instance, the distribution of two predictors are depicted in Fig. 1, time on the left and amount spent on the right. The first one, left, is the time for both true or normal transaction and fraudulent or false for the response (class) whether a transaction was fraud or normal.

As can be seen from Fig. 1 on the left, the two transactions have almost similar distribution indicating that the time variable would not be much of a help for prediction of transaction types. On the other hand, for the "amount" predictor as expected the two distributions are different across different transactions as a response.

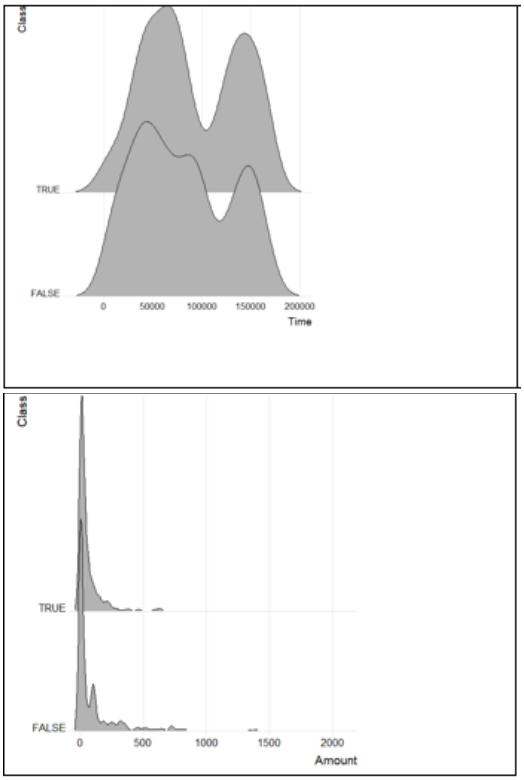


Fig. 1. Distributions of Time and Amount Spent as a Transaction Versus different Class Categories as Response.

Machine learning techniques tend to produce unsatisfactory predictions for the minority class when the data is imbalanced. Generally, when the minority class events accounts for less than 5% of the all response category in the dataset, that dataset is called imbalanced. For this dataset, fraudulent observation accounts for 0.17% of all observations. For imbalanced dataset even best of algorithms are incapable of detection of fraud from legitimate transactions, and they would face the problem of identifying too many false positive, legitimate transaction, as fraudulent ones [21]. In order to address the problem of imbalanced dataset, under sampling technique was used to convert the dataset into a balanced dataset. It has been found that random under sampling of the majority class to be generally batter than other sampling methods [22]. This method has been used in credit card fraud detection often [11]. Thus his in this study, this method was taken advantage for balancing the dataset.

IV. Method

The following section would describe different methods applied in this study to identify credit cards frauds.

A. One-Class SVM

One-class SVM is particularly useful for imbalanced dataset where there are many cases of normal data and not many cases of outliers (anomalies). The objective if this method is to see if test data is a member of a class of the training dataset or not.

The method could be viewed as a quadratic optimization problem, minimizing an objective functionω, with an objective of identifying a best algorithm to maximize the accuracy of the training dataset. To enhance an application of the model trained on test dataset, the distance between the margins, support vectors, needs to be maximized.

In SVM, the anomalies in the positive data which was used in negative samples would be identified. The one-class problem could be written as follows [23]:

$$f(x)\begin{cases}+1, if\ x \in S\\ -1, if\ x \in S\end{cases} \quad (1)$$

For this method, the algorithm maps the data into a feature space H using kernel function, and then a hyperplane would try to separate the mapped vector with maximum margin. The hyperplane could be written as:

$$\omega^T x + b = 0 \quad (2)$$

Where ω is the normal vector to the hyperplane, x is a feature, and b is an intercept.

For this model all the data points for class –1 are on one side and all the data points for class 1 are on the other side. The hyperplane searches for the maximal margin between the classes.

The training dataset is $(x_1, x_2, \ldots, x_n) \in R^n \times (\pm 1)$, where a Kernel map could be written as $\Phi: R^n \rightarrow H$, which transforms the data into feature space H. The model, then, minimize the objective function as follows:

$$\min_{\omega,b,\varepsilon} \frac{||\omega||^2}{2} + C\sum_{i=1}^{n} \varepsilon_i \quad (3)$$

Slack of ε_i would be introduced to the model to prevent the SVM from overfitting with noisy data, on the other hand, the constant of $C > 0$ governs the trade-off across maximizing the margin and the number of training data points within the margin, and $\frac{2}{||\omega||}$ is a distance between the two support vector, which is subject to:

$$y_i(\omega^T(\varphi(x_i) + b) \geq 1 - \varepsilon_i \ \ for\ all\ i = 1, \ldots, n \quad (4)$$

$$\varepsilon_i \geq 0\ for\ all\ i = 1, \ldots, n \quad (5)$$

The equation for on-class SVM is a bit different. This model separates all the data points from the origin with the objective of maximizing the distance from the hyperplane to the origin. The results of the above function would be positive for a small region and negative elsewhere. The quadratic minimization function could be written as follows:

$$\underset{\omega,b,\varepsilon}{min} \frac{||w||^2}{2} + \frac{1}{\nu n}\sum_{i=1}^{n} \varepsilon_i - \rho \quad (6)$$

$\nu \in (0,1)$ is an important Parameter that characterize the solution for the machine by controlling the trade-off between maximizing the distance from the region and containing most of the data in the region created by hyper plane.

B. Deep Autoencoder Network

Autoencoder is a branch of neural network that could be used to learn data in an unsupervised manner. The goal of this model is to learn a representation (encoding) of a dataset. This method can be applied on various objectives such as dimensionality reduction or anomaly detection. Typically, this method is trained over number of iterations with an optimizer and an objective of minimizing the cost function such as mean square error (MSE) or reconstruction error. This model performance in fraud identification is based on the assumption that the distribution of normal transaction is different that the distribution for fraudulent ones.

This model has two parts: the encoder f (mapping $X\ to\ F$), and the decoder g (mapping $F\ to\ \acute{X}$) [24]. Generally, autoencoders are symmetric, with the first half of the autoencoder is considered as encoder and the other half is considered as decoder.

It can be said that autoencoder follows the same principal as feed forward neural network and the same technique could be applied to this model. The goal of this model is to construct the inputs by minimizing the difference between the input and output, compared with feed forward neural network having an objective of predicting output as Y given input x. It should be noted for this model, the number of inputs is equal to the number of outputs. The model can be depicted in a simple way consisting of encoder and decoder functions:

$$\theta: X \to F \quad (7)$$

$$\varphi: F \to \acute{X} \quad (8)$$

$$\theta, \varphi = \underset{\theta,\varphi}{argmin} ||X - (\varphi.\theta)X||^2 \quad (9)$$

Where θ and φ are transactions where θ map the input X to F (encoder), and transaction φ which map F back to the input (decoder), and the objective is to find transactions that would minimize the objective function, which is MSE.

F, which is a Map of X, is referred as code and can be written as follows:

$$F = \sigma(Wx + b) \quad (10)$$

where $\sigma, W, and\ b$ are activation function, weight and bias, respectively. These values are initialized randomly and then updated during back propagation. After this stage the decoder, maps back F to the reconstruction $\acute{X}$

$$\acute{X} = \acute{\sigma}(\acute{W}F + \acute{b}) \quad (11)$$

As mentioned earlier the objective function or OF is to minimize the cost function such as MSE"

$$OF = minimize||X - \acute{X}||^2 = ||X - \acute{\sigma}(\acute{W}(\sigma(Wx + b)) + \acute{b}||^2 \quad (12)$$

Where $\acute{\sigma}$ and σ are activation functions and the other parameters were defined earlier.

Before conducting a main analysis, different hyper parameters were tuned to get a model with a better performance or lower error rate. In this model, fraud observation from normal observation was distinguished by mean square error (MSE), with the assumption that fraudulent transaction would have a higher MSE's. However, for the sake of a comparison, there should be a unique threshold k, MSE>K, so an observation higher than this threshold would be considered as a fraud transaction.

It is expected that frauds transaction to have a different distribution than normal transaction. In other words, autoencoder will have higher reconstruction errors on frauds than on normal transactions. Thus, the reconstruction error can be used to distinguish fraudulent transactions from the normal ones.

Again, as this method uses the same concept as neural network, and neural network models assign a higher importance to variable with higher values, the data needs to be reprocessed before conducting any statistical modeling. This scaling was conducted by dividing each cell by the difference of maximum and minimum of that column.

Two autoencoder layouts are commonly have been employed in the literature review: a bottleneck or under complete, where the number of nodes in the hidden layer is less than the number of nodes in the input layer. The second option is called over-complete layer where the number of nodes in the hidden layer is higher than the number of nodes in the input layer. Beside these two methods, denoising autoencoder (DAE) could be considered. This method objective is to achieve a good representation through changing the reconstruction criteria [25]. After creating partially corrupted input, the model would try to recover the original uncorrupted input. The amount of corrupted input that would be added to the model would be typically 30%. There are three common methods used for denoising. In the first approach, isotropic Gaussian noises would be added to the input layer based on Gaussian model. The second approach randomly selects a fraction of input variables and set them as either zero or one. For masking

noise, the third approach, a fraction of input values would be randomly selected, and their values would be masked at zeros.

Although a basic version of the autoencoder consists of three fully connected layers including one input, one output, and one hidden layer in between, the performance of auto encoder might be improved by adding new hidden layers or increasing the number of nodes. The autoencoder deployed in this study consists of four fully connected layers including one input layer, two hidden layer and one output layer. Generally, the performance of this model due to being unsupervised is evaluated by the normality of reconstruction residuals and the quality of the features extracted by autoencoder. For instance, shapiro-wilk test could be applied on reconstructed residuals to see if the distribution is normal or not.

In this study, hyperbolic tangent function (tanh) function was used for encoding and decoding the input to the output. Then, the backpropagation was used to reconstruct the error. The possible activation functions could be Relu, tanh, or sigmoid. As data contain negative value, in case of using Relu activation function, negative values would be converted to zero which, in return, would block gradient information from learning. On the other hand, since there were negative input data, the sigmoid activation function is not appropriate as it is likely to face a vanishing gradient due to its formula $(\frac{1}{1+\exp(-x)})$.

For this model stacked autoencoder was used where the number of nodes per layer decrease from a layer to a layer and would increase back in the decoder. Number of hidden layers and number of nodes are dependent on the structure of the data, number of features and number of observations. As undersampling technique was used, resulting in a reduction of the observation, only one hidden layer was selected. On the other hand, number of nodes for input layer often follows the below equation:

$$NN=2*N+1 \quad (13)$$

Where NN is a number of nodes, and N is number of features. As there were 28 features in the data set, a value of 15 was chosen for the number of input node and output layers. The number of nodes in hidden layer was selected as slightly smaller than the number of input nodes as 10.

As the dataset become balanced, in order to identify the threshold between normal and fraudulent transaction, the quality of the prediction could be evaluated by recall or precision. In the analysis not much different was identified across these two measures so the results for precision would be presented in Fig. 2. The equation of precision can be written as follows:

$$Precision = \frac{Real\ fraud\ category \cap Predicted\ as\ fraud}{Fraud\ transaction+True\ transaction} \quad (14)$$

The threshold which would result in highest precision would be selected as a boundary between fraud and normal transaction. However, after conducting few trial and errors, a value slightly lower than an identified value in Fig. 2 was identified for the threshold. For this study a value of 0.42 was chosen, Thus MSE greater than this value would be considered as abnormal (fraud) transaction and a value less than this cutting point would be considered as normal.

C. Multivariate Outliers' Detection

Usually a traditional outlier detection refers to measuring the distance between a point and distribution to which that point belongs. A classic Mahalanobis distance is a common method of measuring this distance. This method measures how many standard deviations away a point is from the mean of a distribution. The Mahalanobis distance of an observation x=(x1, x1, x1,…, x1) from a set of observation with mean of $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_n)$ would be written as:

$$MD = \sqrt{(x-\bar{x})^T C^{-1} (x-\bar{x})} \quad (15)$$

where, C is the estimated covariance matrix. $\bar{x}$ Is the estimated multivariate location, or the multivariate arithmetic mean or the centroid.

One of the methods that could be used to identify outliers based on Mahalanobis is distance-distance plot, which plots the classical Mahalanobis distance of the data against robust Mahalanobis distance based on the minimum covariance determination (MCD). The MCD is a highly robust estimators of multivariate location and scatter [18], [26]. It should be noted one of the main advantages of MCD is its resistance to outliers which also makes its use practical for different multivariate techniques such as principal component and factor analyses. It should be noted that Mahanabolis method is applicable when all the variables are continuous. However, when comparison is conducted on categorical predictors another proposed method in the literature could be applied [27]. This method is based on simple matching coefficient (SMC) which could be used for comparing similarity and diversity of sample sets for a binary predictor as follows:

$$SMC = \frac{Number\ of\ matching\ attribuates}{Number\ of\ attribuates} = \frac{M_{00}+M_{11}}{M_{00}+M_{01}+M_{10}+M_{11}} \quad (16)$$

M's measures similarity between different binary predictors, for instance, M_{00} is the total number of attributes where both binary attributes A and B have a same value of zero.

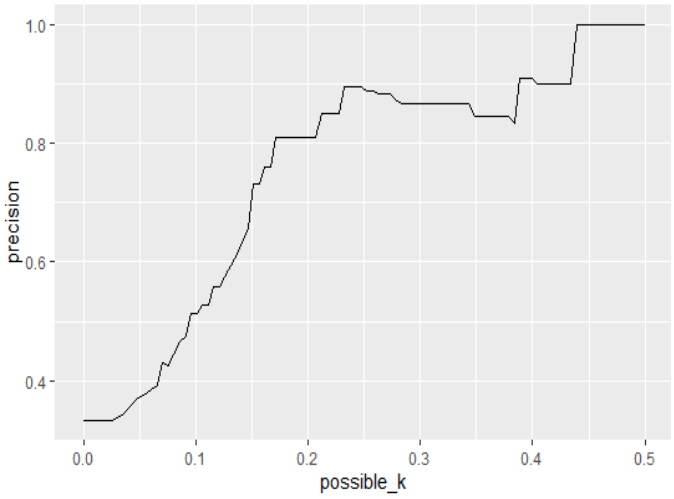


Fig. 2. Identification of Threshold for Autoencoder.

On the other hand, a Gower distance should be used for distance between two entity whose attributes have a mixed of categorical and numerical [28]. This model uses Manhattan distance for calculating distance between continuous datapoints, and Sørensen–Dice coefficient (DSC) for calculating distance between categorical datapoints. The DSC could be calculated as follows:

$$DSC = \frac{2|X \cap Y|}{|X|+|Y|} \quad (17)$$

where, |X| and |Y| are the cardinalities of the two sets.

For autoencoder dataset was normalized with minmax normalization. For one-class SVM, normalizing the data was conducted with scale function in the syntax of the model.

V. Performance Measure

Different metrics could be implemented to evaluate the performance of different models. The performance measures were used not to compare the performance of models to come up with a better model, since these models are performing differently, but to have some ideas how each model is performing in identification of frauds observation. It should be noted that although SVM and auto-encoder needs true positive labeled data for training, Mahanabolis method does not need labeled data and this model would be conducted on the whole dataset. Therefore, it would not be a fair comparison to compare these three models.

These metrics are "true positive" (TP), "false positive" (FP), True negative (TN), and "false negative" (FN). For instance, TF represents the number of normal transaction that are classified/predicted as normal. On the other hand, false positive, are the crashes that are classified as fraud while they were normal transaction.

VI. Results

A. One-Class SVM

For this algorithm, the features were scaled and centered with a logical vector of "scale" in R. Different kernel types were available for this model including linear, and nonlinear such as polynomial and radial basis. The results of using different kernel function indicated that linear basis function kernel would result in an optimal result. This model is based on squared Euclidean distance between two features vectors. As this model needs to be trained on the normal transaction, the data labels were used and divided into two sets of data, normal versus fraudulent transactions. The model was trained on the normal dataset. It, then was examined across the whole dataset including normal and fraudulent observations.

B. Autoencoder

Keras package was used for construction of autoencoder. The construction of this model in Keras is similar to multilayer perceptron models. Similar to SVM, the model was trained over normal dataset and then tested over the whole dataset. The inputs were transformed into 15 nodes and they were transferred into two hidden layer with 10 and 15 nodes. The model, then, recreated the input from the transferred output. In the compilation section of the model, the mean square error (MSE) was set as a loss function to identify the outlier based on higher values of MSE, while adam was set as an optimizer. Due to highlighted reasons in a previous section, tanh activation functions were set for the three layers, input, hidden and output layer. For the model fit section of Keras package, the input and output were set as the normal transactions. After identification of the threshold of normal and possible fraudulent transactions, each transformed input was defined based on its value versus the threshold. The MSE of each observation (transaction) was calculated from the below equation:

$$\text{MSE}_{observation} = \sum_{i=1}^{966}\sum_{j=1}^{29}(Real\ observation - predicted\ observation)^2 \quad (18)$$

where, i is a number of observations, which was 966 transactions, and j is a number of columns or features.

After construction of predicted input, a decision of normal versus fraudulent transactions would be made based on following equations:

$$\text{Normal transaction: MSE}_{observation} < K \quad (19)$$

$$\text{Fraudulent transaction: MSE}_{observation} > K \quad (20)$$

Where k is a threshold calculated from Fig. 2.

C. Multivariate Outlier Detection

This section would highlight the application of Multivariate Outlier Detection for credit card fraud detection. The ordered squared robust Mahalanobis distances of the observations against the empirical distribution function is presented in Fig. 3. The Mahalanobis distance of the data point against the robust Mahalanobis distance was plotted based on MCD estimators [29].

Alpha, amount of observations used for calculating the adjusted quantile, was set as .80. Value of 0.8 was chosen as a lower value was resulting in computationally singular output. This resulted from invertible design matrix which in return is due to multicollinearity across predictors being used for MCD estimations.

On the other hand, amount of observation used for calculating the adjusted quantile was also set as 0.8. This value, Adjusted quantile, is a new threshold that separate outliers from non-outliers [30]. An approximation of 97.5 percentile would be obtained by estimation of mean and standard deviation of each variable and computing the values of mean $\pm$standard deviation [30]. An observation would be considered as an outlier if that observation falls in this extreme 2%. In Fig. 3, the horizontal and vertical lines are plotted at values equal to the cutoff, where the default is square root of the 97.5 distribution.

Table I presented different error rates for different categories across the three implemented machine learning techniques. This study is not discussing the input of Table I in detail as these tree models are performing differently. It should be noted although robust Mahanabolis resulted in a worst performance contrary to the other two methods this method did not used any label for training.

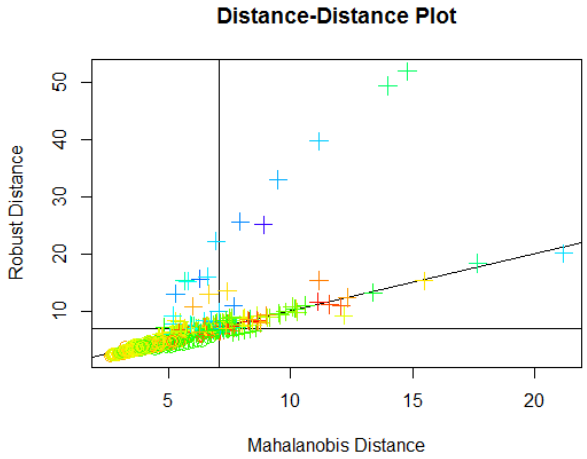


Fig. 3. Distance-Distance Plot: the Robust Distance Versus the Classical Mahalanobis Distance.

TABLE. I. ERROR RATE OF DIFFERENT INCLUDED MODELS

One-class SVM			Autoencoder			Robust Mahanabolis		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
	0	1		0	1		0	1
0	279	195	0	430	44	0	360	114
1	67	425	1	63	429	1	245	247

VII. CONCLUSION

Due to huge loss to banks, individuals and insurance companies, credit card fraud detection is considered as one of most explored domains of fraud detection. In data evaluation, anomaly is referred to any observation that does not conform to the expected distribution/pattern of the other items. At the age of computer, this could refer to an adverse event such as network intrusion, bank or credit frauds. The problem with supervised leaning techniques is that they need labels for all the observations to predict the future transactions. This would create a problem when fraud transactions need to be detected and no label is available for these observations. Moreover, fraudsters change their habit constantly which make it difficult for supervised techniques to be prepared for those transactions. However, unsupervised techniques need only labels for one-class, usually normal class, and it could predict the future observations based on distance from normal observations.

On the other hand, some supervised learning techniques even do not need the label for one-class, and they can identify the outliers on the whole dataset with no labels. In this study we took advantages of the aforementioned techniques and we used the available data labels to check the model performance.

This paper presents application of three unsupervised methods in detection of credit card frauds. For unsupervised methods SVM and autoencoder the training would be achieved from past normal transactions to predict future transactions, normal versus fraud. However, the advantage of mahalanobis method over the other two method is that this method does not need to be trained on labeled data and it can identify the anomalies just based on the minimum covariance determinant matrix. As the performance and training the three models are different, no comparison was made across these three models. However, to have a vision about the performance of these models the available labels were used for models' performance evaluations.

For the future studies the information related to cardholder behaviors and their historical transaction history need to be taken into consideration for achieving a higher accuracy. In other words, both global and local outliers need to be considered for those studies.

REFERENCES

[1] P. K. Chan et al, "Distributed data mining in credit card fraud detection," IEEE Intelligent Systems, (6), pp. 67-74, 1999.

[2] A. Hobson, The Oxford Dictionary of Difficult Words. 2004.

[3] Y. G. Şahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," 2011.

[4] X. Niu, L. Wang and X. Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised," arXiv Preprint arXiv:1904.10604, 2019.

[5] A. Srivastava et al, "Credit card fraud detection using hidden Markov model," IEEE Transactions on Dependable and Secure Computing, vol. 5, (1), pp. 37-48, 2008.

[6] F. Carcillo et al, "Combining unsupervised and supervised learning in credit card fraud detection," Inf. Sci., 2019.

[7] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," Credit Scoring and Credit Control VII, pp. 235-255, 2001.

[8] S. Jiang et al, "A clustering-based method for unsupervised intrusion detections," Pattern Recog. Lett., vol. 27, (7), pp. 802-810, 2006.

[9] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," Inf. Sci., vol. 177, (18), pp. 3799-3821, 2007.

[10] Y. Tian et al, "Ramp loss one-class support vector machine; A robust and effective approach to anomaly detection problems," Neurocomputing, vol. 310, pp. 223-235, 2018.

[11] S. Bhattacharyya et al, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, (3), pp. 602-613, 2011.

[12] M. Hejazi and Y. P. Singh, "One-class support vector machines approach to anomaly detection," Appl. Artif. Intell., vol. 27, (5), pp. 351-366, 2013.

[13] C. Fan et al, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data," Appl. Energy, vol. 211, pp. 1123-1135, 2018.

[14] T. Sweers, T. Heskes and J. Krijthe, "Autoencoding Credit Card Fraud," 2018.

[15] P. Filzmoser, R. G. Garrett and C. Reimann, "Multivariate outlier detection in exploration geochemistry," Comput. Geosci., vol. 31, (5), pp. 579-587, 2005.

[16] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," Journal of the American Statistical Association, vol. 85, (411), pp. 633-639, 1990.

[17] F. R. Hampel et al, Robust Statistics: The Approach Based on Influence Functions. 2011196.

[18] P. J. Rousseeuw, "Multivariate estimation with high breakdown point," Mathematical Statistics and Applications, vol. 8, (283-297), pp. 37, 1985.

[19] P. Filzmoser, K. Hron and C. Reimann, "Interpretation of multivariate outliers for compositional data," Comput. Geosci., vol. 39, pp. 77-85, 2012.

[20] P. Harris et al, "Multivariate spatial outlier detection using robust geographically weighted methods," Mathematical Geosciences, vol. 46, (1), pp. 1-31, 2014.

[21] M. Krivko, "A hybrid model for plastic card fraud detection systems," Expert Syst. Appl., vol. 37, (8), pp. 6070-6076, 2010.

[22] J. Van Hulse, T. M. Khoshgoftaar and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in Proceedings of the 24th International Conference on Machine Learning, 2007,.

[23] B. Schölkopf et al, "Estimating the support of a high-dimensional distribution," Neural Comput., vol. 13, (7), pp. 1443-1471, 2001.

[24] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. 2016.

[25] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. 2016.

[26] P. J. Rousseeuw, "Least median of squares regression," Journal of the American Statistical Association, vol. 79, (388), pp. 871-880, 1984.

[27] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining." DMKD, vol. 3, (8), pp. 34-39, 1997.

[28] J. C. Gower, "A general coefficient of similarity and some of its properties," Biometrics, pp. 857-871, 1971.

[29] P. Filzmoser, R. Maronna and M. Werner, "Outlier identification in high dimensions," Comput. Stat. Data Anal., vol. 52, (3), pp. 1694-1711, 2008.

[30] P. Filzmoser, R. G. Garrett and C. Reimann, "Multivariate outlier detection in exploration geochemistry," Comput. Geosci., vol. 31, (5), pp. 579-587, 2005.

On the Possibility of Implementing High-Precision Calculations in Residue Numeral System

Otsokov Sh.A[1]
Dept. of Computing Machines, Systems and Networks
National Research University "Moscow Power Engineering Institute" Moscow, Russian Federation

Magomedov Sh.G[2]
Dept. of Intelligent Information Security Systems
MIREA Russian Technological University
Moscow, Russian Federation

***Abstract*—This article proposes a method for accelerating high-precision calculations by parallelizing arithmetic operations of addition, subtraction and multiplication. The proposed approach allows us to apply the advantages of the residue numeral system: absence of carry-overs when adding, subtracting, multiplying and reducing high-precision calculations with numbers of high digit capacity to parallel and independent execution of arithmetic operations with numbers of low digit capacity across many modules. Due to the complexity of performing non-modular operations such as: inverse transformation into a positional numeral system, number comparisons, sign identification and number rank calculation in a residue numeral system, the effect of acceleration of high-precision calculations is possible when solving some computational problems with a small number of non-modular operations, for example: determination of the scalar product of vectors, discrete Fourier transformation, iterative solution of systems of linear equations by the methods of Jacoby, Gaussa-Zeidel, etc. Implementation of the proposed method are demonstrated by the example of finding the scalar product of vectors.**

Keywords*—High-precision calculations; residue numeral system; positional numeral system; number conversion; rank determination

I. Introduction

The double-precision floating-point format, supported by modern computer processors, is sufficient to solve many computational problems. However, certain tasks exist in computational practice, for example, in the fields of nanoelectronics, nuclear physics, robotics, computational geometry and others, where high-precision computer calculations are required [1-4] and traditional double-precision floating-point computer calculations provide the wrong result [5-6].

High-precision calculations have been programmatically implemented and for various programming languages there are libraries and packages supporting floating-point calculations of arbitrary accuracy, for example, ZREAL, MParith, GMP etc. [7-11]

A significant drawback of such libraries and packages is a sharp decrease in the computational speed with increasing accuracy or the length of the mantissa of a floating-point number. Attempts to speed up high-precision calculations at the level of floating-point arithmetic algorithms do not provide significant gains due to the fact that arithmetic operations of addition, subtraction and multiplication of numbers in a positional numeral system are poorly parallelized due to inter-digit carry.

These studies were conducted with the support of RTU MIREA within the framework of the initiative research work of MSEC-5 "Development of an automated procurement management system".

The purpose of this article is to propose a method for speeding up high-precision calculations by switching to a residue number system (RNS) in which the operations of addition, subtraction and multiplication can be parallelized. In RNS integers are represented by their values modulo several pairwise coprime integers called the moduli, and arithmetic operations are performed in parallel and independently for each of the moduli [12-16]. The final result of these calculations is converted to a positional numeral system.

Speeding up of arithmetic operations is achieved due to the fact that parallel calculations are performed with low-digit numbers.

The disadvantages of RNS include the difficulty of performing such operations as division, comparison, left and right shifts, rounding, converting a number to a positional numeral system and others, that are called non-modular operations [12]. Therefore, modular arithmetic is mainly used for tasks in which no or a small number of non-modular operations are required, for example, determining the scalar product of vectors, etc. [17].

Modular arithmetic is integer, but arithmetic operations with rational numbers can also be carried out in RNS [16].

Authors in [18-19] describe the method for representing numbers and the algorithms for performing arithmetic operations in RNS, including division. In this article two ways of representing floating-point numbers in RNS are discussed and the possibilities for parallelizing arithmetic operations of addition, subtraction and multiplication are explored.

In the next section a possible way of representing numbers in RNS and a mixed numeral system is considered.

II. Residue Number System

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size.

Consider a set of integers p_1, p_2, ..., p_n, called the moduli, such that the following inequality holds:

$$2=p_1<p_2<, ..., <p_n \quad (1)$$

Let P be the product of all the moduli:

$$P=\prod_{i=1}^{n} p_i \quad (2)$$

According to the Chinese remainder theorem [14], all integers belonging to the range.

$$[0, \ldots, P-1] \quad (3)$$

have a unique representation in the residue number system by the moduli (1).

Since p_1, p_2, ..., p_n are prime numbers and one of them equals 2, their product P is an even number and the following ranges are used to represent positive and negative numbers [12]:

$$[0, ..., P/2-1], [P/2, ..., P-1] \quad (4)$$

for positive and negative numbers respectively.

Thus, in RNS all integers from the following range are unambiguously represented:

$$[-P/2, ..., P/2] \quad (5)$$

Let $P/2>q\text{^}(n_f)-1$.

Any integer belonging to the range (2) has a unique representation in the mixed number system.

III. Representation of Floating-Point Numbers in RNS

Consider the following representation of floating-point numbers:

$$A=K\cdot q^t \quad (6)$$

where A is a floating-point number, K is the mantissa of A, an integer such that $|K|\le q\text{^}(n_f)-1$, q is the base of the numeral system, t is the order, an integer such that $|t|\le k_f$, n_f is a natural number characterizing the length of the mantissa of the floating-point number, k_f is a natural number characterizing the maximum order of representable numbers.

The floating-point format (6) differs from the traditional floating-point number format [6] and is more convenient for representation in RNS.

Table I shows the maximum and minimum positive and negative numbers representable in (6):

The range of representable numbers in (6) is the following:

$$(-q\text{^}(n_f+k_f), q\text{^}(n_f+k_f)) \quad (7)$$

Consider the following representation of the floating-point format (6) in RNS:

$$A=[(\alpha_1,\alpha_2,...,\alpha_i,...,\alpha_n),t] \quad (8)$$

where $\alpha_i=|K|_{pi}$.

TABLE. I. Representable Number Ranges

Maximum positive	$q\text{^}(n_f+k_f)$
Minimum positive	$q\text{^}(-k_f-n_f)$
Minimum negative	$-q\text{^}(n_f+k_f)$
Maximum negative	$-q\text{^}(-k_f-n_f)$

Given the complexity of performing non-modular operations in RNS, in particular the left shift, alignment of orders during addition and subtraction is performed by the right shift.

For the same reason, because of the complexity of the effective implementation of the normalization operation, this article describes an unnormalized floating-point format (8) and there is no normalization operation.

In the next section the rules for performing arithmetic operations of addition, subtraction and multiplication are considered.

IV. Rules for Performing Arithmetic Operations with a Floating Point in RNS

Consider two floating-point numbers in the format (8):

$$A_1=[(\alpha_1,\alpha_2,...,\alpha_i,...,\alpha_n),t_1] \quad (9)$$

$$A_2=[(\beta_1,\beta_2,...,\beta_i,...,\beta_n),s_1]$$

1) The product of the numbers $A_1\cdot A_2$ is

$$A_3=A_1\cdot A_2=[(\chi_1,\chi_2,...,\chi_i,...,\chi_n),t_1+s_1] \quad (10)$$

where

$$(\chi_1,\chi_2,...,\chi_n)=(\alpha_1\cdot\beta_1 \bmod p_1,\alpha_2\cdot\beta_2 \bmod p_2,...,\alpha_n\cdot\beta_n \bmod p_n)$$

2) The sum of the numbers A_1+A_2 is

$$A_4=A_1+A_2=K_4\cdot q^z$$

Let $s_1>t_1$. Then

$$A_1\pm A_2=K_1\cdot q\text{^}(t_1)\pm K_2\cdot q\text{^}(s_1)=q\text{^}(t_1)\cdot(K_1\pm q\text{^}(s_1-t_1)\cdot K_2)=$$

$$=[(\delta_1,\delta_2,...,\delta_n),t_1] \quad (11)$$

where

$$(\delta_1,\delta_2,...,\delta_n)=(\alpha_1\pm q\text{^}(s_1-t_1)\cdot\beta_1 \bmod p_1,\alpha_2\pm q\text{^}(s_1-t_1)\cdot\beta_2 \bmod p_2,...,\alpha_n\pm e\cdot q\text{^}(s_1-t_1) \bmod p_n),\ z=t_1$$

The diagram of high-precision calculations in modular arithmetic is shown in Fig. 1.

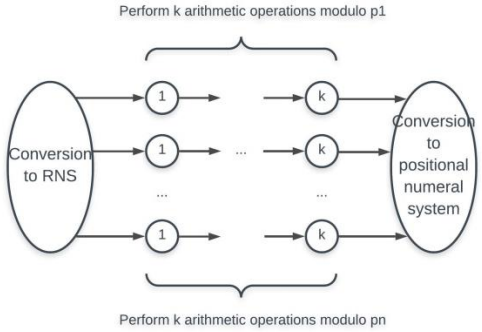


Fig. 1. Diagram of High-Precision Calculations in Modular Arithmetic.

In accordance with this diagram, the source data is converted from a positional numeral system to RNS by moduli (1). Then k arithmetic operations are carried out and the result is converted to a positional numeral system.

Using formulas (10) and (11), it is possible to carry out arithmetic operations with floating-point numbers according to the rules of modular arithmetic in parallel and independently for individual moduli. In this case, the mantissa and orders are calculated in parallel and independently from each other by formulas (10) and (11). Complexity arises when the result is outside of the allowed range defined by the moduli. In this regard, two questions arise:

1) How to choose the moduli so that the result is not outside the allowed range?

2) How to convert results from RNS to a positional numeral system?

To answer the first question, it is enough to estimate the top boundary of the result and choose the moduli so that in RNS all numbers smaller than this estimate are uniquely represented. In the next section an estimation method using the example of calculating the scalar product of two vectors is considered [20].

V. Method for Choosing the Moduli

Let us estimate the order of the result of the scalar product of two vectors.

Consider two vectors $X=(x_1,x_2,...,x_k)$, $Y=(y_1,y_2,...,y_k)$

The scalar product is determined by the formula:

$$(X,Y)=x_1\cdot y_1+x_2\cdot y_2+...+x_k\cdot y_k \quad (12)$$

Consider two floating-point numbers in the format (6).

Then for multiplication the following is true:

$$A_3=A_1\cdot A_2=K_1\cdot K_2\cdot q^{(t+s)}<q^{(n_f)}\cdot q^{(n_f)}\cdot q^{(t+s)}=$$

$$=q^{(2\cdot n_f)}\cdot q^{(t+s)}$$

which means

$$A_3=K_3\cdot q^{(t+s)},\ |K_3|\le q^{(2\cdot n_f)},\ |t+s|\le 2\cdot k_f \quad (13)$$

Substituting $x_i\cdot y_i=K_i\cdot q^{(t_i)}$ in (12) using the format (6) and considering (13) the scalar product can be presented as:

$$(X,Y)=K_1\cdot q^{(t_1)}+K_2\cdot q^{(t_2)}+...+K_k\cdot q^{(t_k)}$$

Let $t_1=\min(t_1,...,t_k)$. Then

$$(X,Y)=q^{(t_1)}\ (K_1+K_2\cdot q^{(t_2-t_1)}+...+K_k\cdot q^{(t_k-t_1)})\le$$

$$\le q^{(t_1)}\ (K_1+q^{(2\cdot n_f)}\cdot q^{(4\cdot k_f)}+...+q^{(2\cdot n_f)}\cdot q^{(4\cdot k_f)})\le$$

$$\le q^{(t_1)}\ k\cdot q^{(2\cdot n_f)}\cdot q^{(4\cdot k_f)}$$

From this expression it follows that the maximum value of the mantissa for the scalar product (12) is

$$k\cdot q^{(2\cdot n_f)}\cdot q^{(4\cdot k_f)}$$

Then to represent the mantissa of the result of the scalar product in RNS the moduli should be chosen so that the following inequality holds:

$$P/2\ge k\cdot q^{(2\cdot n_f)}\cdot q^{(4\cdot k_f)} \quad (14)$$

To answer the second question regarding the conversion of numbers from RNS to a positional numeral system, an auxiliary method for determining the rank of a number that is used in the conversion process is considered.

VI. Auxiliary Method to Determine the Rank of a Number

Any integer $A\in[0,P-1]$ can be represented as:

$$A=\sum_{i=1}^{n} B_i\cdot\gamma_i - rank\cdot P \quad (15)$$

Where B_i are orthogonal bases, *rank* is the largest positive integer such that $A<P$.

Orthogonal bases B_i are constants for RNS with given moduli and are determined by the formulas:

$$B_i=m_i\cdot P/p_i, \quad (16)$$

$$m_i=\left|\frac{P}{p_i}\right|_{p_i}^{-1}$$

Maximum possible value of *rank* is determined by the following equations:

$$\sum_{i=1}^{n} B_i\cdot\beta_i \le \sum_{i=1}^{n} B_i\cdot(p_i-1)<$$

$$<\sum_{i=1}^{n} B_i\cdot p_i=\sum_{i=1}^{n} m_i\cdot\frac{P}{p_i}\cdot p_i=\sum_{i=1}^{n} m_i\cdot P=$$

$$=P\cdot(\sum_{i=1}^{n} p_i-n)$$

This shows that

$$rank<(\sum_{i=1}^{n} p_i-n)$$

Consider an auxiliary method to determine the rank of a number.

An additional module is introduced

$$p_{n+1}>\sum_{i=1}^{n} p_i-n$$

From (15) follows:

$$rank=\left(\sum_{i=1}^{n} B_i\cdot\gamma_i - A\right)\cdot P^{-1}$$

then:

$$rank=\left(\sum_{i=1}^{n}|B_i|_{p_{n+1}}\cdot|\gamma_i|_{p_{n+1}}-|A|_{p_{n+1}}\right)\cdot|P^{-1}|_{p_{n+1}} \quad (17)$$

In (17) all calculations are performed modulo p_{n+1}.

The diagram for quick rank calculation is shown in Fig. 2.

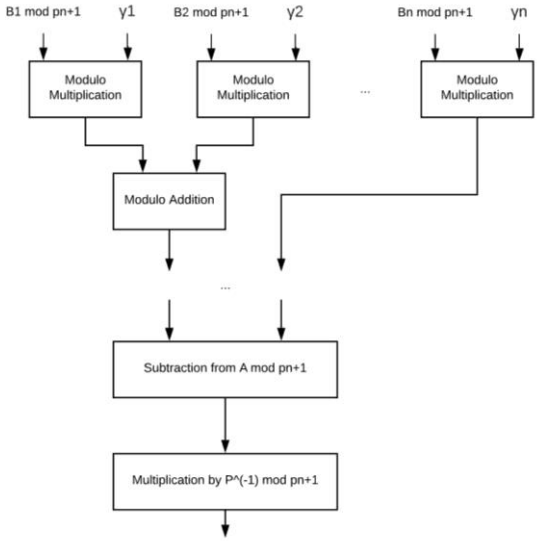


Fig. 2. Diagram for quick rank calculation.

According to the above diagram, the rank value can be calculated by the formula (17) using $\lceil \log_2 n \rceil + 2$ modular operations of addition and multiplication.

Next, a method of converting numbers from RNS to a positional numeral system is considered.

VII. Converting Numbers from RNS to a Positional Numeral System

Let A be a number in the format (8) to be converted to a positional numeral system in the floating-point format with the mantissa length n_f

$A=[(\gamma_1,\gamma_2,...,\gamma_i,...,\gamma_n), t]$

A method for accelerated calculation of the value of expression (15) in a signed-digit numeral system with numbers in the range [-6, ..., 6] is considered.

Let B_i, P be constants in the floating-point format with the mantissa length n_f.

The expressions

$$j \cdot B_i, j=1..p_i-1, \quad (18)$$

$$j \cdot P, j = 1..(\sum_{i=1}^{n} p_i - n)$$

Are constants and can be stored in computer memory, then the conversion process can be represented in the form of the following diagram (Fig. 3).

According to the diagram, the result of the conversion can be calculated by the formula (15) in $\lceil \log_2 n \rceil + 1$ steps for a known value of *rank*. The adjustment block shown in Fig. 3 normalizes the mantissa of the result. If the mantissa contains more than n_f digits, then the order of the result should be reduced by the number of discarded digits.

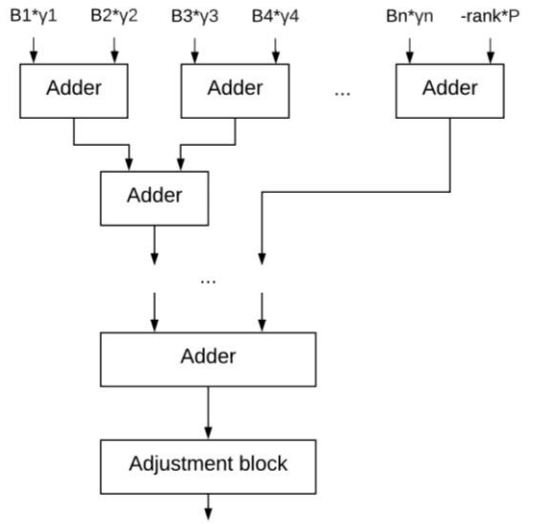


Fig. 3. Diagram for Conversion from RNS to a Positional Numeral System.

VIII. Example

Given:

nf=2, kf=2, q=10

Two vectors X=(1· 102,5· 10-1,3· 102), Y=(1,2,3)

Then according to (14)

P/2≥3· 104· 108=3· 1012

The chosen moduli are:

p1=2269, p2=2437, p3=2791, p4=3169

P=48907121298487

B1=35737332354734

B2=32129791218251

B_3=42756494435080

B_4=36097745887397

The result of the scalar product in RNS is

$(X,Y)=[(1736,1586,1214,836), -1]$

After conversion from RNS to the floating-point format with a given length of the mantissa the following can be obtained:

$(X,Y)=40\cdot 10^1$

IX. Conclusion

This article proposes an approach to speeding up high-precision computing with the following limitations:

1) It should be used for tasks with a predetermined number of operations for which the order of the result can be estimated.

2) Three modular arithmetic operations are considered: addition, subtraction and multiplication (as well as division by a constant, which can be replaced by multiplication).

3) The remaining non-modular operations such as rounding and comparison are difficult to implement in RNS. Therefore, the proposed approach will be effective for the tasks that contain a small number of non-modular operations. In addition to the scalar product, such tasks include, for example, the discrete Fourier transform, iterative solution of linear equation systems by Jacoby, Gaussa-Zeidel methods, solution of the Cauchy problem by Euler method, etc.

4) Possible direction of further research is the software implementation of high-precision calculations in RNS arithmetic on GPU.

REFERENCES

[1] D.H. Bailey, R. Barrio, J.M. Borwein, “High-precision computation: Mathematical physics and dynamics” in Applied Mathematics and Computation, Vol. 218, No. 20, 2012, pp. 10106-10121.

[2] D. H. Bailey, J. M. Borwein, “High-Precision Arithmetic in Mathematical Physics” in Mathematics, 3 (2015), pp. 337–367.

[3] D.H. Bailey, “High-Precision Computation and Mathematical Physics” Lawrence Berkeley National Laboratory, 2009.

[4] Demidova L., Nikulchev E., Sokolova Yu. (2016). BIG DATA classification using the svm classifiers with the modified particle swarm optimization and the svm ensembles. international journal of advanced computer science and applications T.7, №5, P. 294-312.

[5] “Non-obvious features of real numbers” [www.delphikingdom.com/asp/viewitem.asp?catalogid=374]

[6] A.A. Amosov, Y.A. Dubinskiy, N.V. Kopchenkova, Computational Methods for Engineers. Moscow: “High School”, p. 544, 1994.

[7] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, P. Zimmermann, “MPFR: a multiple-precision binary floating-point library with correct rounding” in ACM Transactions on Mathematical Software, Vol. 33, No. 2, Article No. 13, 2007.

[8] MParithm - package for high precision computation, 2015 [www.wolfgang-ehrhardt.de/mp_intro.html]

[9] GNU Scientific Library 2.5 released, 2018 [https://savannah.gnu.org/forum/forum.php?forum_id=9175]

[10] “Operations with multi-bit real numbers of ZReal type” [http://ishodniki.ru/list/index.php?action=name&show=pascal-math&cat=11]

[11] D.H. Bailey, X.S. Li, B. Thompson, “ARPREC: An arbitrary precision computation package” Sep 2002 [http://crd.lbl.gov/~dhbailey/dhbpapers/arprec.pdf].

[12] I.I. Dzegelenok, Sh.A. Otsokov, “Algebraization of numerical representations in providing high-precision supercomputer calculations” in Vestnik MPEI, No 3, 2010, pp. 107-116.

[13] I.Ya. Akushskiy, D.I. Yuditskiy, Machine arithmetic in residual classes. Moscow: Soviet Radio, p. 440, 1968.

[14] R. Graham, D. Knuth, O. Patashnik, Concrete math. Computer Science Foundation. Williams, p. 784, 2015.

[15] A.R. Omondi, B. Premkumar, “Residue Number Systems: Theory and Implementation”, Imperial College Press, 2007.

[16] Magomedov Sh. Organization of secured data transfer in computers using sign-value notation. ITM Web of Conferences. 2017. T. 10. DOI: 10.1051/itmconf/20171004004.

[17] R.A. Solovyev, E.S. Balaka, D.V. Telpukhov, “A device for calculating the scalar product of vectors with error correction based on a system of residual classes” in Problems of developing promising micro- and nanoelectronic systems. Vol. IV, A.L. Stempkovskiy Ed. Moscow: The Institute for Design Problems in Microelectronics RAS, 2014, pp. 173-178.

[18] E. Kinoshita, H. Kosako, Y. Kojima, “Floating-point arithmetic algorithms in the symmetric residue number system” in Computers, IEEE Transactions on, vol. C-23, No. 1, 1974, pp. 9–20.

[19] K. S. Isupov, A. N. Mal’tsev, “A parallel-processing-oriented method for the representation of multi-digit floating-point numbers” in Vychislitel’nyye metody i programmirovaniye, 15:4 (2014), pp. 631–643.

[20] A. Lebedev, S. Magomedov. “A tool for automatic parallelization of affine programs for systems with shared and distributed memory”. Russian Journal of Technology. 2019;7(5):P.7-19. https://doi.org/10.32362/2500-316X-2019-7-5-7-19.

A Hybrid Model of Autoregressive Integrated Moving Average and Artificial Neural Network for Load Forecasting

Lemuel Clark P. Velasco[1], Daisy Lou L. Polestico[2], Gary Paolo O. Macasieb[3]
Michael Bryan V. Reyes[4], Felicisimo B. Vasquez Jr[5]
Mindanao State University-Iligan Institute of Technology
Premier Research Institute of Science and Mathematics
Iligan City, The Philippines

***Abstract*—The complementary strengths and weaknesses of both statistical modeling paired with machine learning has been an ongoing technique in the development and implementation of forecasting models that analyze the dataset's linear as well as nonlinear components in the generation of accurate prediction results. In this paper, autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN) were implemented as a hybrid forecasting model for a power utility's dataset in order to predict the next day's electric load consumption. ARIMA and ANN models were serially developed resulting to the findings that out of the twelve evaluated ARIMA models, ARIMA (8,1,2) exhibited the best forecasting performance. After identifying the optimal ANN layers and input neurons, this study showed that out of the six evaluated supervised feedforward ANN models, the ANN model which employed Hyperbolic Tangent activation function and Resilient Propagation training algorithm also exhibited the best forecasting performance. With Zhang's ARIMA and ANN hybridization technique, this study showed that the hybrid model delivered Mean Absolute Percentage Error (MAPE) of 4.09% which is within the 5% internationally accepted forecasting error for electric load forecasting. Through the findings of this research, both the ARIMA statistical model and ANN machine learning approaches showed promising results in being implemented as a forecasting model pair to analyze the linear as well as non-linear properties of a power utility's electric load data.**

Keywords*—Hybrid model; autoregressive integrated moving average; electric load forecasting; Artificial Neural Network (ANN)

I. INTRODUCTION

Use of individual machine learning and statistical modeling has been in the forefront of predictive analytics due to their promising abilities to deliver close to accurate forecasting results. Autoregressive Integrated Moving Average (ARIMA) is a statistical modeling approach which has been widely used in forecasting with promising accuracy brought about by efficient linear representation exhibited by non-linear systems [1-4]. With its strength in modeling time series data such as consumed electric load as well as datasets with validated linearity properties, ARIMA forecasting applications is a growing body of researches with various applications in different fields [3]. Artificial Neural Networks (ANN) is a machine learning tool that finds patterns in the same way that biological neural networks develop association along with mathematical equivalent elements capable of processing like that of the human brain [2, 4, 5]. Compared to that of ARIMA, ANN has the ability to learn from non-linear datasets due to its strength of being adaptively formed from the implemented features of its own dataset. Despite their differences in the kind of data that they can accommodate, ARIMA and ANN hybrid forecasting methodologies as well as modelling techniques are being widely developed due to the potential of generating better predictive performance than individually utilizing each model [2, 6]. For the purpose of optimal predictive performance, the main challenge in the hybridization of these machine learning and statistical modeling approaches relies on the optimal match between the data they are processing along with the forecasting ability that they enforce in their inherent unique advantages. This gives data modelers the challenge beyond the functions of data preparation and explore on the performance analysis of the ARIMA and ANN hybridization technique that can yield optimal predictive results.

Datasets such as electric load data bearing recorded consumption behavior through time has linear along with non-linear properties [5, 7, 8]. A combinatory modelling technique of these two models with ARIMA to handle the dataset's linearity and ANN to handle the non-linearity can result to more efficient forecasting outputs than just independently using one of them. This combined model is also suitable for both one-step ahead and multistep ahead predictions in generating better hybrid model performance for natural and economic datasets [1, 9, 10]. Power systems from the functions of generation, transmission and distribution can make use of historical load data in the development of load forecasting models that can aid decision makers in operations planning. Forecasting the load consumption in different time frame granularities whether it be week-ahead, day-ahead or hour-ahead predictions can benefit power systems in the provision of demand information used to carry out required actions and planning processes that ensure reliable power systems [5, 11]. With historical electric load consumption data bearing both linear and non-linear properties, a specified hybrid model imploring the unique strengths of both ARIMA and ANN can be a potential implementation in the generation of close to accurate electric load forecast. The presence of historical load datasets as well as the absence of forecasting models that can

process these datasets for tactical and strategic operations planning is a common problem faced by power generation, transmission and distribution entities. This research aims to present a hybrid model of ARIMA and ANN in the processing of electric load data in order to generate optimal day-ahead forecasting results, Despite availability of literatures that explore the hybridization of ARIMA and ANN among non-electricity related datasets, this study aims to present a personalized hybrid model specific in the processing and prediction of electric load [1, 4, 6, 10]. With an exploration on electric load data preparation and hybridization technique in ARIMA and ANN, the results of this study hopes to aid data modelers and decision makers in the development and implementation of a forecasting model that utilizes both machine learning and statistical modeling for better management of power systems.

II. Methodology

A. ARIMA Modeling

Historical electric load data of a power utility from December 27, 2013 to October 21, 2014 for a total of 28, 704 records were chosen as inputs to create both ARIMA and ANN models. As shown in Table I, the dataset comes in 15-minute records of the date, time, kilowatt delivered (KW_DEL), kilowatt per hour delivered (KWH_DEL) and kilo volt amps reactive hours delivered (KVARH_DEL). The column utilized as input to the models was kilowatt delivered (KW_DEL) since this is represents the consumed electric load which can be used to determine the predicted values for the next day. The entire 15-minute data of October 21, 2014 was used as the testing set for the overall testing in ARIMA forecast, ANN forecast, and Hybrid forecast.

Since the data contains unscheduled power interruptions, several zero values were found in the historical data causing the dataset to become inefficient and out-of-range. Data correction was then performed in order to replace outlying values that could possibly alter the behavior and final results of the electric load forecasting model and might produce poor results. Outlying values of the electric load consumption data was replaced with the respective preceding day with the same time frame of the outlying value. This process of replacing outlying values by its preceding day's respective data was supported by studies since the missing data per day has a tolerable occurrence [4, 5, 12]. Since the electric load consumption data was recorded in fifteen-minute interval, the maximum consumed electric load among the hour's four fifteen-minute recordings was chosen to represent the hour's consumption as suggested by researchers [5, 13]. By doing this the new number of observation would be 7, 176 with 7, 152 observations to be used for training the model and the last 24-hourly observation for testing the model. The number of observations to be used is just efficient for the hybrid model because a larger amount could lead to overfitting the ARIMA model while a smaller and inappropriately minimal amount could possibly lead to underfitting the ANN model [9, 14].

ARIMA modeling was conducted through model identification, model estimation, diagnostic checking, and forecasting phases. The model identification phase involves determining the order of the ARIMA model *p, d,* and *q,* where *p* represents the autoregressive terms, *d* represents the non-seasonal differences needed for stationarity, and *q* represents the lagged forecast errors in the prediction equation [10, 12, 14]. Modeling an ARMA (*p, q*) process requires stationarity in order to fit this model easily. A time series data is said to be stationary if its statistical properties do not depend on the time at which the series is observed having both mean and variance that do not change over time with the process not having trends [6, 15]. A method called differencing where differences are computed between consecutive observations must be done to achieve stationarity. Equation 1 shows the process for first order differencing where *t* is the differenced variable and y_t is the time series variable.

$$\Delta y_t = y_t - y_{t-1} \quad (1)$$

A stationarity test called the Dickey-Fuller test was then used to determine if the variable is stationary or not [14]. The Dickey-Fuller test tests the null hypothesis of whether a unit root or non-stationary is present in an autoregressive model. First regular differencing as seen in Equation (1) was applied to meet the condition for stationarity and if by doing this the electric load data is still not in a stationary condition, the second regular differencing will be applied. Once stationarity have been addressed, the next step is to identify the order *p* and *q* of the autoregressive and moving average terms respectively [14, 15]. This involves plotting the data over time and the corresponding Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). By plotting the ACF and PACF, the researchers can come up to possible ARMA models that will be used later in estimation. After the nomination of one or more appropriate models to describe the view of the time series data, parameters of the model was estimated using an estimation method used in the original ARIMA Box and Jenkins methodology. This estimation method followed a guiding principle of parsimony that is the total number of parameters of the model should be as small as possible which makes a model a good fit [14]. In addition to estimating a model for the electric load consumed data, the model with the smallest parameters is more promising to the efficient forecast due to more stable parameters obtained. Using a high quality statistical software, the coefficients of the parameters was determined to come up with the final ARIMA model that would fit the original data. These ARIMA models were further examined to check if all the parameters are significant using the assumptions of the criterions called the Akaike Information Criterion (AIC) and p-value. The AIC is an index used in a number of areas as an aid to choosing between competing models which defined as where L_m is the maximized log-likelihood and *m* is the number of parameters in the model. Among the set of suggested models that are being estimated, a model with the smallest AIC and has significant coefficients was chosen as the final model that will be validated in the diagnostic checking phase.

TABLE. I. Electric Load Data

DATE	TIME	KW_DEL	KWH_DEL	KVARH_DEL
XXX	XXX	XXX	XXX	XXX
XXX	XXX	XXX	XXX	XXX

Once a model has been identified, the result was then considered as the final model and forecasts was obtained accordingly. In a normal diagnostic check in an ARIMA Box and Jenkins methodology, fitted models was checked if it is a satisfactory one to protect against forecasting errors. This was implemented by the use of standard diagnostic checks that examined the correlogram of the residuals from the fitted model to see if the residuals exhibit white noise. A good forecasting method should yield residuals with a white noise in which residuals are uncorrelated and have zero mean. If the fitted model is a good model for the data, the residuals should satisfy these assumptions. If these assumptions are not satisfied, a more appropriate model should be fitted and the whole process will go back to the model identification step and try to develop a better model. In the case of passing the model for these tests, a final model was adopted which is used to estimate linear predictions of the electric load data. Moreover, the residual series was analyzed using the Box-Pierce Q-statistics as recommended by studies to make clear that the values are normally distributed [6, 14, 15]. When the residual series finally meets the condition of normal distribution, then the model could be used in predicting future values. After the model has been estimated and validated, this model was identified as the final model for ARIMA forecasting. Using the provided electric load dataset, the final model will obtain the corresponding linear forecast L_t at time t and the residual of the observed value where y_t is the observation at time t. The residuals dataset was then modeled by ANN.

B. ANN Modeling

Simple perceptron can only identify sets of data that are linearly separable, and when the input data to be classified are not linearly separable, learning and classification will never reach an optimum point of distinct separation [16]. In situations like this, Multilayered Perceptron (MLP) model is used in learning and classification. Thus, a MLP having input, hidden and outputs units as a type of ANN was used in formulating an ANN Model to resolve the problem of electric load forecasting [5, 7]. As a feed-forward ANN having input neurons, hidden neurons and outputs neurons, it allows signals to travel one way only, from input to output [7, 16-17]. In this study, the number of inputs of the neurons depends on the residuals of the ARIMA model, which has non-linear relationship. The hidden layer, on the other hand, serves to encode the input and map it to the output. Identifying the ANN's hidden neurons in the hidden layer does not have any standardized or theoretical approach but there are some empirically-derived rule-of-thumb approaches [5, 18]. Trial and error was then used to determine the optimum neurons in the hidden layer of the network using formulas from different researchers on how to crosscheck the efficient number of hidden neurons [4, 17-19]. Since the output layer is where the outcome of the network can be seen, the number of output neurons solely depends on the problem that the neural network wants to learn.

Backpropagation is considered as one of the original training algorithms for feedforward neural networks that uses both learning rate and momentum with learning rate as the variable of learning agility and the momentum as the variable that helps the network get out of the established local minima [5, 12]. Manhattan Propagation on the other hand uses a delta value in updating its weight values. Resilient propagation training algorithm is unique compared to the two earlier mentioned ANN training algorithms since it does not require training parameters making it much easier to model and utilize, but has shown better performance efficiency than that of Manhattan Propagation and Backpropagation [6, 17]. As shown in Table II, this study used different pairs of training algorithms with Hyperbolic Tangent and Logarithmic activation functions since these activation functions has the ability to produce results between -1 and 1 fit to the datasets that were transformed in a scale of -1 and 1. This was done conservatively for the purpose of avoiding ambiguous values.

After the architecture of the ANN model, the next phase involves training the forecasting model. During the training process, test inputs were implemented to the electric load consumed training dataset using the training algorithms. These training algorithms were used to update the network weights and adjust biases of the network until the error is less than the desired limit [9, 17, 19]. The learning parameters used in this study were the desired error, and number of iterations. The desired error was set to 0.0001 to help the network reach an optimal solution with the smallest amount of error. The standard range for the desired error should be between 0.0005 and 0.0001. The lower the desired error implies optimal the result. Since the goal of this study is to get an error rate below the international error rate of 0.0005 and to have efficient and optimal results, the desired value was set to 0.0001. In choosing the number of iteration, the number of records to be used is accounted for resulting to10, 000 as the identified number of iterations.. If the number of records is 500 then the iteration is between the range of 500+250 and 500-250 or 750 to 250. In this study, since the amount of records to be used is 7,152 for training, then the iteration to be used is within the range 10, 728 and 3576. The researchers then decided to use the value of 10000. These parameters were used in order to have an equal credibility in comparison with the forecasting models. To measure accuracy, error measure in terms of Mean Square Error (MSE) was calculated to determine the predictive capability of the models. For this error measure, resulting values will always be non-negative and values closer to zero are better [7]. Thus, smaller value from the resulting calculation would indicate a consistency of performance in the neural network.

TABLE. II. FORMULATED MLP MODELS

Model	Training Algorithm	Activation Function	Model	Training Algorithm	Activation Function
Model 1	Backpropagation	Hyperbolic Tangent	Model 4	Backpropagation	Logarithmic
Model 2	Manhattan Propagation	Hyperbolic Tangent	Model 5	Manhattan Propagation	Logarithmic
Model 3	Resilient Propagation	Hyperbolic Tangent	Model 6	Resilient Propagation	Logarithmic

C. Hybrid Model Implementation

In implementing the ARIMA model, Dickey-Fuller test was said to test stationarity followed by the identification of a candidate model to be used. The chosen model was estimated along with variations of the model for comparison. The residuals of the chosen model underwent a diagnostic check by doing a portmanteau test. The result of the final model was a linear forecast of the hybrid model, along with the residuals calculated by subtracting the actual dataset with resulting dataset which was then fed to the ANN model. The resulting linear forecasted data from the ARIMA model would be stored in a database. The ANN model was implemented in Encog, a java-based system for ease of simulation and calculation of the training and testing. The residuals from the ARIMA model would be read in Encog which training set and testing set being partitioned from the residuals. After partitioning the datasets, the next step is configuring the neural network and adding a hidden layer. After the hidden layer is added, the network would be structured and undergo a reset. After the network undergo a reset, it would then be trained along with the training set and then iterated. Once the iteration is finish, a finalized neural network is created. The last step is to calculate the error of the finalized neural network on the testing set. After the calculation, Encog would store the data in a database and exit. The result from neural network would be the nonlinear forecast of the hybrid model.

In the Hybrid Implementation phase, the linear and nonlinear forecasted dataset obtained from the previous phases stored in the database would be added manually. As shown in Equation (2), the process of adding the linear and nonlinear forecasted dataset was based on Zhang's ARIMA-ANN hybridization in which it is assumed that the given time series data is a sum of two components: linear and non-linear [4].

$$y_t = L_t + N_t \quad (2)$$

On the given time series (y_t), ARIMA is fit and the linear predictions are obtained, ($\hat{L}_t$) by Equation (3).

$$\hat{L}_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + b_1 e_{t-1} + \cdots + b_q e_{t-q} + e_t \quad (3)$$

The difference series is obtained by Equation (4) on which ANN is fit and the predictions ($\hat{N}_t$) are obtained using Equation (5).

$$n_t = y_t + \hat{L}_t \quad (4)$$

$$\hat{N}_t = f(n_t, n_{t-2}, \ldots, n_{t-A}) + v_t \quad (5)$$

The hybrid model predictions are now obtained by summing the ARIMA and ANN predictions as shown in Equation (6).

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \quad (6)$$

To summarize the hybridization process, a block diagram is shown in Fig. 1. The sums were stored in the same database and was read in the Java-based system. In the Java Interface, it would show the hybrid forecasted dataset along with the specific time and date for each data.

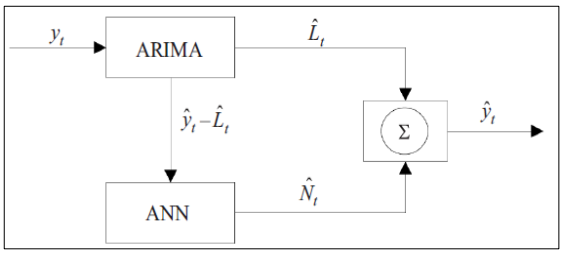


Fig. 1. A Block Diagram of the ARIMA and ANN Hybridization.

After the implementation of the ARIMA and ANN hybrid model, the results from the hybrid model were assessed using MSE. Electric load consumption data for October 21, 2014 was used in testing the hybrid model.

III. Results and Discussion

A. ARIMA Modeling Results

The results of the ARIMA modeling were divided into four stages from model identification, model estimation, diagnostic checking, and forecasting phases. The model identification phase involves determining the order of the ARIMA model *p*, *d*, and *q*. Before modeling the data, the entire 7,152 hourly observations from the raw electric load data from December 2013 to October 2014 was first plotted in a basic plot to visualize the behavior of the data. The plot as shown in Fig. 2 revealed compactness of the series with huge randominity which signifies the nonlinearity of the data. This constant trend shows that the series is in nonstationary condition and there's a need to apply differencing technique to make the series stationary [15].

Moreover, to support this claim, the test for stationarity called the Dickey-Fuller test which tested the null hypothesis of whether a unit root is present was applied which obtained a p-value of 0.81 from a lag of 24 which is equivalent to the 24 hourly values per day data. A p-value more than 0.01 signifies a non-stationary data [14]. Thus, the test signifies a nonstationary condition and the null hypothesis was not rejected. The dataset, in order to achieve stationarity of the series, was applied with the differencing technique of $y_d = y_t - y_{t-1}$ where y_t is the load at time t and y_d is the differenced load. After the technique was applied, the Dickey-Fuller test showed a result of 0.01 p-value with a lag value of 24 which signifies a stationary condition. The data was plotted as shown in Fig. 3.

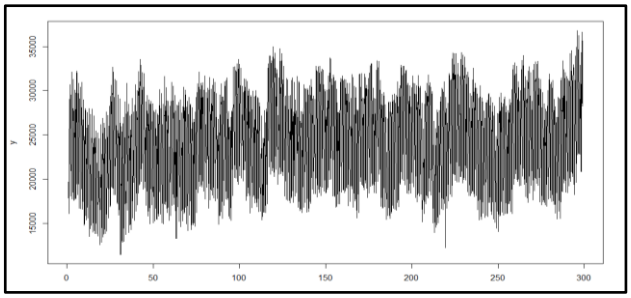


Fig. 2. Electric Load Data Plot.

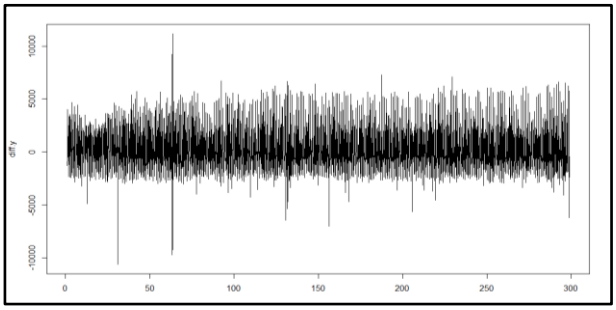


Fig. 3. Diffrenced Series.

The differenced data as clearly seen in the plot is already in a stationary condition with a constant mean and variance and so the series was ready for model identification. The ACF and PACF of the differenced series was being plotted as shown in Fig. 4 and Fig. 5, respectively which shows a tail-off behavior in the ACF and the PACF cuts off after the 9th lag. This behavior shows a stationary data and a tentative autoregressive-moving average model [15]. Thus, this behavior signifies stationarity of the series and an autoregressive (AR) and moving average (MA) model. The feasible number of AR and MA term is the lag that is zero or approximately equivalent to zero [12]. Based on the ACF, the AR term can be lags 5, 6, 7, or 8 which can be determined by looking at which lag is zero or close to zero. While based on the PACF plot, MA term can be identified the same as the AR do, so the MA term could be 2, 4, or 5. Based on the ACF and PACF results, the model that will be estimated will be ARIMA(5,1,2), ARIMA(5,1,4), ARIMA(5,1,5), ARIMA(6,1,2), ARIMA(6,1,4), ARIMA(6,1,5), ARIMA(7,1,2), ARIMA(7,1,4), ARIMA(7,1,5), ARIMA(8,1,2), ARIMA(8,1,4), and ARIMA(8,1,5) [14-15].

The purpose of ARIMA model estimation is to select a parsimonious model from the generated models which will base on the lowest total number of parameters and AIC. When comparing models fitted by maximum likelihood to the same data, the smaller the AIC implies better fit [1, 3]. After the nomination of the appropriate models based on the ACF and PACF, the AIC test was done on the generated models. The one with the smallest AIC had been chosen as the final model for forecasting. This estimation method follows a guiding principle of parsimony that is the total number of parameters of the model should be as small as possible which makes a model a good fit [14]. Table IV presents the possible ARIMA models with their corresponding criteria in terms of Root Mean Squared Error (RMSE) and the AIC. During the run, some models produced an error in fitting which outputs no AIC and RMSE which indicates that the model is not good for the feature of the data. These models are ARIMA(6,1,4), ARIMA(6,1,5), ARIMA(7,1,5). In this case, the model with the least criterion as a whole is ARIMA(8,1,2) with an RMSE of 1247.582 and AIC of 122283 as indicated in Table III. Based on the results, the effective order of the AR terms is found to be p = 8, the MA terms is equal to q = 2, and the differentiation parameter is i = 1 since the raw series is differenced in the first order. The final model of order ARIMA(8,1,2) was chosen as the model for validation in the next phase.

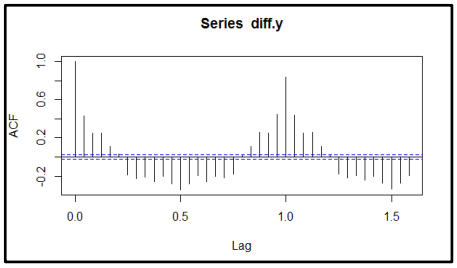


Fig. 4. Autocorrelation Function Plot.

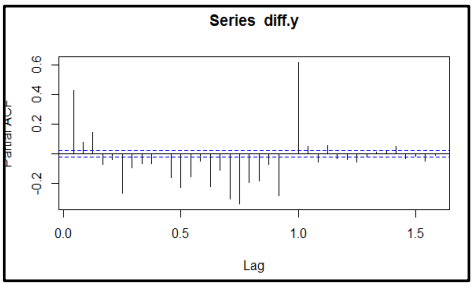


Fig. 5. Partial Autocorrelation Function Plot.

TABLE. III. POSSIBLE ARIMA MODELS

Model	RMSE	AIC	Model	RMSE	AIC
ARIMA(5,1,2)	1396.169	123881.2	ARIMA(7,1,2)	1357.848	123487.5
ARIMA(5,1,4)	1438.601	124315.4	ARIMA(7,1,4)	1259.149	122420.7
ARIMA(5,1,5)	1449.254	124419.9	ARIMA(7,1,5)	N/A	N/A
ARIMA(6,1,2)	1357.947	123486.5	ARIMA(8,1,2)	1247.582	122283
ARIMA(6,1,4)	N/A	N/A	ARIMA(8,1,4)	1259.695	122424.4
ARIMA(6,1,5)	N/A	N/A	ARIMA(8,1,5)	1249.388	122312.8

Before a forecast was generated using the ARIMA model, the model was first checked to test the adequacy of the overall model to prevent forecast errors. Residuals of the model was generated and checked by examining the correlogram to see if the residuals process a white noise. A residual in forecasting is the difference between an observed value and its forecast based on other observations [4, 15]. The residuals were plotted in a standard plot and autocorrelation plot as shown in Fig. 6 and Fig. 7, respectively. The residuals standard and ACF plot of the fitted model did not satisfy the diagnostic checking phase. The standard plot shown in Fig. 6 shows a slight pattern of data which rejects the white noise assumption. While the ACF plot in Fig. 7 shows significant spikes especially in lag 24 which correlates to the 24 hourly data and not all lags fall outside the confidence interval.

The Box-Ljung test was also applied to the residuals and the test showed a p-value of 2.2e-1. Since this study is a hybrid process of ARIMA and ANN for linear and nonlinear components respectively, the results of the diagnostic checking phase shows that it does not satisfy the claim to be a random

distribution, thus was fixed by feeding the residuals to ANN for it to model and forecast a residual data for hybridization. Based on the results, the researchers found out that there was no evidence to claim that the residuals are random. Thus, the assumption that the electric load data comprised of nonlinear component was proved and the residuals data needs to utilize ANN modeling for residuals forecasting to later be joined in ARIMA forecast. The final chosen model as a whole in ARIMA modeling phase was ARIMA(8,1,2) which has a coefficients per term shown in Table IV.

The model equation is where B is the backshift operator and a_t is the white noise. A 24-hour forecast was generated using this model which is shown in Fig. 8. The blue line is the point forecast from ARIMA(8,1,2) with its high and low boundaries. While the red line represents the actual 24-hour value of October 21, 2014. A shown in Fig. 5, the 24-hour forecast values of this model were imported in a database for later use in comparison.

While residuals of the fitted model for a total of 7,152 data points were passed on to the next process for ANN modelling and then will forecast residuals to be merged in the hybrid process. Fig. 9 shows that the raw data of 7,152 values being fed to the ARIMA generated the same number of residuals to be used then in ANN modeling and a 24-ahead forecast to be test in the hybridization process.

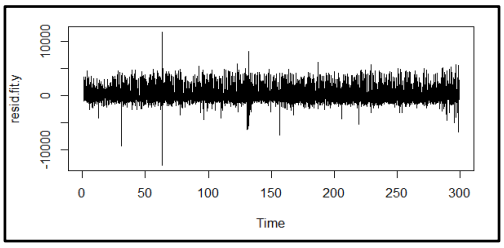


Fig. 6. Residuals Standard Plot.

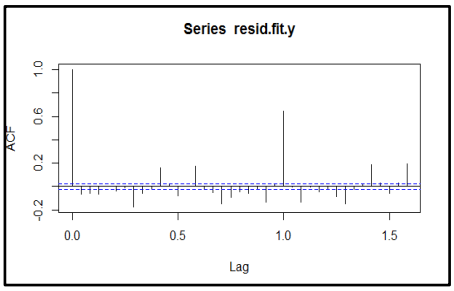


Fig. 7. Residuals Autocorrelation Plot.

TABLE. IV. ARIMA COEFFICIENTS

AR1	AR2	AR3	AR4	AR5	AR6	AR7	AR8	MA 1	MA 2
2.0165	-1.3577	0.4737	-0.4029	0.3233	-0.4254	0.5849	-0.3244	-1.9283	0.9956

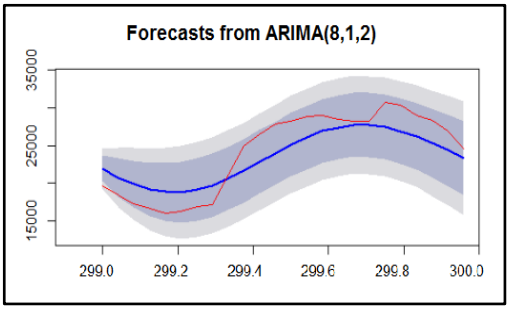


Fig. 8. Actual vs. Predicted Plot.

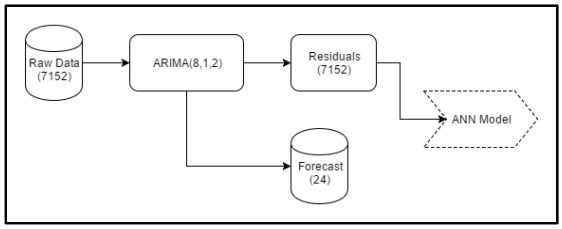


Fig. 9. ARIMA Modeling Data Flow.

B. ANN Modeling Results

The results of the ANN model were divided into four phases namely ANN data preparation, ANN model formulation, ANN model training, and residuals forecasting. The model formulated in the ARIMA phases generated residual data for the ANN to model. After the residuals were generated in the ARIMA modeling, the data were plotted in the graph as shown in Fig. 10.

The plotted residual values shows a random distribution of values with minimum value of -2680.893454503 and maximum value of 3652.1309335015 as these values are within the range of -2680.893454503 and 3562.1309335015, the boundaries for the residual transformation process. It is common in ANN modeling to undergo data transformation between a specified range e.g. -1 to 1 or 0 to 1 since this makes the training of the network efficient to yield accurate predictive results [6-8]. It was found out that the residuals generated from the ARIMA model did not satisfy this requirement resulting to a need for the data to be transformed. Shown in Table V are sample of the residuals dataset that underwent a transformation technique using Min-Max normalization which scaled down the dataset to a range of -1 to 1.

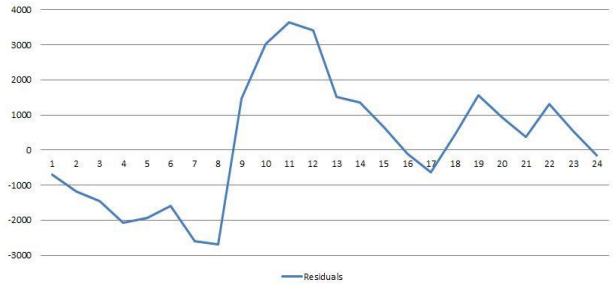


Fig. 10. Residuals Dataset.

TABLE. V. SAMPLE ACTUAL RESIDUALS AND NORMALIZED RESIDUALS

Actual	Normalized	Actual	Normalized	Actual	Normalized
-1459.63138	-0.524629072	3652.130934	1	1564.143068	0.340603245
-2596.55235	-0.973364668	648.3317892	0.051385575	522.33606	0.011595509
-2680.89345	-1	-110.760525	-0.095252 67	-154.224759	-0.807376 23

Designing the architecture of an ANN model includes the identification of the number and size of the input, hidden, and output layers. Although neural networks are used for the purpose of unsupervised learning, classification, or regression, in this study, regression was the appropriate application for the problem since residuals have time series structure [1, 7, 10]. A regressor neural network used two window sizes for the model training which are the input window size and the output window size [9, 12]. The input window used the number of input nodes while the output window size is equivalent to the number of the forecasting results being 24 hourly values in a day. The block diagram shown in Fig. 11 implies that all the models that were created were of a feedforward type of ANN where there is 1 layer containing 24 input neurons for the input layer to represent the 24 data points of the 24 hourly residuals.

If the neural network is used for the purpose of regression, then the output layer has a single node [1]. This was also supported in a study which also used 1 output neuron for the output layer [19]. Since the data used in this study is a time series data, the output layer used contains 1 neuron. A neural network with one hidden layer has the tendency to perform very well depending on the problem [5, 19]. The researchers used 1 layer for the hidden layer. In order to select the appropriate number of hidden neurons, the researchers conducted a test with the result shown in Table VI. Each identified number of hidden neurons was tested using the final ANN model and yielded error measures in terms of MSE. From the 4 hidden neurons results, 17 hidden neurons had the lowest MSE while 4 hidden neurons had the highest MSE. Thus, this study used 1 hidden layer containing 17 of hidden neurons.

The ANN model architecture composed 24 input neurons, 17 hidden neurons and 1 output neuron. This research produced 6 models and used different kinds of training algorithms, namely, Backpropagation, Manhattan Propagation, and Resilient Propagation which were Hyperbolic Tangent and Logarithmic functions. In comparing the different ANN models, the model which had the lowest error is the Model 3 which has the combination of Resilient Propagation and Hyperbolic Tangent and generated a network error of 0.003668849. As shown in Table VII, better precision and accuracy of prediction were seen in Model 3 and Model 6 which uses Resilient Propagation as training algorithm, however Model 3 outperformed Model 6 in terms of lesser MSE or prediction accuracy. On the other hand, Model 2 and Model 5 which uses Manhattan Propagation as training algorithm produced the neutral values of MSE which means it has the capability of performing predictions, however, these models were not enough to have a better prediction.

Model 1 and Model 4 which used Backpropagation as Training Algorithm produced bigger value of MSE which only means that it has less capability of achieving good prediction or low quality of performance. Moreover, results showed that models which had Resilient Propagation as the training algorithm produced the smallest amount of error which only means that better predictions can be seen in these models. Resilient Propagation exhibited more efficient performance than Manhattan Propagation or Backpropagation for supervised feedforward neural networks. This made an advantage for the Resilient Propagation training algorithm since there were no learning rates, momentum values or update constants that need to be determined. On the other hand, the Backpropagation training algorithm used two parameters in conjunction with the gradient descent which may result to a problem in the algorithm because the gradient descent algorithm should be able to seek out local minima. These local minima are points of low error but may not be a global minimum. With Manhattan Propagation training algorithm, the sign of the gradient and the magnitude is discarded. This means that it is only important if the gradient is positive, negative, or near zero. When all the propagation training algorithms were paired with the Hyperbolic Tangent activation function, it stood out and resulted lesser error than the other activation function. This is because Hyperbolic Tangent activation function works with both negative and positive numbers. It has a derivative function which can be used with propagation training making it a common choice for feedforward and simple recurrent neural networks. Using Model 3, the graph shown in Fig. 12 shows the forecasted values of the residuals for the next 24 hours.

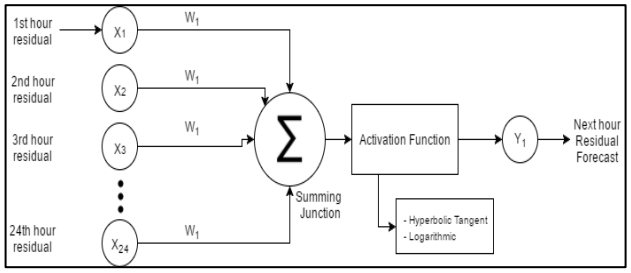


Fig. 11. Block Diagram for ANN Architecture.

TABLE. VI. HIDDEN NEURON RESULTS

Hidden Neurons Technique	Hidden Neuron	MSE
The square root of input neuron plus the output neuron	5	0.000149
The average of input and output neurons	13	0.000119
The square root of input neuron multiplied with the output neuron	4	0.000153
2/3 of the input neuron plus the output neuron	17	0.000110

TABLE. VII. ANN MODEL RESULTS

ANN Models	MSE	ANN Models	MSE	ANN Models	MSE
Model 1	0.00490	Model 3	0.00111	Model 5	0.00426
Model 2	0.00406	Model 4	0.00505	Model 6	0.001273

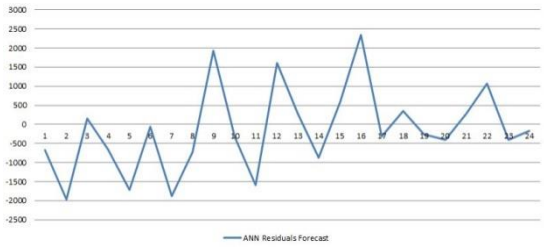


Fig. 12. ANN Residuals Forecast Plot.

C. Hybrid Model Implementation Results

In the ARIMA-ANN hybrid model implementation, the actual dataset was passed to the ARIMA(8,1,2) model. This resulted to two dataset outputs which is the ARIMA forecast or the linear forecast and another which is the ARIMA residuals. Consequently, the ANN model that processed the ARIMA residuals employed Hyperbolic Tangent activation function and Resilient Propagation training algorithm. As shown in Table VIII, this ANN model generated the nonlinear forecast which was correspondingly added to the linear forecast to generate the ARIMA-ANN forecast.

This hybrid model was based on Zhang's hybridization process using ARIMA and ANN but focuses on a different aspect or field of study [4]. As presented in Table IX, the result still shows that the hybrid technique has outputs which are more accurate than using the individual models.

Similar to a study which used ARIMA and GRNN, the hybrid forecast is much close to the actual data rather that the individual models used [12]. Another study using ARIMA and ANN on predicting traffics accidents also found that the hybrid model has a higher accuracy in prediction [10]. The results along with these studies presented that if the dataset is fully linear or fully nonlinear, then the hybrid model would not be beaten by the individual models. However, the researchers were not able to find any studies yet, disproving the prediction capability of the ARIMA-ANN hybrid model since as Zhang has assumed, there is no dataset that is fully linear and fully nonlinear.

TABLE. VIII. ADDING THE LINEAR AND NONLINEAR FORECASTS

	ARIMA Forecast		ANN Forecast		ARIMA+ANN Forecast
1	26510.853	+	-183.272	=	26327.581
2	25565.625	+	-1245.971	=	24319.654
…	…	+	…	=	…
24	28439.424	+	-463.978	=	27975.445

TABLE. IX. MAPE OF THE MODELS

Model	**Mean Absolute Percentage Error**
ARIMA Model	5.11%
ANN Model	16.90%
ARIMA+ANN Hybrid Model	4.09%

IV. CONCLUSION AND RECOMMENDATIONS

This study attempted to present a hybrid model of ARIMA and ANN in load forecasting. The methodology focused in the formulation and performance evaluation of twelve ARIMA models and six ANN models with different combination of training algorithm and activation functions. Since ARIMA (8,1,2) exhibited the best forecasting performance, its residuals were then processed by the best performing ANN model employing Hyperbolic Tangent activation function and Resilient Propagation training algorithm which generated the ANN forecast. The independent results of both ARIMA and ANN models were then processed following Zhang's hybridization technique that generated a MAPE of 4.09% which is generally lower than the internationally accepted 5% MAPE for electric load forecasting.

This study only focuses on a single hybridization technique of ARIMA and ANN models. Despite the very limited literature that attempted to develop and implement ARIMA and ANN models in forecasting, the researchers still recommend that other hybridization techniques should be explored along with the performance evaluation of these hybridization techniques. The possibility of implementing other hybrid implementation frameworks other than that of Zhang's ARIMA and ANN hybridization process can yield to a fundamental rethinking of how statistical and machine learning models can process linear and non-linear datasets for forecasting functions. The results of this study clearly suggest that a forecasting model that utilizes both statistical modeling and machine learning can perform promising results that can be used for better management of power systems.

ACKNOWLEDGMENT

The authors would like to thank Mr. Art Louie Bangcale for his technical assistance in the models evaluation. Additionally, the authors would also like to thank the support of the MSU-IIT Office of the Vice Chancellor for Research and Extension and PRISM-Premiere Research Institute in Sciences and Mathematics for their assistance in this study.

REFERENCES

[1] B. C. Narendra, and R. C. Eswara, "Performance Comparison of Four New ARIMA-ANN Prediction Models on Internet Traffic Data," Journal of Telecommunications and Information Technology, 2015.

[2] Y. Yang, J. Wu, Y. Chen, and C. Li, "A New Strategy for Short-Term Load Forecasting," Abstract and Applied Analysis, 1-9, 2013, doi:10.1155/2013/208964.

[3] N. Mohamed, M. Ahmad, Z. Ismail, and S. Suhartono, "Short Term Load Forecasting Using Double Seasonal ARIMA Model," World Applied Sciences Journal 15(2), p 57-73, 2010.

[4] P. G. Zhang, “Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model,” Neurocomputing, 50, p 159-175. 2003, doi:10.1016/s0925-2312(01)00702-0.

[5] L. C. Velasco, D. L. Polestico, G. P. Macasieb, M. B. Reyes, F. Vasquez , “Load Forecasting using Autoregressive Integrated Moving Average and Artificial Neural Network,” International Journal of Advanced Computer Science and Applications, 2018.

[6] M. Kumar and M. Thenmozhi, ”A Comparison of Different Hybrid ARIMA-Neural Network Models for Stock Index Return Forecasting and Trading Strategy,” International Journal of Financial Management, 1(1), 2012.

[7] H. Chen, C Canizares, and A. Singh, "ANN-based short-term load forecasting in electricity markets," Conference Proceedings of the IEEE Power Engineering Society Winter Meeting, 2001, doi:10.1109/ pesw.2001.916876.

[8] F. Liu, Z. Wang, J. Wu, and J. Wang, "A Hybrid Forecasting Model Based on Bivariate Division and a Backpropagation Artificial Neural Network Optimized Chaos Particle Swarm Optimization for Day-Ahead Electricity," Abstract and Applied Analysis, p 1-31, 2014, doi:10.1155/2014/249208.

[9] I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, M. C. Pereira, "Multiple linear regression and artificial neural networks based on principle components to predict ozone concentrations,"Environmental Modelling & Software, Volume 22, Issue 1, p 97-103, 2007.

[10] L. Barba, N. Rodriguez, and C. Montt, "Smoothing Strategies Combined with ARIMA and Neural Networks to Improve the Forecasting of Traffic Accidents," The Scientific World Journal, p 1-12, 2014, doi:10.1155/2014/152375.

[11] H. Cui, and X. Peng, "Short-Term City Electric Load Forecasting with Considering Temperature Effects: An Improved ARIMAX Model," Mathematical Problems in Engineering, p. 1-10, 2015, doi:10.1155/2015/589374.

[12] W. Wei, J. Jiang, H. Liang, L. Gao, B. Liang, J. Huang, and H. Chen, "Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China," PLOS ONE PLoS ONE, 11(6), 2016 doi:10.1371/journal. pone.0156768.

[13] J. E. Gardner and R. L. Lehr, “Enabling the Widespread Adoption of Wind Energy in the Western United States: the Case for Transmission, Operations and Market Reforms,” Journal of Energy & Natural Resources Law, 31(3), pp. 237-285, 2013, doi:10.1080/02646811. 2013.11435333.

[14] A. Alexandru and I. Dobre, "Modelling Unemployment Rate Using Box-Jenkins Procedure," Journal of Applied Quantitative Methods, 2008.

[15] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice", Monash University, Australia.

[16] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," .Informatica 31, University of Peloponnese, 2007.

[17] S. Param, M. Chowdhury, D. Lamp, E. Kendall, and E. Nygard, "Energy Demand Prediction Using Neural Networks" Proceedings of the Conference: 28th International Conference on Computer Applications in Industry and Engineering, 2016.

[18] D. Palchak, S. Suryanarayanan, D. Zimmerle, "An Artificial Neural Network in Short-Term Electrical Load Forecasting of a University Campus: A Case Study," Journal of Energy Resources Technology, p 135(3), 2013, doi:10.1115/1.4023741.

[19] J. Lipae, and E. Deligero, "On Forecasting Water Consumption in Davao City Using Autoregressive Integrated Moving Average (ARIMA) Models and the Multilayer Perceptron Neural Network (MLPNN) Processes, " Ateneo de Davao University, 2012.

The Use of Microwave Drying Process to the Granular Materials

Francisc Ioan Hathazi[1], Vasile Darie Șoproni[2], Mircea Nicolae Arion[3], Carmen Otilia Molnar[4]
Department of Electrical Engineering, University of Oradea, Oradea, Romania

Simina Vicaș (Coman)[5]
Department of Computers and Information Technology
University of Oradea, Oradea, Romania

Olimpia Smaranda Mintas[6]
Department of Animal Science and Agritourism
University of Oradea, Oradea, Romania

***Abstract*—The use of electro thermal technologies based on microwave energy represents an important step in the development of new innovative solutions. The numerical modelling allows to study the influence of the high frequency electromagnetic and thermal field on the dielectric materials during the drying process before achieving practical installation. So, when is developed the experimental model will be already known some of the phenomena that characterize the system, being eliminated a number of unknown issues. This paper describes experiments conducted to gather data on production parameters in order to improve the stored corn seed quality. The interpretation and dissemination of results triggers the description of "recipes" for drying corn seeds. The described method is flexible and can be applied to near any agricultural seeds in further researches.**

Keywords*—Microwave energy; microwave field; experimental models; laboratory models; drying and seed treatment

I. INTRODUCTION

The corn seeds storing represents an important question because the harvest is performed in different weather condition and their humidity is different from crop to crop. In order to preserve their quality, it is important to reduce their humidity to requested storage value through drying. There are various technologies used for the granular materials drying, such as: conventional drying, high frequency drying, vacuum drying, infrared drying, and various combinations of these technologies [1]. At present, the most complex microwave and radio frequency wave patterns that produce the inverse temperature gradient so if there is water inside the material, there is a moisture gradient, with its magnitude:

$$d = \frac{2 \cdot (u_i - u_{he})}{s} \quad (1)$$

where u_i is the moisture content inside the material, u_{he} is the moisture content corresponding to the surrounding environment in accordance with the hygroscopic balance and s is the thickness of the material.

The conventional drying rate depends on the temperature, the relative humidity and the speed of air circulation. The problem is complicated by the condensation phenomenon. In the first phase of conventional drying, free water, which exists in the cell cavities, is removed and finally the saturation point of the fibres is reached. When the humidity is 25-30% (depending on the type of material), the saturation point is reached. Therefore, the cavities contain no water, but the fibres are saturated with water. When the moisture value has been reached at equilibrium, the second drying phase moves from the surface to the inside. The drying rate at this phase is a linear function of the square root of time that indicates the drying rate controlled by the heat transfer through the material.

The distinctive features of microwave drying consist in the fact that is created a movement of moisture produced by pressure, where water will be removed by hot air [2]. Internal evaporation can generate significant pressure depending on the strength of the solid material when transporting water or vapours. This current is a different diffusion mechanism; it does not depend on the humidity concentration, and therefore cannot be described by the diffusion equation. The resulting pressure stream becomes a mechanism for transporting internal moisture and heat [3], resulting the drying.

Microwave energy affects the entire volume of the material, and due to the large differences in the dielectric loss tangent, it is absorbed mostly by the water. Because ε' and ε" decrease with temperature, the heating process is stable. Absorption decreases as the temperature rises and the penetration depth increases [4], so the water is removed from the layers of the seed core.

It is clear from the literature that hybrid drying is the most flexible and economical, although some research suggests that it is more convenient to use conventional drying, while others propose the application of the hybrid technology drying, which use microwaves and air stream [3]. There are also applications where the vacuum system has a real success [5]. An important study is about determining the maximum energy that should be applied to seeds to effectively remove free water without affecting the integrity of the grain. The research carried out has concerned on different drying and treatment methods of granular agricultural products by using microwave technology for different material samples. Starting from the results obtained in previous researches [6], the solution presented in the paper proposes to analyze the possibility of drying processes using microwave systems with continuous feeding of granular material. Therefore, the microwave drying process proposed in the paper represents an important issue

for future researches in the field, due to its advantages of improving the drying process efficiency and the quality of the seeds.

II. Materials and Methods

Accordingly to von Hippel [7], the dielectric properties of a material are defined by:

$$\varepsilon^* = (\varepsilon' - j\varepsilon'')\,\varepsilon_0 \quad (2)$$

where: ε^* - complex permittivity; ε' - dielectric constant; ε''- dielectric loss factor; ε_0 - permittivity of free space; j - complex operator.

The dielectric properties are dependent upon the moisture content, temperature of the material and on the frequency of the field. The larger the loss factor, the more easily the material absorbs the incident microwave energy [2]. Both the dielectric constant and the loss factor increase with moisture content, but decrease with frequency. The dielectric properties of corn over a wide range of moisture contents (11.1 to 53.9%, dry basis) at 2.45 GHz have been determined by Nelson [4]. The dielectric properties of a material also affect the power attenuation of the electromagnetic waves as they penetrate the loss material. Penetration depth is defined as the distance from the surface of the material at which the power drops to 1/e (36.8%) from its value at the surface and is expressed by:

$$p_d = \frac{\lambda_0 \sqrt{\varepsilon'}}{2\pi\varepsilon''} \quad (3)$$

where, p_d - penetration depth; λ_0- free space wavelength; ε' - dielectric constant; ε'' - dielectric loss factor.

The penetration depth increases with longer wavelength or, in other words, decreases with increasing frequency. The effect of temperature on the penetration depth of different foods has been studied [8]. They reported higher penetration depths at lower temperatures and frequencies, but this effect was less important at temperatures above 0°C, than for frozen conditions [9]. Uniform distribution of the thermal field in the volume of material and precise control of power consumption are essential conditions to be met to process food. The microwave applicators are designed so that to meet these objectives, microwave technology [10], being increasingly used in the food industry.

In the case of corn seeds, the irregular changing of the dielectric loss factor with temperature and moisture content makes the prediction of the penetration depth during the drying process a tedious task. In a seed bed with high moisture, water migrates from the region with high temperatures to areas with lower temperatures, leading to condensation in cold spots [8].

The proposed presented solution with in the paper (Fig. 1) implies the analysis of the complex problem of drying the corn using the hybrid method. This presume the exposure of corn seeds to microwave radiation in order to reduce the moisture content to the desired value through heating. In order to avoid the appearance of condensation to the surface of the seeds, which would lead to the uncontrolled increase of the temperature at surface, the installation is provided with an exhauster that ensures the evacuation of the water excess from the surface through the air stream. The temperature of the exposed seeds to high frequency electromagnetic field is monitored in real time by means of the temperature sensors, thus being controlled the time exposure to microwave of the processed corn seeds. The microwave energy supplied by the microwave generator is transferred to the grains upon heating thereof by means of the special design of the radiating antennas [11]. The microwave system was designed so as to allow the continuous processing of the granular products, their displacement in the microwave field being realized through a helical screw with blades which ensure also the homogenization of the thermal field. As follow the main constructive elements and their role in the microwave drying solution studied in the paper are presented.

The drive motor with it adjustable rotation speed, is coupled to the helical screw carrying the seeds in the microwave field. It is being used a frequency phase converter for infinitely variable speed with a constant mechanical drive motor torque.

The grain supply system allows both the continuous granular material supply subjected to microwave drying and prevents the microwaves leakage to outside.

The technical solution adopted for preventing leakage of microwaves to the outside consists of a dosing device with four blades which rotate inside one cylinder. By rotating blades, one helical auger dispenser transfers the seeds and three block access to the outside of the microwave field. Rotating blades dispensers are provided by a gear motor speed control that provides a steady transfer of seeds inside auger conveyor screw.

To protect against leakage of microwaves in external fixation in the housing supply system was used a copper wire braid with cross-cutting dimensions of 3×30mm, arranged for the whole fixing area.

The microwave generator is designed to emit microwave energy at a frequency of 2.45 GHz, energy which is dissipated in seeds located inside the microwave applicator with screw conveyor, accomplishing heating.

The main constructive element of the source is the radiant type microwave horn antenna, which is coupling the microwave generator (magnetron). Microwave source contains all elements for producing microwave energy and radiation, respectively microwave generator (magnetron), fan, anode voltage transformer and filament, double voltage circuit (capacitor, high-voltage diode rectifiers).

The helical screw with blade rotated by an electric motor represents the constructive solution designed to move seeds in the microwave field, so as drying is accomplished. By moving the seeds in the microwave field occurs thermal dehydration, so that after drying, they will be within imposed humidity standards.

(a) (b) (c)

Fig. 1. The Experimental Installation. (a) Air Exhauster (b) Rotating Helical Screw Blade (c) Grain Supply System.

The exhauster has the role to evacuate humidity formed in the enclosure through the air stream produced by the blower, guiding the air flow and heat/cold air source in the microwave dryer. The air is supplied by a fan and the temperature can be raised to the desired level if so by passing it through a heating coil, the temperature control being accomplished through the thermocouple.

Prolonged exposure to microwave radiation at an intensity greater than 10 W/m2 is a dangerous concern to the health of the users. To minimize radiation exposure to personnel working in the immediate vicinity of experimental plant, it's used perforated metal sheet through which the air jet enter in the facility. The leakage of radiation in the vicinity of equipment was constantly monitored with a leak microwave detector, in order to determine that the existing value does not exceeding safety standards.

Overall and constructive technical characteristics, the microwave drying system has the following specifications:

- Supply voltage 3×380 V / 50 Hz.
- Maximum power absorbed 15 kW.
- Maximum power microwave generators. 7.5 kW.
- Microwave frequency generators 2.45 GHz.

Before and after each experiment, the weight of the sample was determined with a high precision digital scale. With the measurement devices are monitored the process parameters: the microwave power input, direct power, humidity, airflow exit, adaptation of the load impedance, the temperature of the air stream, which is set so as the temperature will not exceed 55°C ± 5% in the seed bed.

The temperature in the seed bed was determined using an optic thermometer for the seed bed. Humidity and air temperature at the output gap was determined by using the analyser Lutron YK-90HT. At the end of each experiment the stored samples of dry seeds are marked and preserved in paper bags and sent in a specialized laboratory to determine the percentage of germination.

The air supply and heat sink are turned on at least 3-5 minutes before starting the experiment to bring the system to a steady state. The temperature of the air entering the system is regulated with an accuracy of ± 0.5°C with a thermocouple. The air temperature measuring thermocouple was installed at a sufficient distance from the applicator and was protected with aluminium foil in order to avoid the induction interference that may occur in high-frequency electromagnetic field.

Thermocouples were tested for correct operation electromagnetic field by comparing the temperature read when microwave power was on and off. The readings were appropriate and validated. The relative humidity of the air at the input is not stable which causes the measurement accuracy to be in the range 5-15%.

III. Numerical Modeling

The problem was initially numerically analysed, and the results obtained being used to set the optimum operating parameters for the corn seeds microwave drying system. For the numerical simulation a professional software was used, defining a Frequency - Transient type of Study. The electromagnetic field solutions are obtained by solving Maxwell's equations, in sinusoidal state. The source field on the port is supposed known as $\mathbf{E}_t$ (transversal electric wave) component, the internal and external sources are imposed null and on the rest of the superconducting walls the $\mathbf{E}_t$=0. So as for the considered mathematical model, this will presume the solve of equation:

$$\nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) - k^2 (\varepsilon_r - j\sigma/\omega\varepsilon_0)\boldsymbol{E} = 0 \quad (4)$$

where was considered: μ_r - the relative magnetic permeability; k - the wave factor; ε_r - the relative electric permittivity; σ is the electric conductivity (S/m); ω is the pulsation of sinusoidal quantities (s^{-1}).

For the considered model, the input data are the applicators work frequency 2.45GHz, the dielectric's characteristics, the cavity excitation mode and the above boundary conditions.

An important factor for the dielectric materials processed in microwave field is represented by the thermal field and its distribution in the considered sample. During numerical computation in order to analyze the electromagnetic and thermal field is considered Frequency Transient working mode.

A complicated water diffusion problem appears, by considering the inner evaporation, where a nonhomogeneous pressure field interferes due to the water vapours [2]. Also, in the drying processes, the fast occurrence of the vapours from the inside of the dielectric sample can cause its destruction. From this consideration the maximum temperature inside the seeds needs to be limited (below 65°C). From this consideration the inner evaporation can be neglected, taking into account only the one of the surface of the granular material. The surface evaporation speed depends on the difference between the surface temperature of the seeds and the ambient temperature, with respect on air pressure, to the degree of saturation of vapours, on air flow in the proximity of the charge etc.

The thermal field problem is computed for the transient regime for the heating of the supposed sample placed in the applicator. The thermal field developed in the sample is represented by the heat source associated to Joule's effect [12], [13], with its both components the heating of the material (the first term), and the heat transfer by conduction (the second term).

$$p = \rho C_p \frac{\partial T}{\partial t} - \nabla \cdot (\lambda \cdot \nabla T) \quad (5)$$

where: p – volume power density (W/m3), ρ - density (kg/m3), C_p - the specific heat (J/kg°C), T – temperature (°C), λ – thermal conductivity (W/m°C).

The imposed boundary condition for thermal field is:

$$-\lambda \frac{\partial T}{\partial n} = \alpha (T - T_0) \quad (6)$$

where α is the thermal convection coefficient (W/m^2°C) and T_0 is the temperature on the dielectric's boundary and in the air domain (°C).

In the boundary conditions for the thermal field problem we consider: as initial values $T = T_0$ ($T_0 = 22$°C that represents the temperature of the dielectric's boundary in the air domain). Thermal Insulation: $-\mathbf{n} \cdot (-\lambda \nabla T) = 0$ [11].

The space discretization of equation (4) in performed by using finite element method, using the same mesh as in the electric problem where the computation domain was meshed by using tetrahedral network elements. It was assigned an extra fine mesh for the dielectric material, and for the rest of the geometry was defined a coarse mesh.

The geometry of the cavity was made, having the dimensions: 240×345×250mm. The helical screw was constructed with 3 cylinders and below was placed a block representing the seed bed. The walls of the cavity and waveguide were defined as Aluminium, its interior being defined by air with its properties: electric conductivity σ=0 and $\mu_r=\varepsilon_r=1$. The dielectric material is corn seeds (with 26% humidity), described by ε'=3.5, ε"=0.66 and tgδ=0.19. The thermal properties of the seeds are: ρ=1300kg/m3, λ=0.17W/mK and C_p=1.5 kJ/kgK [4], [14].

As input data were used different values of the power and time: microwave power of 25, 50, 75, 100, 125, 150 and 175 W for 300 s; microwave power of 20, 40, 60, 80, 100, 120 of a processing time of 600 s; and microwave power of 20, 30, 40, 50, 60, 70 W for 900 s. For each numerical simulation was noted the maximum value of the temperature measured in the whole volume of the dielectric material. The objective of the study was finding the most appropriate value of the microwave power and processing time in order to dry the mass of seeds without affecting its structure.

Fig. 2 present the maximum value of the temperature measured in the whole volume of the dielectric material for each of the above cases of microwave power. When using the processing time of 300 s with microwave power ranging from 25 to 175 W the maximum value of the temperature measured in the whole volume of the dielectric material showed values from 44°C (for 25 W) to 135°C (when using 175 W). Analysing the values presented in Fig. 2, the most suitable value of the microwave power when using a processing time of 300 s was around 75 W – 100 W.

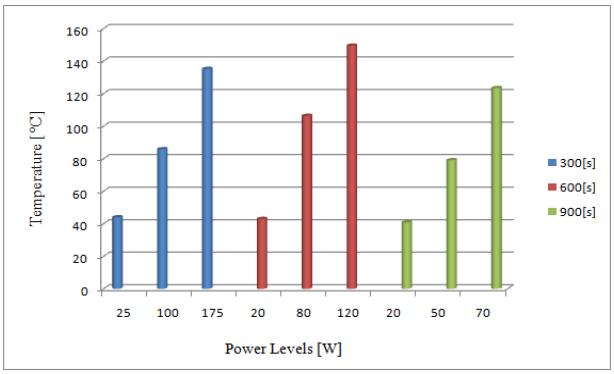


Fig. 2. The Maximum Temperatures Calculated in the Mass of the Seeds According to Microwave Power at different Values of the Processing Time.

In the case of using a processing time of 600 s next conclusions are being presented (Fig. 2):

- A microwave power of 20 and 40 W is too low for drying seeds, maximum value of the temperature measured in the whole volume of the dielectric material showed values of 42.97°C, respectively 63.16°C;
- A microwave power of 80, 100 and 120 W is too high for drying seeds, maximum value of the temperature measured in the whole volume of the dielectric material showed values of 106.32°C, 127.9°C respectively 149.48°C;
- The most suitable value of the microwave power considering keeping the quality of the seeds and decreasing humidity is 60W, the maximum value of the temperature measured in the whole volume of the dielectric material reaching 84.74°C.

When using a processing time of 900s the maximum value of the temperature measured in the whole volume of the dielectric material showed the best values for a microwave power of 50W – 79.1°C. For values of the microwave power of 60 W and 70 W the temperature was 93.86°C and 123.41°C (Fig. 2).

The numerical results obtained from the simulation of the drying process, allowed to know the electromagnetic and thermal field parameters from the microwave drying equipment. Starting from the obtained values it was sought to adapt the operation parameters of the installation, thus validating the theoretical model.

IV. Experimental Data

The purpose of these experiments was to find optimal conditions based on influence of that heating in a microwave field hot / cold air has on maize seeds with high humidity. Were analysed various parameters (temperature, humidity, power) in the drying maize seeds with high humidity using a mixed process microwave / hot air. Considering the conclusions presented in the numerical simulations the effects of microwaves on drying characteristics were studied in the following cases [13]: by using a variable power microwaves with cold/hot air and by using a constant microwave power without airflow.

Samples were performed on untreated maize corn variety Turda super (semi early variety), which is a common maize. To see and study the influence that microwaves have on seeds treated in a microwave field in the two cases above, the grains of maize were germinated and it was observed the daily evolution of the germination process.

The aim is to determine what conditions of temperature, humidity and power are favorable for germination of maize seeds treated in a microwave field. To determine the percentage of moisture removed from the seeds (STAS 10349/1-87) bed using weight of the sample before drying m_i and after drying m_u:

$$U = \frac{m_i - m_u}{m_u} \times 100[\%] \quad (7)$$

Experimental data were conducted on corn, the variety Turda Super (semi early variety) by using microwave power or mixed methods microwave power/jet of cold air. The purpose is to see what happens to the temperature and humidity seed bed where constant power use of about 1.25W/g. Measurements were performed every 30 to 30 seconds without closing down operations and they required continuous monitoring of the power generated to not reach an excessive temperature in the seed bed. During the experiments the reflected power was kept to 0.

1) For the first sample were used mixed methods - microwave power / jet of cold air, by using a weight of 127.2 kg initial corn. After drying in a microwave field for 15 minutes to obtain a final weight of 114.07 kg of maize with a moisture proper for storage, only 13.13 kg standing out of water evaporated from the seed bed. Moisture removed from the seed bed is only U = 11.51%. It was obtained a small amount of water released since it was used jet of cold air and thus less water was removed. At first, was applied a microwave power of 2.5 W/g than decreasing it to 1 W/g and furthermore, from minute 3 to a rather small power 0.50W/g-0.10W/g (see Fig. 3). Germination for this sample has a value of 75% whereas for the control sample was 92%.

2) For the second sample it was used the same amount of corn seeds, as described above, but it was applied microwave power with hot air in the first 4 minutes, then cool shot was used until the end of the test period (see Fig. 4).

The initial mass was 127.2 kg of wet corn after drying for obtaining 112.94 g dried corn seeds in 15 minutes. There is a considerable difference in weight, 14.26 kg of water evaporated, humidity eliminated from the seed bed is U = 12.62%. It was obtained this small amount of moisture removed from the seed bed because there were used low levels of power: starting with 1.3W/g - 1.0W/ g decreasing to 0.1W/g. In this case, was obtained an 80% germination rate. The first two samples stand a good plant development.

3) A new experiment was performed on the same amount of corn seeds, but only using microwave power. Note that after drying for 15 minutes, there was a difference between the initial and final weight of only 7.16 kg of water evaporated from the seed bed. Seeing a film of water on the surface of the seed bed, a phenomenon explained by the fact that there was used only the microwave power without airflow and thus condense occurred. Moisture removed from the seed bed is U = 5.96%. Variation in temperature is very high in the first minute of 83.6°C – 94°C – applied due to high power 2.0 W/g (see Fig. 5). Note that at a power of 0.1 W/g appears quite high humidity fluctuation in corn seed bed. Germination rate for this sample is G = 25%, seeds that germinated are developing normally.

4) For this samples it was used only the microwave power without air flow. Because it was used only the microwave power, without airflow to eliminate vapours from bed, was noticed a film of water on the surface of the seed bed. During the test period, the output air temperature and the temperature in the seed bed have steady growth. It was used a constant power of 0.20W/g/15 minutes, with an initial weight of 105.42 kg of wet corn and a final weight of 102.87 kg of dry corn. Was noticed little difference between the initial and final mass, 2.55 kg of water evaporated from the seeds bed, humidity eliminated from the seeds bed is U=2.47%. Notice that, when using a power of 0.20W/g in the seed bed, the temperature increases from 21°C to 63°C value (see Fig. 6). Because these seeds have a higher moisture percentage, more energy is absorbed. Germination rate in this case is G = 70%, a good development of seeds. From the experimental data using only the microwave power, this sample shows the highest rate of germination. This can be explained by the fact that when using a low power, constant 0.2W / g / 15 minutes in the seed bed the temperature did not exceed the amount of 63°C. Germination rate of the 5 samples of corn is shown in Fig. 7.

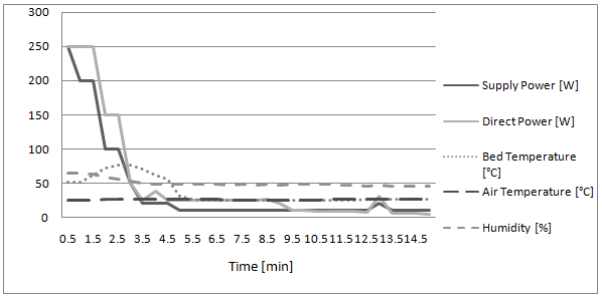


Fig. 3. The Parameter Values using a Variable Microwave Power and Cold Air Jet (2.5÷1.0÷0.10W/g, U=11.51%, G=75%).

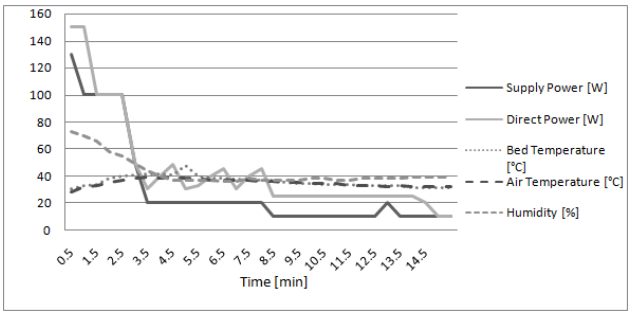


Fig. 4. The Parameter Values using a Variable Microwave Power and Cold/Hot Air Jet (1.3÷0.2÷0.1W/g, U=12.62%, G=80%).

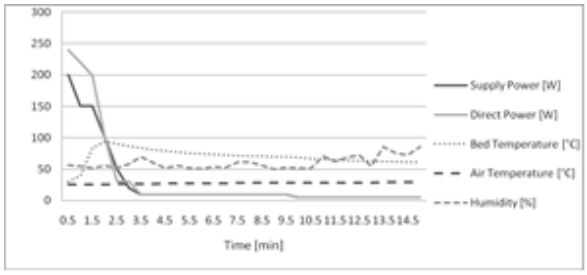


Fig. 5. The Parameter Values using a Variable Microwave Power without Air Jet (2.5÷0.1÷0.05W/g, U=5.96%, G=25%).

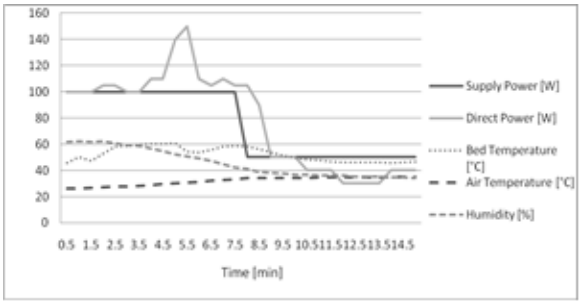


Fig. 6. The Parameter Values using a Constant Microwave Power without Air Jet (0.2W/g, U=2.47%, G=70%).

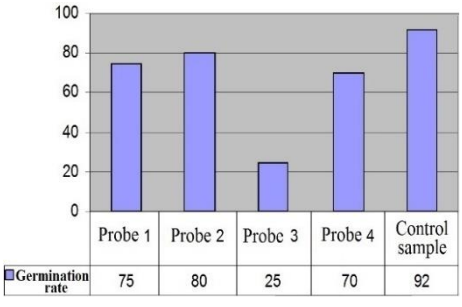


Fig. 7. Corn Seeds Germination after 14 Days.

V. Conclusions

The experimental results presented the process of drying corn seeds using a variable and constant mode microwave power generation. High levels of power should be used as long as the drying speed is high enough. Then power should be applied by lower values for the remainder of drying. By changing the power, density value of 0.75W/g - 0.50 W/g to 0.25 W/g increases drying time and thus yields a higher energy required to operate in an alternating mode. This study focused on fundamental aspects of research on the microwave drying of maize seed. Based on the results of studies were reached the following conclusions:

1) The applicator has been used successfully to couple microwave energy to the seed bed, the following parameters were monitored during drying: the microwave power, the temperature in the seed bed, the temperature and humidity of the air leaving the outlet of the laboratory installation.

2) On drying of corn seeds for a microwave power density generated by 2.5W/g – 0.1W/g was obtained a more than 92% germination.

3) Drying time can be substantially reduced by using high values of microwave power absorbed by seeds, but there are detrimental effects on the product quality. Because large temperature and humidity variations adversely affect quality of seeds treated in a microwave field, it is important to continue monitoring the generated power to avoid high temperatures in the seed bed.

Analyzing the obtained results, both by numerical modeling and especially by experimental measurements we can say that the use of the microwave field for drying the corn seeds is useful for reduction of drying time and increasing the quality of the seeds as long as that the maximum permissible values of the drying temperature are not exceeded.

Acknowledgment

This work was co-financed from the European Regional Development Fund through Sectoral Operational Program "Increase of the Economic Competitiveness" – POS – CCE 2007 – 2013, project number POS – CCE 1843 / 48800, "Increasing the capacity of research - development of the interdisciplinary laboratories for the technologies in electrical engineering", developed by University of Oradea.

References

[1] E.T. Thostenson and T.W. Chou, "Microwave processing: fundamentals and applications", Compos. Pt. A-Appl. Sci. Manuf., **30**, pp. 1055-1071, 1999.

[2] T. Leuca, A. Palade, I. Hănțilă, L. Bandici, "The use of hybrid finite element – boundary element method for all the analysis of the operating parameters of a radio-frequency dryin instalation", Revue roumaine des sciences techniques, Série Électrotechnique et Énergétique, **56**, 4, pp. 367-376, 2011.

[3] Hassini, L., Peczalski, R., & Gelet, J.-L., "Combined Convective and Microwave Drying of Agglomerated Sand: Internal Transfer Modeling with the Gas Pressure Effect", Drying Technology,**31**,8,p.898-904, 2013.

[4] S.O. Nelson, "RF and microwave dielectric properties of shelled, yellow-dent field corn", Trans. ASAE, **22**, 6, pp. 1451-1457, 1979.

[5] Wojdyło, A., Figiel, A., Lech, K., Nowicka, P., & Oszmiański, J., "Effect of Convective and Vacuum–Microwave Drying on the Bioactive Compounds, Color, and Antioxidant Capacity of Sour Cherries", Food and Bioprocess Technology, **7**, 3, pp. 829-841, 2014.

[6] F.I. Hathazi, V.D. Soproni, M.N. Arion, C.O. Molnar, "About the use of the high frequency electromagnetic field in order to improve the quality of the agricultural seeds", International Symposium on Fundamentals of Electrical Engineering, Bucharest, pp. 1-4, 2014.

[7] A.R. von Hippel, "Dielectric Properties and Waves", John Wiley Ed., New York, 1954.

[8] Chandrasekaran, S., Ramanathan, S., & Basak, T., "Microwave material processing-a review", AIChE Journal, **58**, 2, pp. 330-363, 2012.

[9] M.S. Venkatesh, G.S.V. Raghavan, "An overview of microwave processing and dielectric properties of agri-food materials", Biosystems Engineering, **88,** 1, pp.1-18, 2004.

[10] T. Ohlsson, N. Bengtsson, "Microwave technology and foods", Advances in Food and Nutrition Research, **34**, pp. 65-140 (2001).

[11] Cusacani Guerrero, Julio Agapito, Roman-Gonzalez Avid, "Antennas of Circular Waveguides", (IJACSA) International Journal of Advanced Computer Science and Applications, **10**, 8, 2019.

[12] S. Coman, T. Leuca, O. Coman, "Optimization of the dielectric position inside the microwave applicator", Revue Roumaine des Sciences Techniques Serie Electrotechnique et Energetique. **58**,4.pp.357-366, 2013.

[13] Hemis, M., Choudhary, R., & Watson, D. G., "A coupled mathematical model for simultaneous microwave and convective drying of wheat seeds", Biosystems Engineering, **112**, 3, pp. 202-209, 2012.

[14] Holtz, E., Ahrné, L., Rittenauer, M., & Rasmuson, A, "Influence of dielectric and sorption properties on drying behaviour and energy efficiency during microwave convective drying of selected food and non-food inorganic materials", Journal of Food Engineering, **97**, 2, pp. 144-153, 2010.

Intuition, Accuracy, and Immersiveness Analysis of 3D Visualization Methods for Haptic Virtual Reality

Taehoon Kim[1], Hyeonseok Song[3], Mee Young Sung[4]
Department of Computer Science and Engineering
Incheon National University, Incheon, Republic of Korea

Chanwoo Kim[2]
JLK Inspection Seoul
Republic of Korea

***Abstract*—The purpose of this study is to analyze the usefulness of 3D immersive stereoscopic virtual reality technology in applications that provide tactile sensations. Diverse experiments show that the haptic 3D visualization method presented in a 3D stereoscopic space using headsets is more intuitive, accurate and immersive than the haptic 3D visualization method displayed in 3D flat displays. In particular, the intuitiveness has been significantly improved in a stereoscopic 3D visualization rather than flat 3D visualization. In spite of the general superiority of the stereoscopic 3D visualization, it is mentionable that the precise operation performance has not been greatly improved in the elaborate object movements. It means that users can recognize and manipulate the position of objects more quickly in 3D stereoscopic immersive VR environments, however, the precise operation does not benefit greatly in the 3D stereoscopic visualization. Note that the degree of game immersion is remarkably augmented in the case of using 3D stereoscopic visualization.**

***Keywords*—3D; stereoscopic; haptic; intuition; immersiveness; recognize; simulation; virtual reality; visualization**

I. INTRODUCTION

The development of virtual reality (VR) and stereographic visualization technology has made a great impact in the field of education. Furthermore, the advancement of head-mounted display (HMD) technology has led to the creation of many immersive simulations using VR HMDs for education and training purposes. It is noted that immersive technology with VR HMDs is advantageous for acquiring cognitive skills related to remembering and understanding spatial and visual information. It is also useful for psychomotor skills related to the movement of one's head, such as visual scanning or observational skills [1]. Haptic interfaces can enhance the immersion of VR simulations by enabling users to "touch" and "feel" objects in virtual environments. Virtual haptic simulations can be implemented with both the two-dimensional (2D) graphics and the three-dimensional (3D) graphics. 3D graphics are divided into 2D flat visualization and 3D visualization.

In this study, the authors only consider methods to visualize virtual haptic simulations implemented in 3D graphics. 3D visualization generally presents stereoscopic images using binocular parallax. There are stereoscopic display and volumetric displays, which are classified according to whether wearing glasses or not. The currently commercialized binoculars are a way to assess the stereoscopic effect by using shutter glasses or polarized glasses. The images taken with two lenses are combined into one image from the brain through two eyeballs again. The resulting binocular disparity can be reconfigured in the brain to give a sense of depth.

Another common 3D graphics visualization method is to use a stereoscopic head-mounted display like Oculus Rift or HTC Vive. This method allows stereoscopic images to be synthesized in the brain by inputting binocular images directly into two eyes. Combined with the head tracking equipment, it has the advantage of providing a constant immersion feeling even when the viewpoint is changed by turning the head. Various previous studies and methods for visualizing 3D stereoscopic images have been suggested. Table I summarizes the diverse visualization methods. This study adopts a 3D visualization method using a head-mounted display due to the easiness of implementation and its efficiency.

TABLE. I. VARIOUS 3D STEROSCOPIC VISUALIZATION METHODS

3D Visualization Method	Visualization Equipment	Summary
Anaglyph Stereoscopy	Color glasses	Two differently filtered colored images are viewed through the "color-coded" "anaglyph glasses", one for each eye.
Film Patterned Retarder	Polarized 3D glasses	The optical element controlling the polarization direction of the light called the Patterned Retarder (PR) is attached to the display front side and the light of the different polarization characteristics comes out from even and odd number line of the horizontal direction.
Shutter Glasses Method	3D active shutter glasses	Two screens separating are shown in order on the display. At the same time, when a screen for the right appears in synchronized shutter glasses, the left closes and the right closes when a screen for the left appears.
Head-Mounted Display	VR Headsets	A 3D rendered virtual environment is shown through one or two small displays with lenses and translucent mirrors embedded in headsets, visors or helmets.
Parallax Barrier	No device	The vertical shield is properly placed in front of the screen, and the image using this shield is only reached by the left eye and the other by the right eye.
Volumetric Display	No device	The light is shot in the real space and the flow of the cubic light is made.

This work was supported by Incheon National University Research Grant in 2015 (No. 2015-1774).

Recently, many studies have been conducted applying 3D stereoscopic visualization methods to various fields. The 3D visualization method using a VR headset is expected to produce superior results in user experience than the 3D visualization method using a flat monitor which is commonly used. If the expected results are actually proven by experiments, it will be a good basis for applying 3D visualization methods in wider and diverse fields.

This study is to analyze through experiments the efficiency of 3D haptic VR technology that provides virtual simulations with 3D stereoscopic images as well as tactile sensations. In order to conduct experiments, the authors developed haptic VR simulations that perform various experiments. These simulations allow for experimenting the object recognition, accuracy of operation, and immersiveness in virtual environments. the authors intend to analyze the effectiveness of 3D visualization methods through three experiments using haptics and virtual reality technology.

The researchers also wish to demonstrate the improvements of the accuracy and the speed of virtual operations by means of various experiments using stereoscopic 3D visualization (3D graphics to visualize on the VR stereoscopic headset) rather than the flat 3D visualization (3D graphics to visualize on a 2D flat display) for virtual simulations.

The rest of this paper proceeds with a short survey of related works. Then, a brief description of virtual simulations used for experiments and some technical considerations are presented. Next, the methods of experiments are briefly explained. Then the results of the experiments are elaborated. Lastly, a comprehensive analysis of our experiments is described. followed on by our conclusion and future work.

II. Related Works

Various core technologies for effective haptic virtual reality simulations are being studied in order to provide training systems for practicing before actual operations. Among many studies, there is a case where a simulation developed to perform elaborate work using a haptic device. This study developed an educational program for the telemanipulation of carbon nanotubes using haptic feedback and a 3D display [2]. Another study proposes a method to evaluate usability and acceptance in the virtual learning environment [3]. There is also a study that reports the development of an interactive immersion environment for educating nanotechnology [4]. And there is a study that the effect of the VR system on academic achievement was evaluated for engineering students. The results of this study showed that academic achievement was significantly higher when the VR system was used [5]. As a result of utilizing a cost-effective smartphone-based VR system for classroom teaching, it is suggested to use VR for leading immersion and realistic learning [6]. An example of an evaluation study is a review of the author's work of 20 years ago that confirmed that achieving experience through VR has a positive impact on worker performance [7].

In the field of medicine, diverse haptic virtual reality surgery simulations are being studied in order to provide medical students and surgeons practicing before the actual surgery. One of the remarkable early studies for surgery training simulators using haptic devices and virtual reality technology can be found in 2007. It is a simulation-based training in minimally invasive surgery (MIS) which allows the trainee touch, feel, and manipulate virtual tissues and organs while viewing images of tool-tissue interactions on a monitor as in real laparoscopic procedures [8]. Also, a dental anesthesia training simulation is developed based on anatomical data [9]. In this study, haptic technology is used for dental treatment technology education. Another example of using haptic equipment and the VR environment can be found in the case of the construction of an online database from the suturing procedures of surgeons [10]. In addition, relevant to cognitive science, a study on the HVDT (haptic visual discriminant test) [11] using haptic virtual reality technology can be noted [12]. This paper presents a possible integration of vision and haptic perception of physically or mentally disabled children.

Another study analyzes the relationship between the 3D position of a three-dimensionally rendered object and the point that touches on the surface [13]. This study demonstrates that users can recognize accurately when touching objects presented in 3D stereoscopic environments then 2D flat environments. There is also a study that addressed the problems that arise when 3D touch applied in a stereoscopic display environment like the flat-screen in a 3D movie theater. This study evaluates the effect of visual conflicts such as vergence/accommodation mismatches and double vision for mid-air 3D selection performance [14]. We can also note a case study for the exploration and information retrieval of scientific articles [15].

Pre-operative training through surgical simulation can have a great effect. However, even if presenting and dealing with 3D objects in a virtual environment represented by a flat monitor, the position of the object on the flat monitor can be much different from what actually seen. The reasons are that there can be distortion on the flat-screen or a lack of perspective. Because of this problem, there is a limit to recognizing the position of objects through a simple 2D plane monitor. In order to practice the operation of the correct movement, intuitive, and precise awareness of the operation position is needed.

III. Virtual Simulations

This study measures how quickly and realistic a user feels if stereoscopic VR is applied to a virtual reality application. In addition, it demonstrates that the factors such as user's immersion, presence, and fun vary depending on the visualization method of 3D VR games. This section describes the implementation issues of the experimental environment and simulations.

A. Implementation Environment

The virtual simulation developed in this study is implemented using a 3D VR head-mounted headset such as Oculus Rift or HTC Vive for 3D visualization. For tactile senses, haptic devices such as PHANToM OMNI[1], PHANToM Desktop[2] , or Novint Falcon[3] are used.

[1] The Touch™ Haptic Device (2019). Retrieved from https://www.3dsystems.com/haptics-devices/touch

[2] The Phantom Premium Haptic Devices (2019). Retrieved from https://www.3dsystems.com/haptics-devices/3d-systems-phantom-premium

Basically, this experimental system is implemented C++ language software development environment, in addition to open software tools, such as UNITY Engine[4], CHAI3D[5], OpenGL[6], and OpenHaptics Toolkit[7]. UNITY Engine is a cross-platform game engine for producing content such as 3D video games, architectural visualization, and real-time 3D animation, and it provides functions that are easy to produce for Oculus Rift device programs. CHAI3D is an open-source framework of C ++ library for real-time simulation of computer touch, visualization, and commercialization. CHAI3D is simple in structure, so it can be easily used by beginners with haptic equipment. CHAI3D has the advantage of supporting the same interface regardless of the haptic device. Graphics are based on OpenGL. OpenHaptics Toolkit is also an open-source framework for real-time simulations such as CHAI3D. However, OpenHaptics Toolkit has the advantage of being easy to apply to a 3D engine because it has separate functions based on OpenGL or DirectX.

B. Simulation Implementation

There are two perspectives in 3D VR: the first-person perspective and the third-person perspective. In fact, VR in the first-person perspective with haptic feedback increases immersion and intuition than VR in third-person perspective. Meanwhile, in the experiment using HMD, there is a work that shows the result of increasing the efficiency of operation by improving accuracy and work speed [16].

The following three experiments are designed for examining whether the flat visualization or the stereoscopic visualization appear more efficient, accurate, and intuitive when performing them using haptic devices and VR headsets:

1) Intuitive Object Movement
2) Precise Manipulation
3) Immersiveness for Horror Games

For the first two experiments, the researchers developed two virtual simulations in the first-person perspective with haptic feedback. The third experiment is undertaken using a commercial 3D VR game Outlast 2 (Outlast by Red Barrels). Some considerations such as free perspective and collision detection should be resolved for implementing simulations using 3D haptic VR technology.

1) Free perspective: In order to implement simulations for experiments, it is necessary to link a haptic camera with a real camera. Existing OpenHaptics library has limitations in setting the point of view of the haptic camera. However, by using VR, simulations can be conducted from a free perspective. The authors solved the point limit of the haptic camera by adding an attribute value of the VR camera. The UNITY game engine is linked to the CHAI3D library to efficiently manipulate haptic devices and the Oculus Rift DK2.

2) Collision detection: A series of object moving processes need to place objects to the target point. However, to prevent the progress of abnormal experiments due to the phenomenon of unrealistic penetration of rigid objects during the movement process, all of the objects must be set up to be a "collision object". In addition, the application of our "haptic collision detection method using subdivision surface and sphere clustering" will provide more precise haptic rendering [17], [18]. This way allows you to touch objects with a haptic device and enable collision interactions between objects, thereby preventing impossible movements in reality.

IV. Experimental Design

This section explains the procedure of three experiments. Each experiment has different purposes. The purpose of the first experiment is to examine the degree of intuitiveness for object recognition in a stereoscopic 3D environment. On the second experiment, the accuracy and efficiency of precise operation is the main measurement of the experiment. The last experiment is about the relationship between immersiveness and stereoscopic vision.

A. Intuitive Object Movement

Define intuitive recognition of objects can influence the operational performance in 3D VR environments. The first experiment places various objects into the corresponding containers of their own shape. In this experiment, examinees move a cube to the open box, a soccer ball to the trash can, and a block of Jenga from the tower of Jenga blocks. Those objects are paired together for moving operations. The experiment compares the time until the first haptic pointer touches the object and the finish time when all objects are located at the given position. Fig. 1 and Fig. 2 show the simulation and a screenshot for the first experiment, respectively.

B. Precise Manipulation

Surgery training, such as vascular sutures, requires a high degree of precision using the appropriate medical instruments. The second experiment is to interconnect two blood vessels for vascular suture experience. The examinees move blood vessel objects on the right using the haptic device. The left blood vessel cannot be moved, and it is for connecting the right part of the blood vessel. Only the right blood vessel can be operated to be connected. This experiment compares the time until the haptic pointer touches the right object first and the final time until the right blood vessel reaches the appropriate position of the left blood vessel. Fig. 3 and Fig. 4 presents the simulation and a screen capture of simulation for the second experiment.

C. Immersiveness for Horror Games

According to a study on the immersion in digital games, immersion is a confluence of different psychological faculties such as attention, planning, and perception that, when unified in a game, lead to a focused state of mind [18]. However, the authors intend to quantitatively measure the immersiveness. In clinical studies, heart rate variability (HRV) is used as one of

[3] The Novint Falcon Haptic Device (2019). Retrieved from https://en.wikipedia.org/wiki/Novint_Technologies
[4] Unity for All (2019). Retrieved from https://unity3d.com/
[5] chai3D (2019). Retrieved from http://www.chai3d.org/
[6] Kessenich, J., Baldwin, D., and Rost, R. 2004. The OpenGL Shading Language. Retrieved from http://www.opengl.org/documentation/oglsl.html
[7] Haptic Device Drivers (2019). Retrieved from https://3dssupport.microsoftcrmportals.com/knowledgebase/article/KA-01460/en-us

the evaluation methods for autonomic nervous system change [19].

The third experiment observes the differences in human emotional responses when playing a horror game equipped with or without a VR headset. The game used in this experiment is Outlast 2, and it is tested twice for each examinee. The first is to play without a VR headset and the second is to play with a VR headset. It takes about 15 minutes after the start of the game including the introductory tutorial. In this experiment, the fear felt by users is measured by the number of blood pulses. the authors capture the change of the users' heart rate to evaluate their emotional immersiveness. The minimum, maximum, and average pulse rates are compared between the two trials.

Fig. 1. Performing Intuitive Object Movement Simulation.

Fig. 2. Screenshot of the Simulation of Intuitive Object Movement.

Fig. 3. Performing Precise Manipulation Simulation.

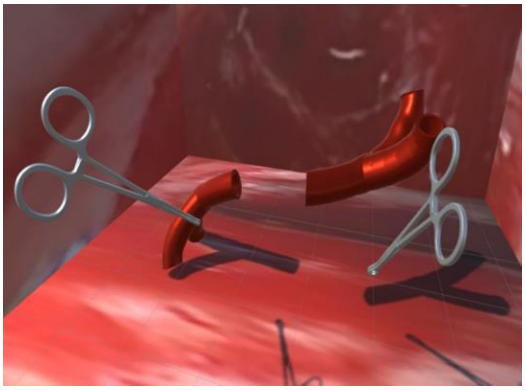


Fig. 4. Screenshot of the Simulation of Precise Manipulation.

V. Experimental Results

The examinees are divided into two experimental groups; Novice and Expert. This grouping is based on their intimacy with 3D environments. Before the actual experiment, the examinees are given a preliminary explanation of the process of experiments. After the experiments are pre-practiced, the actual experiments are repeated three times for each examinee. Then the average value was measured to reduce the effect of the operational errors.

A. Intuition for Object Movement

The measurement data for the first experiment are as follows:

- The time for positioning the haptic pointer to the target object to move.
- The time for positioning the target object to be moved to the exact target point after detection of the starting point.

This experiment is to analyze the operational efficiency by measuring the duration for intuitively recognizing an object, and then moving an object to the target point. The chart in Fig. 5 calculates the average value of the flat 3D and the stereoscopic 3D experimental results. The starting point search time shows that 23.27% (2.76 seconds) in the Novice group, 25.95% (1.98 seconds) in the Expert group, and the average search time in stereoscopic 3D is reduced by 24.33% (2.37 seconds) compared to the flat 3D.

In object movement time, the results of stereoscopic 3D experiments were reduced in all experimental groups. The manipulation time reduction rate was about 13.83% (0.65 seconds) in the Novice group, 19.31% (0.62 seconds) in the Expert group, and 16.16% (0.64 seconds) on average. Note that the average reduction rate of the Novice group is 14.17% (2.32 seconds), while the average reduction rate of the Expert group was 28.02% (3.61 seconds), much higher than the average reduction for the Novice group. In addition, the total experimental time shows an average reduction rate of 20.24% (2.96 seconds).

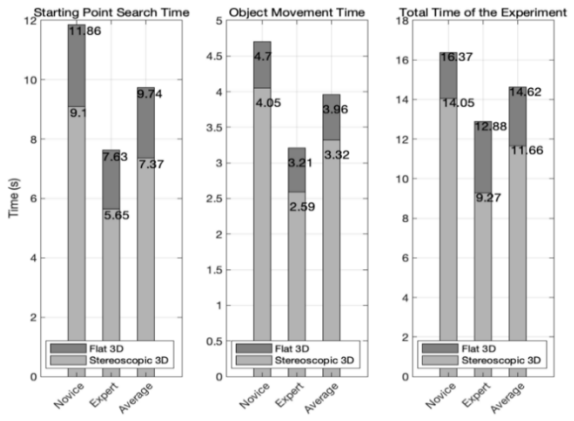


Fig. 5. The Starting Point Search Time, Object Movement Time, and Total Time of Intuitive Object Movement Experiment for Novices and Experts based on their Intimacy with 3D Environments.

The results of the first experiment above show that starting point search time is clearly better in the stereoscopic 3D than the flat 3D. It shows that users can recognize the position of the object more quickly in the stereoscopic 3D VR visualization. However, the time duration of reaching the target point after recognizing the position of the object does not differ greatly. The reason is that object manipulation time differs according to the examinees' ability to search rather than the dimension of visualization.

In this first experiment, the Expert group shows a higher reduction rate than the Novice group. It means that the more familiar the examinee is with the stereoscopic 3D environment, the more intuitive the perception can be.

B. Precise Operation of Blood Vessels

The followings are measurement data of the second experiment:

- The time for positioning the haptic pointer to the right blood vessel to move.
- The time for placing the right blood vessel to reach the junction of the left blood vessel.

Experiments measured the accuracy of the simulations to be required for surgery. In addition, the operational efficiency is also examined through the experiment of moving the object to the target point after recognizing the object.

In the second experiment, the starting point search time is faster in a stereoscopic 3D environment on average. That is, the starting point search time has decreased by 19.17% (1.98 seconds) in the Novice group, 19.97% (1.52 seconds) in the Expert group, and 19.6% (1.76 seconds) on average. In addition, the object movement time showed a reduced rate of 5.87% (0.24 seconds) in the Novice, 12.46% (0.35 seconds) in the Expert group, and 8.7% (0.295 seconds) on average. The average reduction rate has appeared at 18.02 % (1.88 seconds) in the Expert group and 15.33% (2.21 seconds) in the Novice group. The total duration of the operation in the stereoscopic 3D visualization showed an average reduction rate of 16.57% (2.06 seconds) compared to the flat 3D visualization as presented in Fig. 6.

The first indicator shows improvement of intuition in 3D stereoscopic visualization compared to that of flat 3D visualization. From the second indicator, the authors can understand that the time to recognize an object and manipulate the haptic device for moving it in a stereoscopic 3D environment is higher than that of a flat 3D environment. Nevertheless, the absolute value of the time shows a small reduction of less than just one second on average. It means that the stereoscopic 3D visualization does not give a significant improvement in accuracy compared to the flat 3D visualization. This can lead us to conclude that the 3D stereoscopic visualization has a higher improvement in the intuition of recognizing the object than the flat 3D visualization. However, in the process of performing the precise work, stereoscopic 3D visualization does not show significant improvement than the flat 3D visualization.

In the second experiment, the Expert group had a slightly higher value of the reduction rate of time than the Novice group. The actual value of the absolute time decrement was larger in the Novice experiment group. This result shows that there is no correlation between the degree of intimacy and the degree of improvement of accuracy in the 3D environment when the precise work is performed.

C. Immersiveness for Horror Games

In the previous two experiments, data are recorded in units of time to analyze the accuracy of object recognition. The third experiment checks the following data to analyze immersion:

- The average heart rate of examinees at the simulation being performed
- The heart rate in three emotionally sensitive sections of the most fearful parts

The examinees are divided into the following two groups:

- Wearing a VR headset initially
- Wearing a VR headset subsequently

In the third experiment, the heart rates of the examinees are used as the major indicator, unlike the previous two experiments, and recorded in beats per minute (bpm). The reason behind dividing the experiment groups in terms of wearing a VR headset initially or subsequently in the third experiment is to normalize the variability of the heart rate that may possibly arise from being immersed in a virtual environment with or without the prior knowledge of the content.

In Fig. 7, the authors can recognize that the overall heart rate of the group that wore a VR headset initially is generally low in a flat 3D environment compared to the stereoscopic 3D environment. The Novice group showed a reduced rate of 5.34% (4.44 bpm) and the Expert group showed a 9.08% (6.74 bpm) increased. The heart rate was increased by 11.49% (8.75 bpm) on average. For emotionally sensitive sections, the heart rate of the Novice group decreased by 7.55% (6.5 bpm) and that of the Expert group also increased by 13.05% (9.93 bpm). On average, it was increased by 13.3% (10.3 bpm) in emotionally sensitive sections.

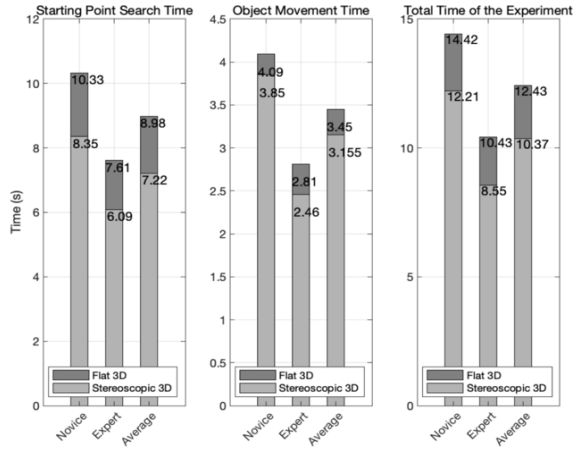


Fig. 6. The Starting Point Search Time, Object Movement Time, and Total Time of Precise Manipulation Experiment for Novices and Experts based on their Intimacy with 3D Environments.

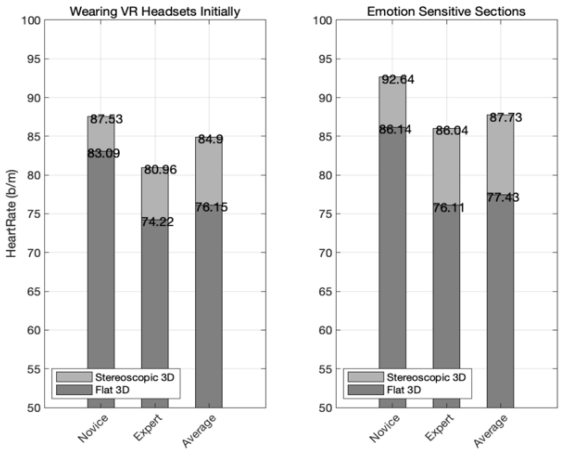


Fig. 7. Results for Heart Rates of the Group Wearing a VR Headset Initially.

As summarized in Fig. 8, the experiment in the experimental group that used a VR headset subsequently, the heart rate of the Novice group increased by 7.95% (6.48 bpm) and that of the Expert group increased by 4.85% (3.6 bpm). The result also showed a 6.74% (5.23 bpm) cardiac rate reduction on average compared to the stereoscopic 3D environment in the flat 3D environment. In addition, the heart rate in the emotion sensitive sections also showed a reduction of 7.74% (6.58 bpm) in the Novice group and a reduction of 4.83% (3.75 bpm) in the Expert group. The heart rate increased by 7.4% (5.9 bpm) on average. Overall, regardless order of wearing headsets, the average heart rate of wearing a VR headset was 83.28bpm and the average heart rate of using a flat monitor was 77.81bpm totally increased by 7.03% (5.47 bpm). And in the emotionally sensitive section, the average heart rate of wearing a VR headset was 87.65bpm and the average heart rate if using a flat monitor was 80.71bpm totally increased by 8.60% (6.94 bpm).

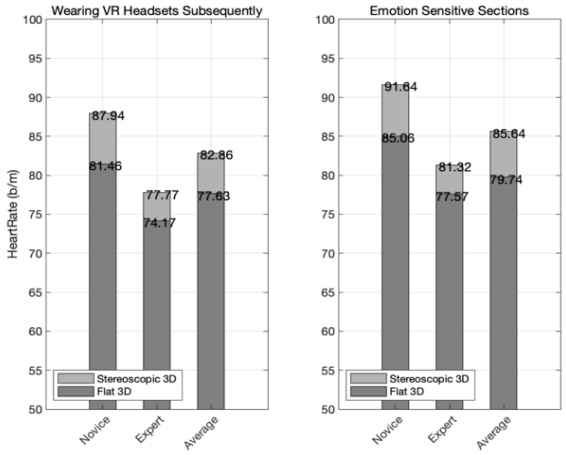


Fig. 8. Results for Heart Rates of the Group Wearing a VR Headset Subsequently.

TABLE. II. AVERAGE OF HEART RATE REDUCTION RATE ACCORDING TO THE ORDER OF WEARING A VR HEADSET

Experiment	**Average of the Entire Operation**	**Average of the Emotion Sensitive Sections**
Wearing a VR headset Initially	11.49% (8.75bpm)	13.3% (10.3bpm)
Wearing a VR headset subsequently	6.74% (5.23bpm)	7.4% (5.9bpm)

The results of the third experiment were summarized in Table II. Both experimental groups showed a higher heart rate in a stereoscopic 3D environment. There is a difference in the reduction rate between the groups. The group that used a VR headset initially showed a greater difference in heart rate in immersion than the group that used VR headset subsequently. In the group with VR headset subsequently, it can be confirmed that the prior experience of the experiment in the flat 3D environment influenced the experiment in the stereoscopic 3D environment.

Based on the above experiment, the authors noticed that the VR immersion in the stereoscopic 3D visualization is stronger than the VR in the flat 3D visualization. In addition, it was confirmed that the degree of heart rate in the stereoscopic 3D visualization is much higher than the flat 3D visualization even in the emotionally sensitive sections.

VI. COMPREHENSIVE ANALYSIS

The three experiments above lead us to conclude that it is more efficient and intuitive to identify the three-dimensional position of a virtual object in stereoscopic 3D visualization rather than the flat 3D visualization in general. However, the authors could not find a consistent reduction rate difference between Novice and Expert groups. This implies that the intimacy of the stereoscopic 3D environment does not significantly influence the difference of immersion in the virtual environment. Table III is a comprehensive summary of the experimental results. In all three experiments, stereoscopic 3D visualization showed excellent overall results. However, in the precision manipulation experiment, the improvement of stereoscopic 3D visualization was relatively insignificant.

TABLE. III. EXPERIMENTAL RESULTS OUTLINE

Experiment	1st Experiment Intuitive Object Movement	2nd Experiment Precise Manipulation	3rd Experiment Immersiveness for Horror Game
Data Acquisition	Starting point search time (second) Object movement time (second)	Starting point search time (second) Object movement time (second)	Heart rate (bpm)
Intuition	The overall average shows a 24.3% reduction of the total operation time in the stereoscopic 3D environment.	The reduction of the total operation time by 19.6% is obtained in the stereoscopic 3D environment.	• Comprehensive evaluations in both groups (with a VR headset initially or subsequently) resulted in a 6.56% increase in heart rate in the 3D stereoscopic environment. • In the 3D stereoscopic environment, an increase in heart rate of 7.92% occurred in three emotionally sensitive sections.

What we need to discuss through the experimental results is that operations in 3D stereoscopic environments can give a similar experience of operations in the real world. This allows us to experience difficult, expensive, or dangerous operations in the real-world with the help of 3D stereoscopic simulations in VR. However, VR has different control variables compared to the real world. For example, VR has not yet such control variables as temperature or noise compared to the real world.

The stereoscopic environment using VR headsets has diverse influencing factors including immersiveness, intuition, and accuracy that studied in this paper. However, there are several other factors that influence users in 3D stereoscopic environments. In future work, the research on the measurement of other influencing factors is needed.

VII. CONCLUSION

In this study, the authors analyzed the improvement of efficiency, accuracy, and immersiveness of stereoscopic 3D visualization through diverse experiments. The plane 2D monitors can be used to represent 3D graphics objects in virtual environments, however, the location of virtual objects that users perceive may be different from the actual location of the virtual object. The degree of distortion, perspective, and immersion make these differences. By presenting a virtual environment as a stereoscopic 3D objects using a VR headset, the intuitive operation performance has been improved and especially operational accuracy has been greatly enhanced.

The results of three experiments conducted in this work showed that the first experiment improved the intuitiveness of the examinees wearing a VR headset. However, the second experiment showed that the accuracy of the examinees who wear a VR headset does not improve meaningfully in the elaborate work requiring accuracy. In the third experiment, the results of the improvement of immersion were observed in the group who wear a VR headset. This study leads us to conclude that the intuition and the immersiveness are outperformed in the case of the stereoscopic 3D visualization using a VR headset. It is also expected that the training effect will be maximized when performing in stereoscopic environments equipped with a VR headset for intuitive and cognitive training using haptic technology.

In this study, the authors experimented only with simple simulations that experience the movement of objects and game simulation. In future work, the authors will examine whether the authors can experience the same effect even if it is applied to complex and collaborative simulations, such as wargame simulation. Meanwhile, some examinees complained of cybersickness in the experiment using a VR headset. Future studies will endeavor to solve the cybersickness that may arise in the stereoscopic visualization method using the HMD.

ACKNOWLEDGMENT

This work was supported by Incheon National University Research Grant in 2015 (No. 2015-1774). The authors would like to thank Mr. Jungwook Kim, Mr. Byeonghun Kang, Mr. Haesung Joo for the help in various experiments.

REFERENCES

[1] L. Jensen and F. Konradsen, “A review of the use of virtual reality head-mounted displays in education and training,” Education and Information Technologies vol. 23, no. 4, pp. 1515–1529, July 2018.

[2] Z. Gao and A. Lécuyer. “A vr simulator for training and prototyping of telemanipulation of nanotubes,” Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST 2008, ACM, pp. 101–104, October 2008.

[3] A. Ewais and O. D. Troyer, “Usability evaluation of an adaptive 3D virtual learning environment,” International Journal of Virtual and Personal Learning Environments (IJVPLE), vol. 4, no. 1, pp. 16–31, 2013.

[4] K. Schönborn, G. Höst, K. L. Palmerius, and J. Flint, “Development of an Interactive Immersion Environment for Engendering Understanding about Nanotechnology: Concept, Construction, and Implementation,” International Journal of Virtual and Personal Learning Environments (IJVPLE), vol. 5, no. 2, pp. 40-56. April-June 2014.

[5] W. S. Alhalabi, “Virtual reality systems enhance students' achievements in engineering education,” Behaviour & Information Technology, vol. 35, no. 11, pp. 919-925, 2016.

[6] A. B. Ray and S. Deb, “Smartphone based virtual reality systems in classroom teaching - a study on the effects of learning outcome,” Proceedings of the 8th IEEE international conference on Technology for Education, IEEE, pp. 68-71, 2016.

[7] L. Alfaro, C. Rivera, J. Luna-Urquizo, S. Alfaro, and F. Fialho, “Virtual Reality Full Immersion Techniques for Enhancing Workers Performance, 20 years Later: A Review and a Reformulation,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 10, pp. 502–510, 2019.

[8] C. Basdogan, S. Mert, H. Matthias, and W. Stefan, “VR-based simulators for training in minimally invasive surgery,” IEEE Computer Graphics and Applications, vol. 27, no. 2, pp. 54–66, March/April 2007.

[9] M. Poyade, A. Lysakowski, and P. Anderson, “Development of a haptic training simulation for the administration of dental anaesthesia based upon accurate anatomical data,” Proceedings of the Conference and Exhibition of the Association of Virtual and Augmented Reality, pp. 143-147, January 2014.

[10] S. H. Shu, E. L. Su, C. F. Yeong, S. Sood, and A. Gandhi, “Development of Virtual Reality Surgical Simulator with Online Database,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC). vol. 9, no.3, pp. 91–94, December 2017.

[11] L. M. Carron and P. W. Horn, "Haptic visual discrimination and intelligence," Journal of Clinical Psychology, vol. 35, no. 1, pp.117–120, 1979.

[12] H. Y. Kim and M. Y. Sung, "Virtual Haptic Visual Discrimination Test," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1, pp. 5–11, 2018.

[13] D. Valkov, F. Steinicke, G. Bruder, and K. Hinrichs, "2D touching of 3D stereoscopic objects," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 1353–1362, May 2011.

[14] G. ruder, F. Teinicke, and W. Stürzlinger, "Effects of visual conflicts on 3D selection task performance in stereoscopic display environments," Proceedings of the 2013 IEEE Symposium on 3D User Interfaces (3DUI), IEEE, pp. 115–118, March 2013.

[15] L. Alfaro, R. Linares, and J. Herrera, "Scientific Articles Exploration System Model based in Immersive Virtual Reality and Natural Language Processing Techniques," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 7, pp. 254–263, 2018.

[16] A. Nomoto, Y. Ban, T. Narumi, T. Tanikawa, and M. Hirose "Supporting Precise Manual-handling Task using Visuo-haptic Interaction," Proceedings of the 7th Augmented Human International Conference (Augmented Human 2016), February 2016.

[17] A. R. Choi, M. Y. Sung, "Performance improvement of haptic collision detection using subdivision surface and sphere clustering," PLoS ONE vol. 12, no. 9, e0184334, September 26, 2017.

[18] A. R. Choi, S. M. Kim, and M. Y. Sung, "Controlling the Contact Levels of Detail for Fast and Precise Haptic Collision Detection," Frontiers of Information Technology & Electronic Engineering (FITEE), vol. 18, no. 8, pp. 1117-1130, August 2017.

[19] P. Cairns, A. Cox, and A. I. Nordin, "Immersion in digital games: review of gaming experience research," Handbook of digital games, Chapter 12, Wiley Online Library, pp. 339–361, March 2014.

[20] M. Malińska, K. Zużewicz, J. Bugajska, and A. Grabowski, "Heart rate variability (HRV) during virtual reality immersion," International Journal of Occupational Safety and Ergonomics (JOSE), vol. 21, no. 1, pp. 47–54, 2015.

Performance Evaluation of E32 Long Range Radio Frequency 915 MHz based on Internet of Things and Micro Sensors Data

Puput Dani Prasetyo Adi[1], Akio Kitagawa[2]
Micro Electronics Research Laboratory (MeRL)
Kanazawa University, Kanazawa
Ishikawa, Japan

***Abstract*—This research discusses how to build and analyze a 915 MHz Long Range (LoRa) E32 Frequency-based Node Sensor network with a Micro Sensor with 3 sensor outputs produced i.e, Temperature (DegC), Air Pressure (hPa), and Humidity (%). therefore, This research succeeded in making a sensor node using the LoRa E32 915 MHz using a mini type ATmega 328p microcontroller with a 3.7 volt, 1000 mAh battery. The display on the receiver uses an 8X2 LCD which will output 3 sensor data outputs. furthermore, the result and analysis of this research are how to analysis of the LoRa Chirp Signal, furthermore, LoRa Chirp Signal obtained from the Textronix Spectrum analyzer in realtime, Quality of Service (QoS), Receive Signal Strength Indicator (RSSI) (-dBm), uplink and downlink data on the Internet Server. Furthermore, The Micro Sensor Graph Output will be displayed on the application server with a sensor data graph. In this research Application Server used is Thingspeak from Mathworks.**

Keywords*—Long range; microcontroller; internet of things; quality of service; micro sensor

I. INTRODUCTION

The development of the Internet of Thing is currently growing rapidly, Sensor Nodes are developed with very low power consumption with long node durability, supported by the right environment and the right battery usage, in research [1] and [2], sensor nodes are used in delivery of Internet-based Pulse Sensors and Blood Pressure sensors with IoT devices with ZigBee sending devices. The advantages of ZigBee are Low Power, with a data rate of 250 kbps, but weak in the distance (m), a maximum of 120 meters on a regular type ZigBee, and 1 km on a ZigBee Pro. So the development is done by changing the radio frequency device with Long Range (LoRa). LoRa center frequency e.g, 433 MHz, 868 MHz, and 915 MHz. for Asia, e.g, Japan and Indonesia use Frequency Long Range (LoRa) 920-923 MHz (Center 915 MHz). The type of signal produced by LoRa is Chirp Signal as in research [3]. In research [4], using Bluetooth type RN-42 in sending LM35 temperature sensor data, and using Raspberry Pi 3 Model B, the resulting analysis is Quality of Services (QoS) on Bluetooth RN42 Master-Slave communication and data management process on Raspberry Pi 3 using Python and MySQL or MariaDB databases. Research [2] was developed with subsequent research using GUI (Graphical User Interface) using JavaScript Object Notation (JASON) and Application Server, e.g, The Things Networks (TTN) and Thingspeak. Fig. 1 shows the comparison between Frequency radio devices e.g, WiFi, Bluetooth, Bluetooth LE, Cellular and LoRaWAN. Using Cellular and GSM technology e.g., in research [5] studies on how to detect water distances using Ultrasonic sensors and GSM SIM 900A. Furthermore, still, on the application of Radio Frequency, namely in research [6], LoRa is used in Street Light monitoring.

This comparison uses two parameters, i.e, the ratio of Power Consumption to distance. With a distance of> = 15 km on Free Space, placing LoRa or LoRaWAN in the best ranking compared to other devices. LoRaWAN can be paired with Cellular Devices in terms of Range but Cellular cannot match LoRaWAN in terms of Power Consumption. Therefore, LoRa and LoRaWAN only use battery power sources with the ability to survive> = 2 years. nevertheless, LoRaWAN is the most ideal device that can be used as a sensor node based on Low Power and with the best data transmission range.

LoRa can> 30 km, in the condition of Free Space, of course, with the right antenna position and an adequate transmitter Antenna height. In research [7], packet loss is calculated for different coverage of SF12 use in indoor and outdoor conditions. In research [8] very detailed in testing the reliability of LoRa signal and LoRa transmission data (bit rate, data rate), Receive Signal Strength Indicator (RSSI), Error Bit Rate (EBR (%)) and EPR (Error Packet Rate (%)), Packet Delivery Ratio (PDR(%) juga dilakukan menggunakan perbandingan simulasi menggunakan LoRaSIM [9].

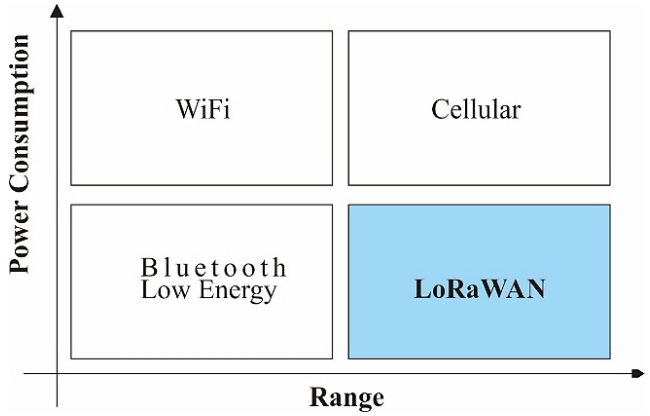


Fig. 1. Radio Frequency Devices with Power Consumption and Range Comparison.

II. Related Works

Research [10] analyzed the Parameter Packet Delivery Ratio (PDR (%)) from direct measurements using Locations A, B, C, D, and E., In theory, these locations show a Spreading Factor (SF) parameter of 7 to 7 12. If the greater the Spreading Factor, the smaller the Packet Delivery Ratio (PDR (%)). The closest distance is A is 650 m and the farthest E is 3400 m. furthermore, the analysis was also carried out by making a comparison of average throughput (bytes / s) with Spreading Factor (SF). The greater the SF value, the smaller the value of Throughput (bytes / s). In SF 7 the average value of 3 payloads is 51 bytes, 25 bytes and 1 byte is 3.96 bytes / s and in Spreading Factor 12, from 3 payloads of 51 bytes, 25 bytes, and 1 byte, the average throughput is 0.22 bytes / s.

Research [3] conducted a detailed analysis of Chirp Signal or CSS. chirp Signal Analysis, including Up chirp and down chirp. This research emphasizes more on the fundamental aspects of LoRa i.e. parameters, demodulation process, and Decoding Process. Furthermore, the CSS Packet Spectogram comparison between Frequency LoRa (125 kHz, 250 kHz, and 500 kHz) against Time (ms) with Spreading Factor (SF) differs from 7 to 12. And the results of the simulation are the comparison between SF and SNR (-dB).

Research [11] analyzed in detail the Signal Interference Ratio (SIR) by comparing the detailed Spreading Factor (SF) Parameters e.g., Co-Spreading Factor Interference. LoRa parameters are obtained with a detailed mathematical approach that results in the Probability of Success (Ps) with parameters e.g., $P_{SNR,}$ $P_{SIR,}$ and Simulation with a Comparison of distances (km). the research compared the value of R = 6 km and R = 12 km.

III. Methodology and Devices Used

A. Chirp LoRa Signal

Compressed High-Intensity Radar Pulse (chirp) is signal modulation technique has long been used in the world of commercial and Military RADAR Systems. therefore, There are 3 types of signal modulation, e.g. Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK) and Phase Shift Keying (PSK). furthermore, Amplitude Shift Keying (ASK) is a modulation signal based on the change in the value of the amplitude is high or the low wave signal is generated, if the amplitude of high-value and low-amplitude digital 1 digital is worth 0. Frequency Shift Keying (FSK) is a modulation signal which is based on the change in frequency indicated by the Pulse Length (t) value or the resulting signal density as shown in Fig. 2 and Fig. 3. In Fig. 2, the parameters are as follows.

t=0:0.05:10;A=1;omega=2;s_t

= t.^2/4;y_t=A*cos(omega*t + s_t)

t is the Pulse Length in Fig. 2 using the value t = 10 and Fig. 3 the value of t = 40.

Like Fig. 2, Fig. 3 has the following parameters

t=0:0.05:40; A=1; omega=2; s_t = t.^2/4; y_t=A*cos(omega*t + s_t);

The signal in Fig. 1 can be formulated with equation 1, while for Fig. 3 it can be expressed with equation 2. If it is formulated with general function the signal for signal density is like equation 3.

$$sin\left(\frac{1}{2}\pi\left(\left(\frac{2x}{3}+1\right)^2-1\right)\right) \quad (1)$$

$$sin\left(\pi((\pi+1)^2-1)\right) \quad (2)$$

$$y(a,b,c,d) = c\sin\left\{\frac{\pi}{b-a}\left[\left((b-a)-\frac{x}{d}+a\right)^2-a^2\right]\right\} \quad (3)$$

while Phase Shift Keying (PSK) is a signal modulation based on signal shifting, not high or low amplitude or signal density. Furthermore, the value of signal modulation is 1 and 0, digital value 1 indicates ON and digital 0 states OFF. of several types of signal modulation, LoRa uses the Chirp type or called the Chirp Spread Spectrum (CSS), there are two types of Chirp namely Frequency Increases (Up-Chirp) and Frequency decrease (Down-Chirp) with time, as shown in Fig. 4.

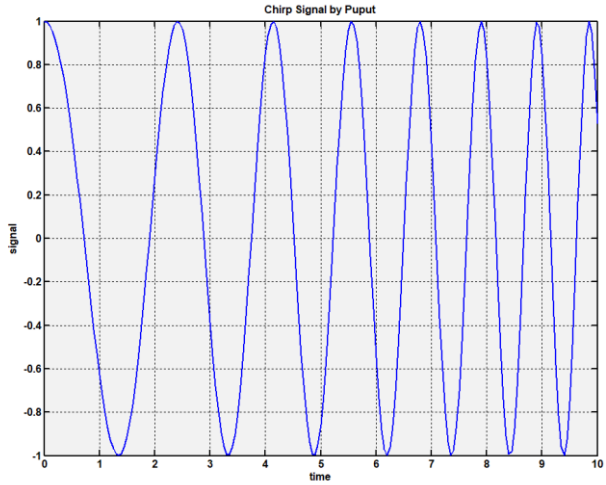


Fig. 2. Chirp with t = 10.

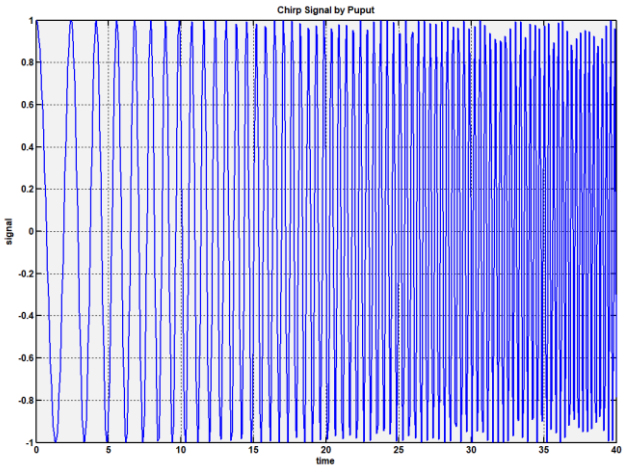


Fig. 3. Chirp with t = 40.

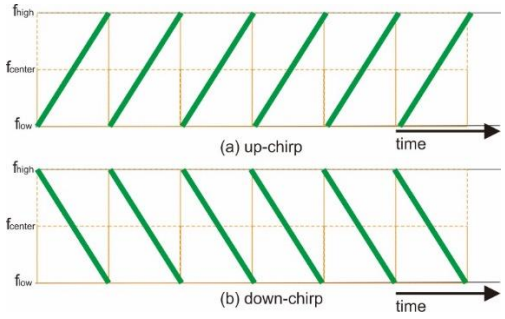


Fig. 4. Up and Down-Chirp Signal LoRa.

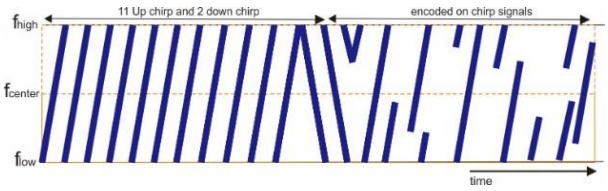


Fig. 5. LoRa Signal with Encoded (data) on.

Fig. 5 contains 2 signal parts i.e., up and down chirp signal on the LoRa signal prefix, consisting of 11 up chirp and 2 down chirp. Next is the encoded signal. Fig. 8 is the realtime LoRa E32 signal that is the Chrip signal captured and analyzed by the Textronix Spectrum analyzer.

In research [12], a decoding technique for LoRa signals was used using a decoding algorithm with two slightly desynchronized superposed LoRa Signals. furthermore, this decoding Algorithm technique is used to increase the effectiveness of sending data e.g, Throughput.

B. BME280 Sensor

The BME280 Sensor is a multi-sensory and applied in the Health Care Application (e.g. Spirometry). in previous research, Health Care Application using different sensors in research [1], [2] i.e. Blood Pressure and Pulse Sensor Internet of Things (IoT) based. BME280 sensor has a three types of Sensors, i.e. Atmospheric Pressure (hPa), Temperature (DegC) and Humidity (%). Atmospheric Pressure (hPa) is pressure at any point in the Earth's atmosphere. Almost the same as the hydrostatic pressure caused by the weight of the air above the measurement point. Table I is the spesifications of the BME 280 Sensor.

TABLE. I. BME 280 SPESIFICATION

No	Parameter	Details
1	Dimension	2.5 mm x 2.5 mm x 0.93 mm metal Lid LGA
2	Digital Interface	I^2C (up to 3.4 MHz) and SPI (3 and 4 wire, up to 10 MHz)
3	Supply Voltage	V_{DD} main supply voltage range 1.7 V to 3.6 V V_{DDIO} Interface voltage range : 1.2 V to 3.6 V
4	Current consumption	1.8 µA @ 1 Hz humidity and temperature 2.8 µA @ 1 Hz Pressure and temperature 3.6 µA @ 1 Hz humidity, Pressure and temperature 0.1 µA in sleep mode
5	Operating range	-40 …+85° C, 100% rel.humidity, 300..1100 hPa

The formula for calculating Atmospheric Pressure is in accordance with the formula 2. Log10 P ≈ 5 - (h -15500), where the value of P is Pressure (Pascal) and h is the height (meters) Furthermore, the temperature has the default unit Degree Celcius. the BME280 sensor, the unit used is Degree Celsius and detects conditions in the environment and results in precision. furthermore, to change to Kelvin (K), Reamur (R) or Fahrenheit (F) C language is used in the Arduino IDE. If the DegC value is known, to be changed to another temperature unit to r = c * 4/5; f = (c * 9/5) +32; k = c + 273.16; furthermore, Humidity (%) is the amount of water vapor in the air that is not visible to the human eye, this amount of water vapor will determine rainfall, dew or fog. Furthermore, Fig. 6 describes the BME280 sensor diagram block, There is a Pressure and Temperature Sensing element that is converted from Analog to Digital using Analog to Digital Converter (ADC) to Logic Gate and continues to 6 Output pins i.e.SDI, SDO, SCK, CSB, VDD and GND, and just 4 pins used i.e. SDA, SCL, VDD (3.3 Volt) and GND.

This sensor will be processed by the MCU (ATmega 328) with BME280 pins on the Arduino Pro mini such as Fig. 3. In Fig. 3, three main components of a sensor node, i.e. MCU ATmega 328, FTDI and LoRa. The LoRa type used is SX1276 a Universal Asynchronous Receiver Transmitter (UART) E32 915T20D, this type of LoRa has the capability of up to 3 km. The LoRa SX1276 UART E32 915T20D uses a default frequency of 915 MHz (920-928 MHz), this Radio Frequency can be used in Japan (920-928 MHz). A block diagram of the BME280 can be seen in Fig. 7.

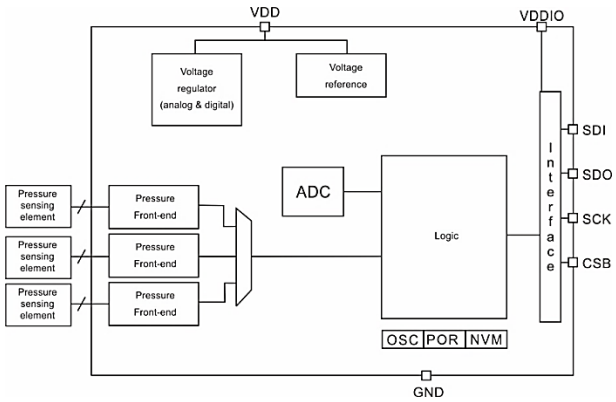


Fig. 6. Block Diagram of BME280 Sensor.

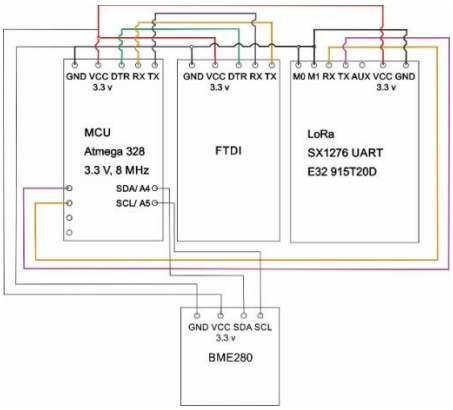


Fig. 7. Block Diagram of BME280 Sensor on MCU ATmega 328.

Battery Life Factor is the most important factor in the LoRa or LoRaWAN network architecture, in research [13] reviewed how to make energy consumption on Sensor nodes in LoRa or LoRaWAN networks can survive well with the approach in Sleep mode. In addition to the Sleep mode approach, the effectiveness of the sensor node and the LoRaWAN network with a multi-node or multi hope is to use a method called the offset-CT Method, as in research [14].

The calculation approach to the node usage on the Receiver is as follows, as in references [1], the calculation of the Power Consumption of Sensor node is based on the total load calculation component used, the Battery life calculation is in accordance with equation 4, furthermore, 0.7 is a value of external factors which can affect battery life. A sensor nodes (Tx and Rx) can be seen in Fig. 8. Furthermore, The type of battery used in the sensor node is the 3.7 Volt 1000 mAh Lithium Polymer Li-Po Rechargeable Battery as shown in Fig. 9.

1) Measurement of battery life (H) of receiver node: In measurements using the Ampere meter in Fig.10, the total value of the Power Consumption mote or node sensor on the receiver (Rx) with LoRa E32 is 22 mA. so that the Battery Life mote according to equation 4 is 1000 mAH / 22 mA = 45.45 x 0.7 = 31.8 hours. accordingly, This formula is the same as finding power (P (watts) = V (Volts) x I (Amperes) = Psensor = 3.7 x 0.022 = 0.0814 watts, so Pbattery = 3.7 x 1 Amperes = 3.7 watts, so the Power Required is 3.7 / 0.0814 = 45.45 x 0.7 = 31.8 Hours.

2) Measurement of battery life (H) of transmitter node: On the transmitter, the total current sensor node is 119.5 mA, then Battery Life mote = 1000 mAH / 119.5 mA = 8.36 x 0.7 = 5.8 hours. This formula is the same as finding power (P (watts) = V (Volts) x I (Amperes) = Psensor = 3.7 x 0.1195 = 0.44215 watts, so Pbattery = 3.7 x 1 Amperes = 3.7 watts, so the Power Required is 3.7 / 0.44 = 8.4 x 0.7 = 5.8 Hours. Furthermore, to reduce the Battery life (H) Sensor node it is necessary to do a strategy on the sensor node by changing the C language programming on the MCU to Sleep mode. With Sleep mode or Reset mode, Battery Life (H) can be longer, in Sleep mode it can be set how long the sensor node will sleep, and when the sensor node will turn on again and send a signal.

C. LoRa 915 MHz (920-928 MHz)

The RF signal used in this research is two types, i.e. Dragino LoRa 915 MHz and LoRa SX1276. LoRa SX1276 uses the default frequency of 915 MHz (Japan: 920-928 MHz), this type of LoRa is intended for Arduino Uno Shield and Arduino Mega. Complete data about the specifications of the LoRa SX1276 915 MHz can be seen in Table I. Furthermore, this research is divided into two analyzes, i.e. the measurement of transmission devices (Transmitter and Receiver) to get the RSSI (-dBm) value and analysis of the 915 MHz Internet Gateway e.g. uplink and downlink of LoRa 915 MHz Internet gateway. In research [8] the Framework of the Distance-Ring Exponential Stations Generator (DRESG) was introduced in handling and providing Multi-hop solutions, e.g, Routing nodes, energy efficiency and consumption among all the STAs in the Network. In research [14] conducted research transmission data using LoRa to find out energy efficiency over very long distance [15] by building communication 6 LoRa nodes which can cover 1.5 ha of network with sensor node specifications ie, 2 AA type batteries, data transmission every 5 seconds and level 80% reliability. The Lora E32 or SX1276 / SX1278 Wireless Module has variants or versions, one of E32 915T20D LoRa type shown in Fig. 5 and this type is used in this research, LoRa E32 868T20D (Permit at Japan 920-928 MHz) this Frequency radio is maximum 931 MHz, moreover the E32915T20D uses a default frequency of 915 MHz, while the E32868T20D uses an 868 MHz frequency. e.g, for the Japanese region using E32915T20D, which works in the range 920-928 MHz, therefore, the maximum frequency for E32868T20D module is 930 MHz.

The dimensions of the LoRa E32 915T20D Wireless Module have dimensions of 21x36 mm in Fig. 11. the voltage used is 3.3 Volt DC (Tipycal), Transmit Power (dBm) is 20 dBm in Fig. 12, more complete can be seen in Table II.

$$Battery\ Life\ Sensor\ Node\ [H] = \frac{Battery\ Capacity\ (mAh)}{Load\ Current\ (mA)}\ x\ 0.7 \quad (4)$$

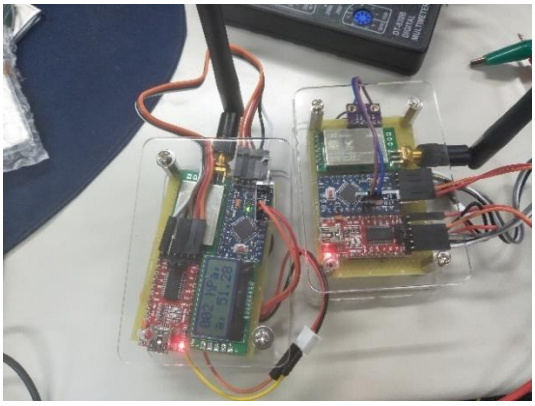


Fig. 8. DIY Transmitter and Receiver LoRa E32 915 MHz Node Sensor.

Fig. 9. Battery 3.7 Volt 1000 mAh Lithium Polymer Li-Po Battery Rechargeable.

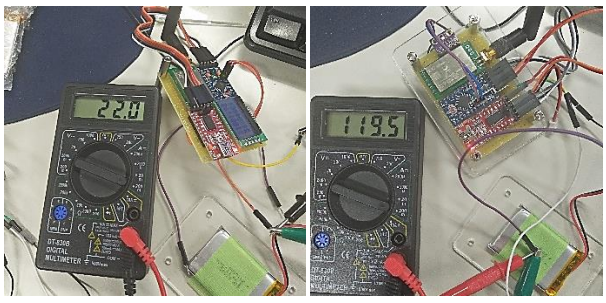


Fig. 10. Total Load Measurement of Current (mA) on Rx dan Tx Sensor Node.

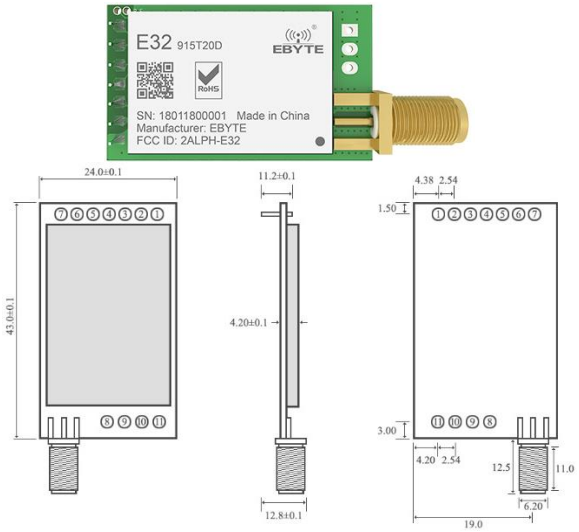


Fig. 11. LoRa E32 915T20D and Dimension.

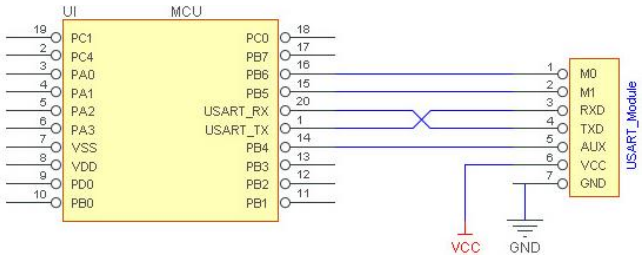


Fig. 12. Circuit Diagram of LoRa E32 915T20D.

TABLE. II. LORA E32915T20D SPECIFICATION

No	Parameter name	Parameter detail
1	Model Wireless Module	E32 (915T20D)
2	Frequency	Min 900 MHz, Typ : 915 MHz : Max 931 MHz
3	Voltage Supply	Min: 2.3 V DC, typ: 3.3 V DC, Min: 5.2 V DC
4	Communication Level	Min: 2.5 V DC, typ: 3.3 V DC, Min: 3.6 V DC If Communication level > Maximum Value, then Module may be damaged
5	TX Power	Min :19.3 dBm, typ: 20 dBm, Max:20.6 dBm
6	TX Current	Min:110 mA, typ :120 mA, Max : 130 mA
7	RX Current	Min:11 mA, typ : 14 mA, Max :15 mA
8	Turn off Current (Sleep mode)	Min:3 mA, typ : 4 μA Max :5 μA
9	Range	Min: 2700 m, Typ : 3000 m, Max :3300 m
10	Packing	DIP
11	Air Data rate (kbps)	2.4 kbps
12	Antenna	SMA-K
13	Reception Sensitivity	-147 dbm
14	Air data rate	0.3 kbps – 19.2 kbps, default 2.4 kbps
15	Test distance	3000 m (3 km)
16	Antenna Gain	5 dBi
17	Communication Interface	UART, boud rate 1200 – 115200, default: 9600 bps
18	Transmitting Power	100 mW
19	Parity	8N1
20	Default Parameter Values	C0 00 00 1A 0F 44

In general, Lora type E32915T20D has 7 core pins that are used in the process of making sensor nodes, namely M0, M1, RXD, TXD, AUX, Vcc and GND, in full about pins and functions explained in Table III. There are 3 modes in sending sensor data to Pins M0 and M1 are Wake up mode (mode 1), if M1 and M0 = 1, then Power-saving mode (mode 2) if M1 = 1 and M0 = 0 and Sleep mode if M1 = 1 and M0 = 1.

A Zigbee using PAN ID (Personal Address Network) Identifiers, e.g, 00 03 04 AA BB CC (00 03 = Address, and 04 = Channel) will send data to other LoRa Wireless Modules, the receiver (Rx) LoRa Wireless Module have the same address and channel, e.g, 00 03 04, if address 00 05 04, message will not be sent or 00 07 04 (different address even though the channel is the same, the message will not be sent) while Broadcast Mode, FFFF will send all addresses and channels such as FF FF 04 AA BB CC can send data to 00 03 04 and 00 05 04 (Different addresses but for channels must be the same), e.g, channel 00 07 06, the data is not sent, nevertheless the channel must be 4, same as Tx (Transmitter) Lora Wireless Module. Writing the address module on the broadcast is 0xFFFF or 0x0000 and channel 0x04, while for Fix mode or point to point is 0x0003 or 0x0001.

In research [15], a study of the basic performance of LoRa, LoRa Physical layer and comparison of Sigfox and LoRa in terms of Bit Error Probability. In research [16] LoRa research on Quality of Services (QoS) e.g, RSSI, SNR, Payload size (bytes) and Spreading Factor. In research [17] LoRa performance testing on image transmission, this is a new idea because LoRa has the ability to transmit small data, in this research the image transmission was successful, the camera used has a SNR 45 dB specification and 38400 bps Boudrate, Micro SD Card as saving data and LoRa shield with Arduino MEGA as a processor (MCU) and transmitter. In research [18] testing and study of the maximum or limit LoRaWAN devices. LoRa test results in a comparison of Payload data of 10, 30 and 50 bytes, and the maximum throughput of the number of different end nodes, namely, 250, 500, 1000 and 5000 end devices.

TABLE. III. LORA E32915T20D PINS

Pin no.	Pin	Pin direction	Application
1	M0	Input (weak pull-up)	Work with M1 and decide the four operating modes
2	M1	Input (weak pull-up)	Work with M0 and decide the four operating modes
3	RXD	Input	Connect to external TXD output pin
4	TXD	Output	Connect to external RXD input pin
5	AUX	Output	To wake up the external MCU
6	VCC	Input	Voltage positive references of module; Power Supply 2.3 Volt – 5.2 Volt DC
7	GND	Input	Ground
8	Fixing hole		Fixing hole
9	Fixing hole		Fixing hole
10	Fixing hole		Fixing hole

In research [10], the LoRa test on the Packet Delivery Ratio of the LoRa field test in SF 7, 9, 12 at 5 different locations. Also obtained a comparison of the average Spreading Factor to Throughput (bytes/s). in research [19] data generated from research on LoRa are AWGN Channel and Rayleigh channel, Loss of Received signal Power to maintain PER = 10^{-1} as in AWGN channel when the end-devices velocity is 30 km / h.

D. Fresnel Zone Approach

In this research, transmitting data from the 915 Mhz LoRa Transmitter to the 915 MHz LoRa Receiver with a long distance (Km). this affects the geographical location of the earth in the form of an ellipse. So a theory about the Fresnel Zone is formed. Fresnel Zone is an elliptical form on the Direct Line of Sight condition between the transmitter and Receiver. furthermore, The formula for calculating radius in the Fresnel Zone is as r=8.657 x √(D/f), r = Fresnel Zone radius (m), D = distance (km), f = frequency (GHz). so that from equation 1 a table of comparison of Fresnel Zone can be made radius (m) at a certain distance and certain Frequency, for example at a distance of 500 m or 0.5 Km and Radio Wave Frequency used is 915 MHz or 0.915 GHz then the Fresnel radius of Zonen r = 8,657 x √ (0.5 / 0.915) = 6.3994 meters, this condition is illustrated in Fig. 5. Conditions when sending BME280 sensor node data from the transmitter to the receiver node which can be seen from the Fresnel Zone Clearance condition. The Fresnel zone is the ellipse part that shows in Fig. 5(a), outside the ellipse part is not the Fresnel Zone. For this reason, the Fresnel Zone is influenced by the transmitting antenna height (H_T) and the receiving antenna height (H_R). analysis of data transmission between Transmitter (T_x) and Receiver (R_x) LoRa node is determined by the Fresnel zone to obtain the correct delivery results, in addition to the influence of the Fresnel zone, other influences, an obstacle (buildings, trees) and interference e.g.weather.

There are four states of the Fresnel Zone described in Fig. 13 and Fig. 14. Fig. 13(a) is the Fresnel Zone Equation Fig. 13(b) is the Condition Transmitter and Receiver without interference or no obstacle in the Fresnel Zone, Fig. 13(c) is Condition Transmitter and Receiver with a reflected signal from the ground surface and Fig. 13(d) is Conditional Transmitter and Receiver with considering factor H, H is heigh from curvature of the earth and to calculate H = 1000 x D ^ 2 / (8 x Rearth), accordingly the theory, H is Height (or earth curvature allowance in m), D is the distance between the end node and the gateway in km, and Rearth is earth, the radius in km equal 8504 km. In full about the effect of Height (m) on the results of the calculation approach of the Fresnel Zone can be seen in Table IV. Moreover, Table IV is a comparison between the distance (km) to the height (m) of the arch of the earth, and this is an approach that is not significant right, considering the different areas when transmitting data.

The essence of adding Fresnel Zone parameters is to take into account the earth's curvature which affects Reflection, Diffraction and Scattering wave signal propagation, in this case, the position of Transmitter is important, tx should be placed at a position far higher than the Fresnel zone. Therefore, Fresnel Zone is used at distances between 5 km to 30 km.

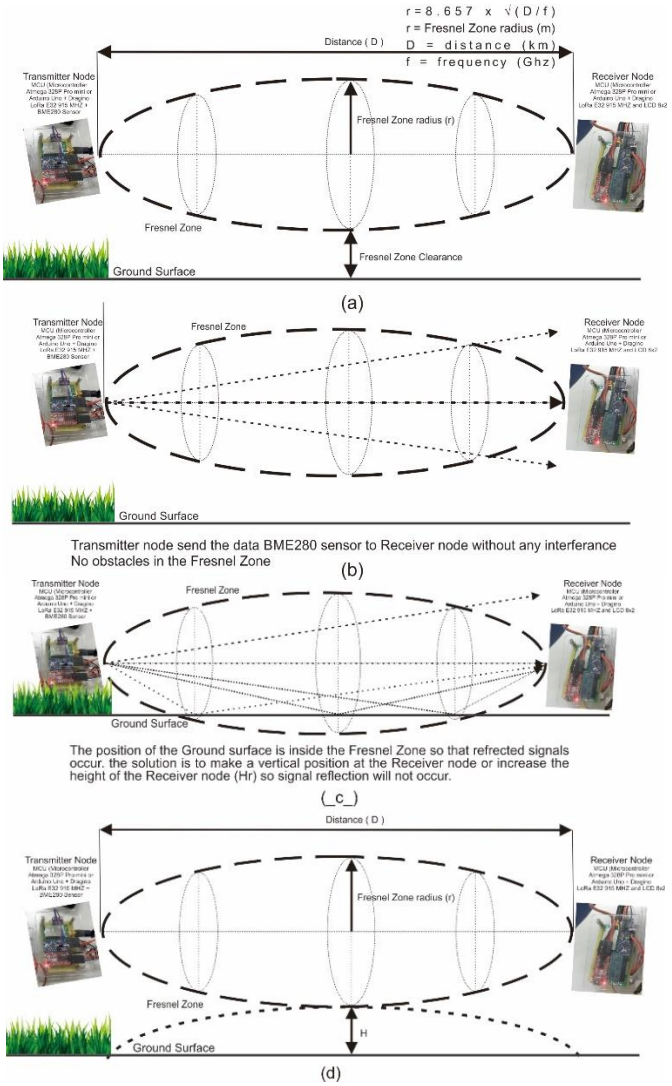


Fig. 13. Conditions of Fresnel Zone Theory.

TABLE. IV. RESULT OF FRESNEL ZONE APPROACH WITH H CONSIDERATION

No	Distance (km)	Height (m)
1	0.1	Negligible
2	0.5	Negligible
3	1	Negligible
4	2	Negligible
5	5	0.4
6	10	1.5
7	15	3.3
8	20	5.9
9	25	9.2
10	30	13.2

Furthermore, the Fresnel Zone area is influenced by Ht and Hr, the higher the value of Ht and Hr, the greater the percentage of Fresnel zone clear, as shown in Fig. 14.

Then from the formula r = 8,657 x √ (D / f), it changes to r = 8,657 x √ ((0.7 x D) / f), this is caused by adding a Fresnel zone value of 70% or 0.7 x D.

Therefore, Fig. 15 shows the Comparation of Fresnel Zone Clear Percentage (%) with the Fresnel Zone approach with a

distance of 5 km produces a value of Fresnel Zone radius 100% clear as far as 639.9 m, Fresnel Zone 70% clear as far as 535.4 m, Fresnel Zone 60% as far as 495.6 m, and Fresnel zone 50% as far as 452.5 m. Furthermore, from Fig. 16 it can be concluded that the farther the distance (m) between the transmitter (Tx) to the LoRa (Rx) receiver, the greater the value of r (Fresnel zone), so that it will be more detailed if the addition of data is done using a comparison of Frequency The different LoRa are 433 MHz and 868 MHz in Fig. 16.

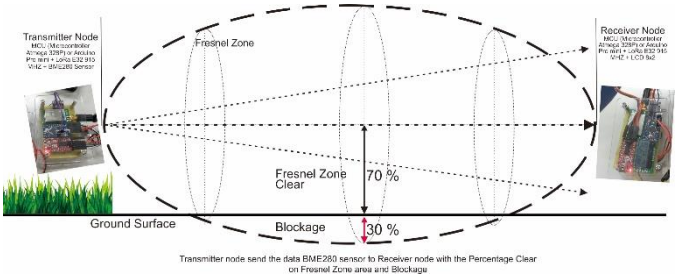


Fig. 14. Fresnel Zone 70% Clear.

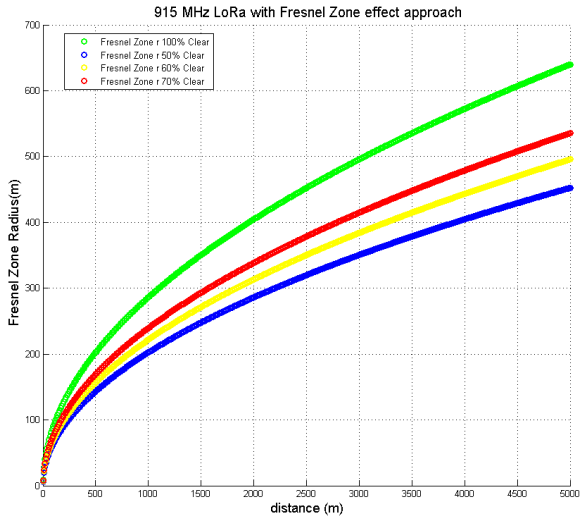


Fig. 15. Comparation of Fresnel Zone Clear Percentage (%).

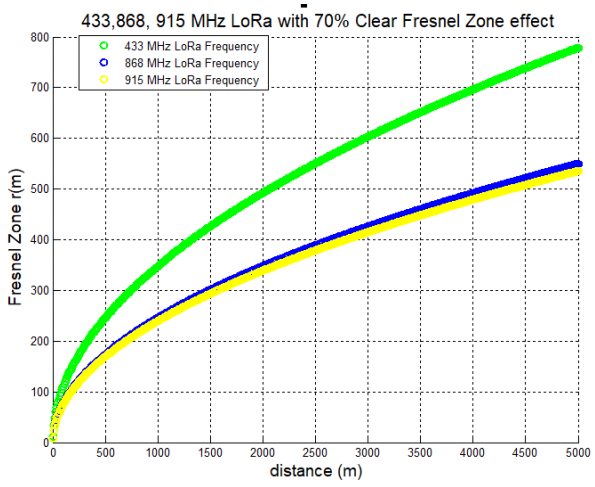


Fig. 16. Comparation of 70% Clear (30% Blockage) Fresnel Zone with different LoRa Frequency (433, 868 and 915 MHz).

In Fig. 16, 433 MHz Frequency of LoRa at a distance of 5 km has a value of Fresnel Zone as far as 778.3 m, 868 MHz as far as 549.7 m and 915 MHz as far as 535.4 m. so it was concluded that by transmitting LoRa data with Low Frequency and with the same Fresnel zone clear percentage condition (70%) has the furthest Fresnel zone (r) value. Please note that the H value is calculated when the distance is km 5 km. so the value of r ≥ 5 km equals r + H, for example, r + 0.4 at a distance of 5 km and r + 1.5 at a distance of 10 km and so on as shown in Table IV.

E. Network Concept

Fig. 17 and Fig. 18 is a concept on this research, end devices/end nodes or mote consists of more than one sensor node connected to the LoRa Internet Gateway. Therefore, this LoRa Internet Gateway functions to store and capture LoRa signals from end devices to be forwarded to the e.g compatible Application Server, The Things Network or Thingspeak and then forwarded to Devices connected to Internet devices. The analysis stage is the data rate (byte/s) of the Uplink and Downlink process, continuous data transmission by the sensor node will make the battery run out quickly and this is undesirable in the LPWAN method.

In Fig. 17, by only using 1 LoRa Internet Gateway, packet data or data bytes will be a buildup of data bits from n-number end-nodes that send data to one source, consequently, this multiplexing method results in increasing the bit rate of Error Ratio (BER (%)) or Packet Error Ratio (PER (%)). therefore, we need an approach to the number of gateways that can dynamically manage packet data (bytes/s) coming in from end-node or end-device. End-node needs to use a sleep mode approach, therefore, it didn't continuously send packet data (bytes/s) without stopping causing bottleneck packet data (bytes) at the gateway. The addition of the gateway an answer, furthermore, the end node can select the destination gateway furthermore, that there no packet data buildup resulting in a reduction in throughput.

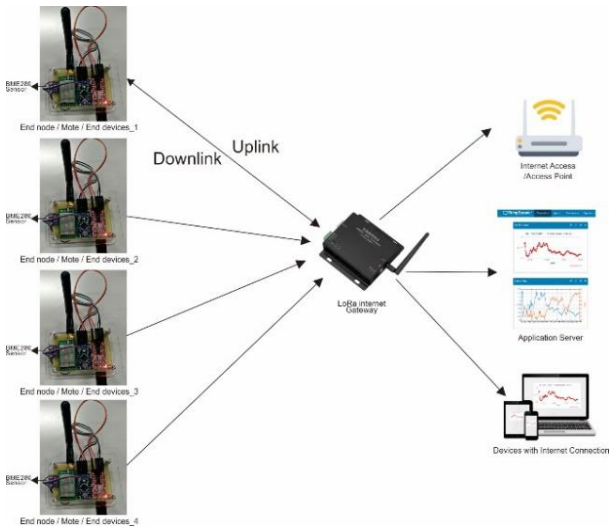


Fig. 17. Simple TT Network Concept.

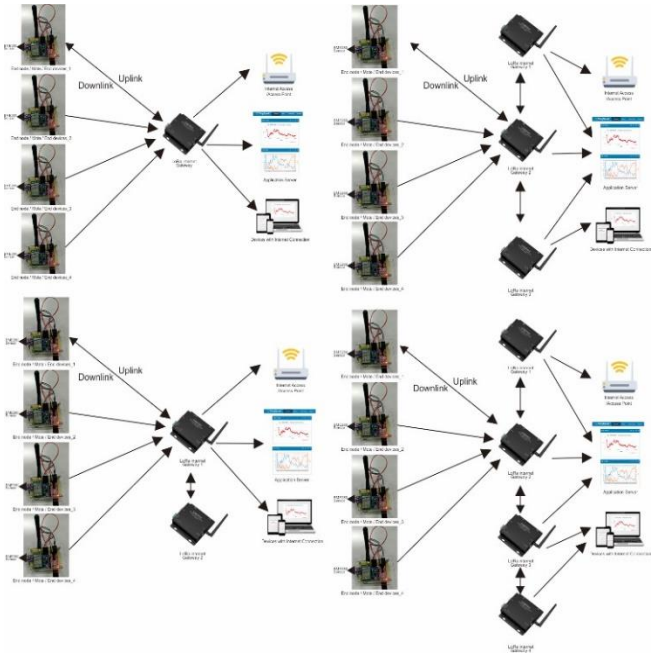


Fig. 18. Adaptive Data Rate (ADR) Algorithm TT Network Concept.

So we need an approach using an algorithm that can regulate the sending of data (byte/s) continuously by the sensor node or end node/mote to the gateway. The core of the ADR algorithm is that sensor nodes can dynamically switch to off or sleep mode so that power consumption can be reduced, therefore the lifetime on sensor nodes is getting longer. Therefore, the Adaptive Data Rate (ADR) algorithm is used to optimize data rates, airtime and energy consumption, the ADR Flowchart in Fig. 19.

F. Sleep Mode

Sleep mode is one way to manage LoRa data transmission using MCU, in this research the Arduino Pro mini MCU type 3.3 v, 8 MHz. in general the structure of the Sleep mode can be seen in Fig. 20.

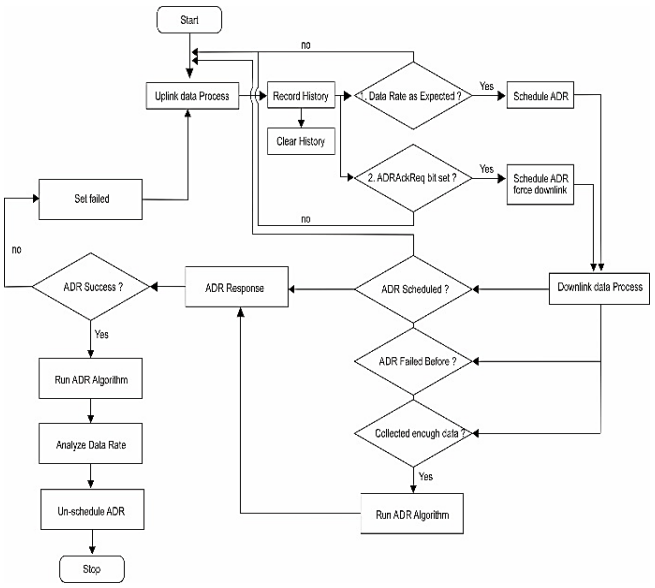


Fig. 19. Adaptive Data Rate (ADR) Flowchart.

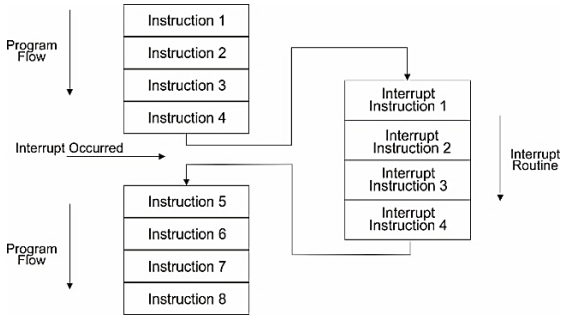


Fig. 20. Interrupt Process.

```
1.  Initialization Sleep mode and LoRa Library
#include <RH_RF95.h>
#include <ThingSpeak.h>
#include <SPI.h>
#include <LoRa.h>
#include <avr/sleep.h>
#define interruptPin 2
2. Initialization Sensor Library
#include <Wire.h>
1.  Initialization  Boud  Rate  or  Speed  Data
    Rate(bps)
void setup() {
Serial.begin(115200);
2.  Initialization Input and Output

Serial.println("LoRa Sender");
 if (!LoRa.begin(915E6)) //915MHz Freq
  {
  Serial.println("Starting LoRa failed!");
}
pinMode(LED_BUILTIN,OUTPUT);
pinMode(interruptPin, INPUT_PULLUP);
digitalWrite(LED_BUILTIN,HIGH);
}
3.  Determine the Sleep Time
void loop() {
delay(10000);
Going_To_Sleep();
}
 4.  Command to Sleep
void Going_To_Sleep(){
 rf95.sleep();
 sleep_enable();
 attachInterrupt(0, wakeUp, LOW);
 set_sleep_mode(SLEEP_MODE_PWR_DOWN);
 digitalWrite(LED_BUILTIN, LOW);
 delay(1000);
 sleep_cpu();
 Serial.println("Stop  Transmit  data  LoRa  and
Sleep! ");
 digitalWrite(LED_BUILTIN,HIGH);
}
 5.  Command to WakeUp
void wakeUp() {
 Serial.println("Transmit data LoRa!");
 sleep_disable();
 detachInterrupt(0);
}
====Pseudocode 1. Sleep Mode Use Interrupt =====
```

IV. Result and Discussion

A. BME280 Sensor Output

Realtime data BME280 Sensor on the Serial monitor is an example of the output generated from the end node transmitter (Tx) to the Receiver (Rx). Shown in Fig. 21 and Fig. 22,

therefore, this output can be added to the analysis parameters, namely Receiver Signal Strength Indicator (RSSI) as in Fig. 23.

Fig. 22 is the Realtime Sensor node taken from the Serial monitor with a 9600 bps baud rate which is indicated by realtime time, temperature, humidity, and Air Pressure data.

Fig. 23 is the sending of Sensor data from the Transmitter (Tx) and Receiver (Rx) accompanied by RSSI Parameters, in addition to RSSI, SNR is also very necessary to determine the strength of the LoRa signal.

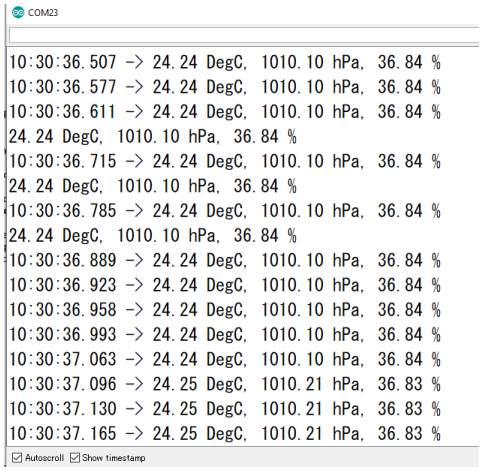
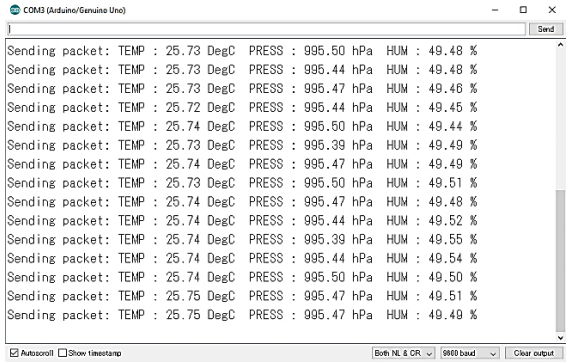
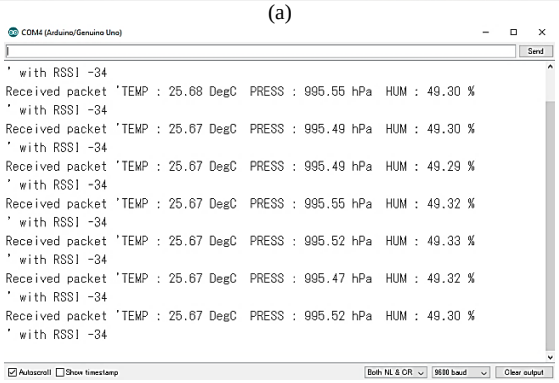


Fig. 21. Realtime Sensor Data from End Node.



(a)



(b)

Fig. 22. Data Transmit (a) and Receive (b) on Serial Monitor with RSSI (-dBm) value.

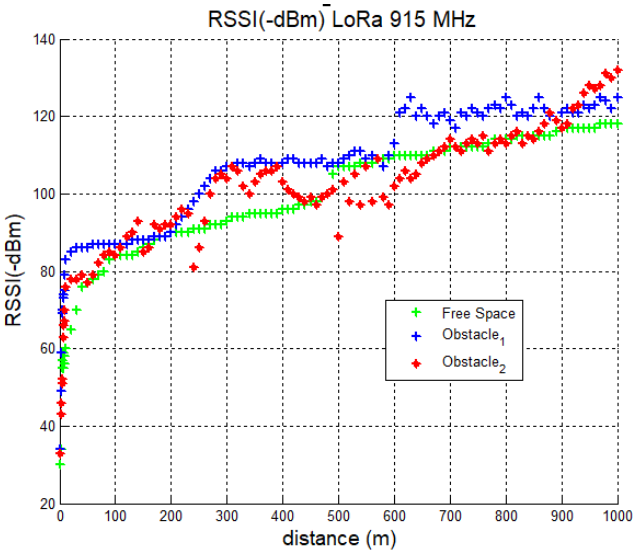


Fig. 23. RSSI (-dBm) LoRa 915 MHz.

B. *Signal Analyze with Analyzer*

Fig. 24 and Fig. 25 shows when the LoRa transmitter is on, it will appear in the Analyzer a noise and signal that moves at the amplitude and signal strength (-dB) at a frequency that matches the radio Frequency LoRa used. (a) is the high LoRa signal when sending sensor data, in this section expressed by the high signal that is the Amplitude signal, while (b) is the Noise, which is part of the LoRa Frequency signal or radio wave. Fig. 24 is the LoRa signal which consists of two parts, the LoRa signal and the noise signal at the bottom. Therefore, Power Noise can be determined from the strength of the LoRa signal.

Fig. 26 is analog demodulation, i.e. IQ versus time, Inphase and Quadrature (IQ) is two amplitude sinusoidal waves with phase is one-quarter cycle (π / 2 radians). Inphase and Quadrature signals are expressed in equation x. in realtime the Inphase and Quadrature signal can be seen in Fig. 8. The yellow color is the amplitude of the In-phase signal and the Green color is the amplitude of the Quadrature signal. Quadrature and phase show the sending of a LoRa signal. therefore, LoRa uses the FSK (Frequency Shift Keying) method if it is shown in the wave signal amplitude in Fig. 27, then the signal density shows the sending of LoRa data in realtime and a low signal or low frequency indicates no LoRa signal.

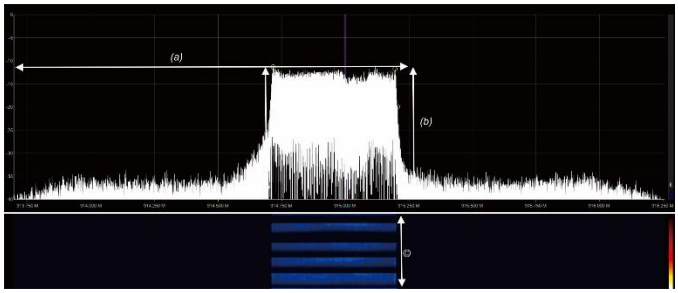


Fig. 24. LoRa Signal Analyze with SDR v.1.0.0.1700 RTL-SDR.

$$s(t) = e^{j(2\pi f_c t + 2\pi \frac{\beta}{2} t^2)} \quad (4)$$

$$\beta = \frac{BW}{T_{symb}} \quad (5)$$

$$T_{symb} = \frac{2^{SF}}{BW}.\text{CR} \quad (6)$$

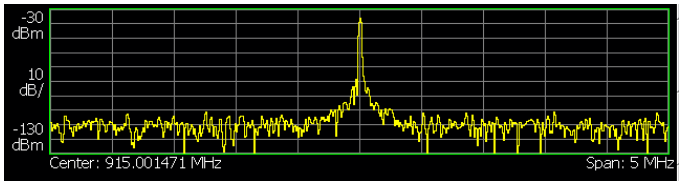


Fig. 25. LoRa Signal Analyze with Textronix RSA 3408B Analyzer.

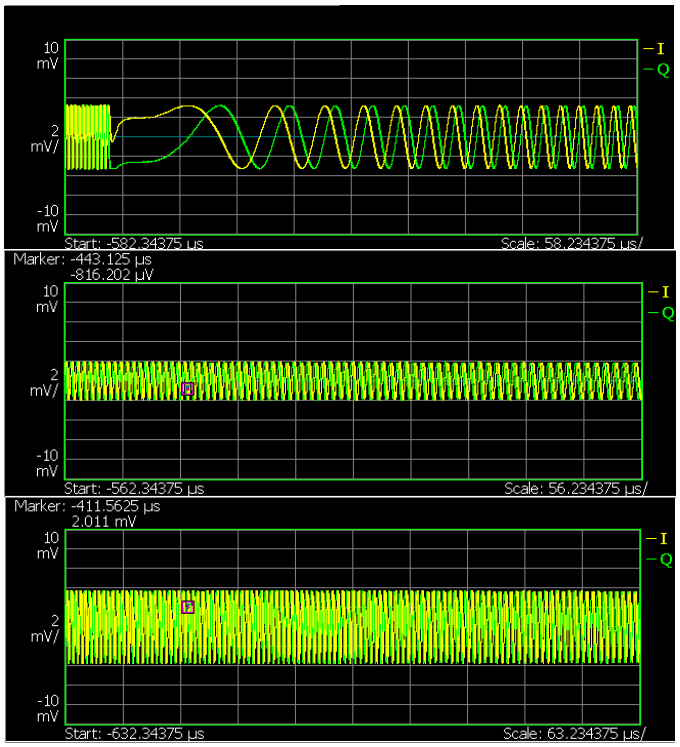


Fig. 26. Inphase and Quadrature (IQ).

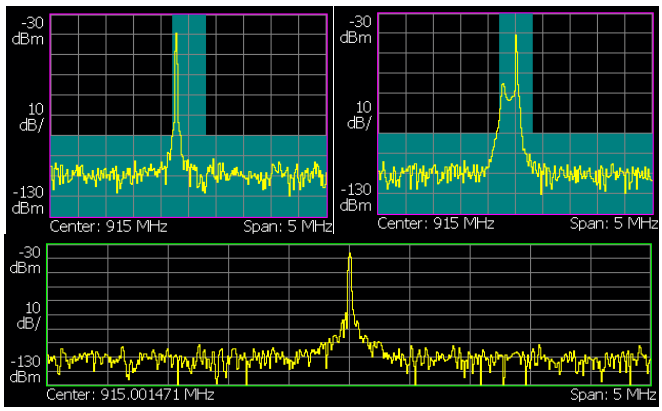


Fig. 27. Transmit Power LoRa Signal (dB).

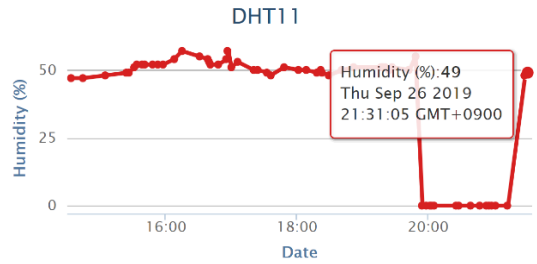
C. Output Sensor on Application Server

The sensor output is shown in the Fig. 28(a) and (b). Taken from the application server e.g, Thingspeak, TTn. This data can be captured in realtime by smartphones, desktop computers, mini PCs or other devices connected to the internet network, so that data can be received easily and quickly, dynamically and freely, wherever they are.

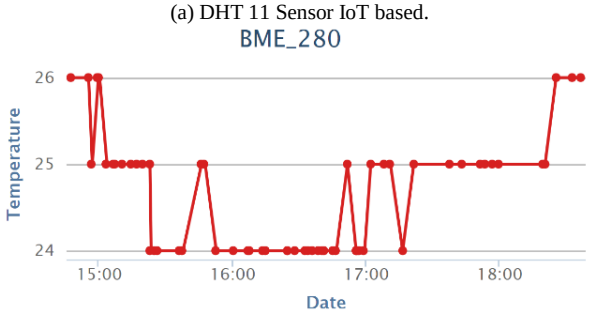
D. Realtime Chirp Sinyal

During data transmission, the LoRa signal called CSS signal modulation (Chirps Spread Signal) can be seen in real-time using a Signal analyzer. Chirps signals consist of 2 Up Chirps and Down Chirps which are also called Preambles. And when the data is received, the Signal Chirps tend to be intermittent, because it contains LoRa data that is being transmitted. furthermore, Fig. 29(a), (b), (c) and (d) are examples of various Chirps LoRa signals.

Sensor node management is important to improve energy savings in the battery. therefore a requirement for analysis of a method using Arduino C Language to change the node to sleep mode and from the Gateway or Server-side is to use the Adaptive Data Rate (ADR) Algorithm method meaning the sensor node can dynamically transmit data on Gateway-1, Gateway-2 to Gateway-n, the adaptive means the data rate can be dynamically controlled by each node, the important function is for the effectiveness and management of the battery and sending data. furthermore, this will also reduce the percentage (%) of the Bit Error Rate (BER) and Packet Error Rate (PER) values. Therefore, This sensor node management pattern is to maintain Long Life or sensor node Lifetime in Wireless Sensor Network (WSN) which has a large number of nodes and by using LoRaWAN will enable Wireless Sensor Network Architecture with a wide range and small bit rate for monitoring data the sensor.



(a) DHT 11 Sensor IoT based.



(b) BME_280 Sensor IoT based.

Fig. 28. (a) and (b), Output Sensor pada Application Server Thingspeak.

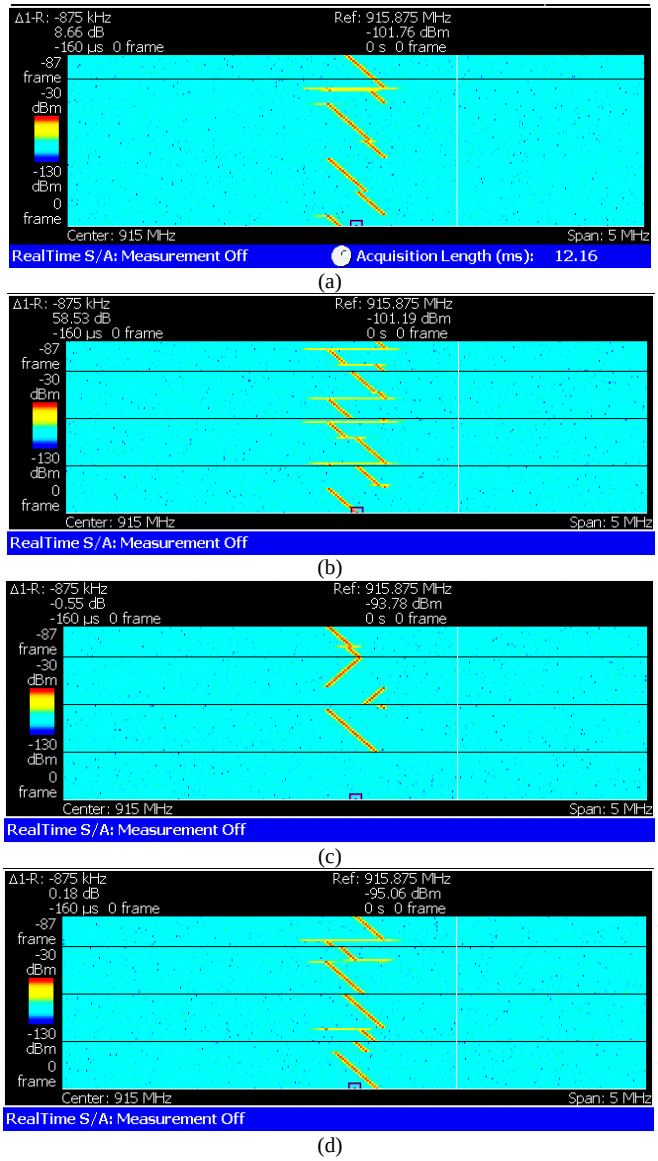


Fig. 29. (a) Start Chirp Signal (Preamble) (b) Realtime Chirp up-Down Chirp and Encoded Signal(c) Up and Down Chirp (d) Encoded Message Signal.

V. Conclusions

Transmission packet data (bytes/s) Long Range (LoRa) from End Devices or End Nodes using BME280 and DHT11 sensors Temperature sensor has been successfully. and furthermore, the packet data (bytes/s) sent to the ThingSpeak application server and can be seen in realtime on devices that have an internet connection. Furthermore, the LoRa signal sending data realtime can be received by the Signal Analyzer according to the LoRa Signal Radio Frequency used in this research and produce a Signal Chirp (CSS). In the next research, it is hoped that it can be further investigated, i.e., broadcast data transmission, therefore, that the weaknesses of the Gateway can be identified and what and what algorithm approaches can be used so that the sensor data analysis is detailed.

Acknowledgment

Thanks for Micro Electronics Research Laboratory (MeRL) Kanazawa University, JAPAN for supporting through a materials e.g., devices, analyzer. and to Professor Akio Kitagawa who has guided the writer and has provided support and role models.

References

[1] Puput Dani Prasetyo Adi and Akio Kitagawa, “ZigBee Radio Frequency (RF) Performance on Raspberry Pi 3 for Internet of Things (IoT) based Blood Pressure Sensors Monitoring” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. DOI: 10.14569/IJACSA.2019.0100504.

[2] Puput Dani Prasetyo Adi and Akio Kitagawa, “Quality of Service and Power Consumption Optimization on the IEEE 802.15.4 Pulse Sensor Node based on Internet of Things” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. DOI: 10.14569/IJACSA.2019.0100518.

[3] V. Fialhoab, F. Azevedoa, "Wireless Communication Based on Chirp Signals for LoRa IoT Devices",i-ETC: ISEL Academic Journal of Electronics, Telecommunications and Computers, IoT-2018 Issue, Vol. 4, n. 1 (2018) ID-6.

[4] Puput Dani Prasetyo Adi and Akio Kitagawa, “Performance Evaluation WPAN of RN-42 Bluetooth based (802.15.1) for Sending the Multi-Sensor LM35 Data Temperature and RaspBerry Pi 3 Model B for the Database and Internet Gateway” International Journal of Advanced Computer Science and Applications(IJACSA), 9(12), 2018. DOI: 10.14569/IJACSA.2018.091285.

[5] Puput Dani Prasetyo Adi, Rahman Arifuddin ,”Design Of Tsunami Detector Based Sort Message Service Using Arduino and SIM900A to GSM/GPRS Module”, JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science) Volume 1, No.2, 2019, ISSN : 2614-4859, DOI:10.26905/jeemecs.v1i1.1982.

[6] Rudy Susanto, Wiedjaja, James Anthony, Jacksen, "Comparison of three LoRa devices and its application on street light monitoring system", 2018 IOP Conf. Ser.: Earth Environ. Sci. 195 012066, doi:10.1088/1755-1315/195/1/012066.

[7] Jetmir Haxhibeqiri, Eli De Poorter, Ingrid Moerman and Jeroen Hoebeke, "A Survey of LoRaWAN for IoT: From Technology to Application", Sensors 2018, 18, 3995; doi:10.3390/s18113995.

[8] Marco Cattani, Carlo Alberto Boano and Kay Römer,"An Experimental Evaluation of the Reliability of LoRa Long-Range Low-Power Wireless Communication", Institute for Technical Informatics, Graz University of Technology, Graz 8010, Austria; J. Sens. Actuator Netw. 2017, 6, 7; doi:10.3390/jsan6020007.

[9] Muhammad Omer Farooq, Dirk Pesch "Extended LoRaSim to Simulate Multiple IoT Applications in a LoRaWAN", International Conference on Embedded Wireless, Systems and Networks (EWSN) 2018.

[10] Aloÿs Augustin, Jiazi Yi, Thomas Clausen and William Mark Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things", Sensors 2016, 16, 1466; doi:10.3390/s16091466.

[11] Amir Mahmood, Emiliano Sisinni, Lakshmikanth Guntupalli, Raul Rondon, Syed Ali Hassan and Mikael Gidlund, “Scalability Analysis of a LoRa Network under Imperfect Orthogonality”, IEEE Transactions on Industrial Informatics, Volume 15, Issue:3, March 2019. DOI. 10.1109/TII.2018.2864681.

[12] Alexandre Guitton, Megumi Kaneko, "Decoding Superposed LoRa Signals", The 43rd IEEE Conference on Local Computer Networks (IEEE LCN), October 1-4, 2018, Chicago, USA, DOI: 10.1109/LCN.2018.8638253.

[13] Taoufik Bouguera,Jean-François Diouris, Jean-Jacques Chaillout, Randa Jaouadi and Guillaume Andrieux, "Energy Consumption Model for Sensor Nodes Based on LoRa and LoRaWAN", Sensors 2018, 18, 2104; doi:10.3390/s18072104.

[14] Chun-hao liao , Guibing zhu, Daiki kuwabara, Makoto suzuki, and Hiroyuki Morikawa, "Multi-Hop LoRa Networks Enabled by Concurrent Transmission", IEEE Access, DOI: 10.1109/ACCESS.2017.2755858.

[15] Guillaume Ferré, Audrey Giremus. LoRa Physical Layer Principle and Performance Analysis. ICECS 2018 25th IEEE International Conference on Electronics Circuits and Systems, Dec 2018, Bordeaux, France. hal-01977497.

[16] Norbert Blenn and Fernando Kuipers, "LoRaWAN in the Wild: Measurements from The Things Network",Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands, arXiv:1706.03086v1 [cs.NI] 9 Jun 2017.

[17] Akram H. Jebril, Aduwati Sali, Alyani Ismail and Mohd Fadlee A. Rasid ,"Overcoming Limitations of LoRa Physical Layer in Image Transmission", Sensors 2018, 18, 3257; doi:10.3390/s18103257.

[18] Ferran Adelantado, Xavier Vilajosana, Pere Tuset-Peiro, Borja Martinez, Joan Melià-Seguí, Thomas Watteyne, "Understanding the Limits of LoRaWAN", IEEE Communication magazines 2017, arXiv:1607.08011v2 [cs.NI] 13 Feb 2017.

[19] Xuan-Chien Le, Baptiste Vrigneau, Matthieu Gautier, Malo Mabon, Olivier Berder. “Energy/Reliability Trade-off of LoRa Communications over Fading Channels”. International Conference on Telecommunication, Jun 2018, Saint-Malo, France. hal-01816574.

e-Learning System using Mashup based e-Learning Content Collection and an Attractive Avatar in OpenSimulator

Kohei Arai
Department of Information Science
Saga University, Saga City, Japan

***Abstract*—Mashup based e-learning content collection and provide system with an attractive avatar in OpenSimulator is proposed for making learning processes attractive and finding the students week points through learning processes with avatar. Through experiments with under graduated students, it is found that the proposed e-learning system is useful for the under graduated students. The students may find their week points through learning processes. Avatar provides the questions of the student's week subjects. The students can enjoy the communications with their avatar. Therefore, the students have the most appropriate e-learning content which provided by the mashup based information retrievals and have lessons with their own avatar attractively using the proposed e-learning system.**

***Keywords*—OpenSimulator; Mashup; e-learning; mobile learning; avatar; search engine**

I. INTRODUCTION

An e-learning system is an information system for carrying out e-learning, and is roughly composed of "learning materials / learning materials" and a "learning management system" (LMS). Note that the term "e-learning" alone, which does not include a system, may mean an information system for e-learning.

The users of the e-learning system are assumed to be "learners" and "teachers", and the functions for learners are different from the functions for teachers. Many e-learning systems have a "system administrator" (system administrator) of the e-learning system, and the system administrator may support learning activities and educational activities.

e-learning can be classified into two types, asynchronous type on-demand (self-learning type) and synchronous type on-demand (live type). For both types of e-learning system has some problems, (1) it is not easy to find an appropriate e-learning content matching to learners' characteristics, (2) most of e-learning contents are not so attractive so that it is hard to continue their learning processes.

E-learning system utilizing learners' characteristics recognized through learning process with OpenSimulator[1] is proposed [1]. Also, avatar utilized Q/A system of e-learning content designed with OpenSimulator is proposed and well reported [2].

Module based content adaptation of composite e-learning content for delivering to mobile learners is proposed [3]. Eye based human-computer interaction allowing phoning, reading e-book/e-comic/e-learning, Internet browsing and TV information extraction is also proposed [4]. Efficiency improvements of e-learning document search engine for mobile browser is proposed and validated the system efficiency [5] together with e-learning document search method with supplemental keywords derived from keywords in meta-tag and descriptions which are included in the header of the first search result [6].

Method for leaning efficiency improvements based on gaze location notifications on e-learning content screen display is proposed [7] while video searching optimization with supplemental semantic keyword for e-learning video searching [8]. On the other hand, e-learning system which allows students' confidence level evaluation with their voice when they answer to the questions during achievement tests is well reported [9].

E-learning system utilizing learners' characteristics recognized through learning process with Open Simulator is proposed and evaluated its effectiveness [10]. Question answering for collaborative learning with answer quality predictor is proposed and validated [11]. Meanwhile, lecture's e-table (server terminal) which allows monitoring the location at which each student is looking during lessons with e-learning contents through client terminals is proposed and reported its effectiveness [12].

Development of learning support software with CG animations for intellectually disabled children is conducted and validated its effectiveness [13]. On the other hand, question answering for collaborative learning with answer quality prediction is proposed [14]. Free Open Source Software (FOSS) based e-learning system together with blended learning system is introduced as a cost effective system [15]. Predicting quality of answer in collaborative question answer learning is also proposed [16].

Yahoo! Search and web API utilized mashup based e-learning content search engine for mobile learning is proposed and evaluated its effectiveness [17]. Meanwhile, e-experimental study of spatial recognition capability enhancement with building blocks learning contents for disable children is conducted and well reported [18].

[1] http://opensimulator.org/wiki/Main_Page

In order for making learning processes attractive and finding the students week points through learning processes, an e-learning system using mashup based e-learning content retrievals and avatar (in OpenSimulator) for attractive lessons is proposed here. Learners can continue their lessons attractively using these OpenSimulator utilized e-learning system. On the other hand, mashup based e-learning content retrievals is getting much popular because it is easy to retrieve attractively and effectively and efficiently. Related contents are also provided from the mashup based content retrievals for overcoming week points of which avatar is found through lessons. Therefore, learners may choose the most appropriate e-learning contents together with the related contents easily.

The following section describes research background of OpenSimulator and mashup technologies followed by the proposed e-learning system. Then some examples of the proposed system as well as experimental results are described. After that, conclusion is described together with some discussions and with future research works.

II. Research Background

The key components of the proposed e-learning system are OpenSimulator and Mashup content retrieval. Therefore, research background of the key components and software used are described in this section.

A. Mashup

API (Application Program Interface) is a set of instructions and functions that can be used when developing software.

For example, suppose you want to display characters in a window. If you have an API, you can use a function that displays characters in a window, but if you do not have an API, you can do it in dozens of lines (possibly hundreds, You may have to write a "function that displays characters in the window" yourself.

The mashup is to create a new service by combining technologies and contents from multiple different providers.

There are various technologies in the world, but each has its own strengths and weaknesses.

For example, yahoo's Yahoo map API displays a map based on latitude and longitude, and if you use the `` Geocoding API " that can search for latitude and longitude from Google's address, site that displays a map of the input address " Can be made.

In this way, creating something by combining multiple technologies (APIs) is called mashup.

B. OpenSimulator

With the OpenSimulator, it is capable to create a virtual environment (virtual world), it is possible to access from various clients and multiple protocols, build own environment using technologies of the right materials, including the construction of avatars, etc. It can be an application server. In addition, it can be visually displayed by using a dedicated viewer. Open Sim is the main OpenSimulator.

OpenSimulator is a 3D application server. It is possible to create a virtual environment (virtual world) and access it from various clients and from multiple protocols. With OpenSimulator, it is capable to build own environment using the right technology at the right moment. – It is designed the software so that it is possible to customize any configuration of the building through the load module. OpenSimulator has been released under the BSD licens[2]: Berkeley Software Distribution License, and it is open source but it is commercially available.

With no difficult settings, OpenSimulator can be used to simulate a virtual environment similar to Second Life[3] ™ (client is also compatible). Other environments, protocols, and functions can be used by installing add - on modules. A list of available modules is on the Forge site[4].

OpenSimulator is still alpha version software. Although OpenSimulator is relatively new software, it already has the following many useful functions:

1) It is capable to create multiple "worlds" in one application instance.

2) Multiple clients and protocols can be used - the same world can be accessed with multiple protocols at the same time.

3) It is possible to customize an avatar by custom made clothes, skins, wearing objects, and so on.

4) It can select from among several engines such as ODE (Open Dynamics Engine) [5], and perform physical simulation in real time.

5) Using in - world construction tools, it can create content in real - time in the environment. "What you see is what you get" (what you see is what you get).

6) It is capable to create in-world applications in many languages including LSL / OSSL[6], C #, VB.NET[7], etc.

It is an OpenSimulator provided by OpenSimulator. Although OpenSim[8] is still under development, it is becoming a 3D virtual world open source project that is drawing attention due to its high degree of completeness and its compatibility with Second Life[9] Clients. A function to display and create 3D objects such as buildings, a function to simulate a change in time, a function to edit an avatar's appearance and to manage inventory, a function to control avatar's appearance between inventory and multiple avatars (Such as chat based on chat and instant messenger, voice chat with VOIP (Voice over Internet Protocol)[10], execution of animation and gesture, exchange of possessions), script support function, and so on.

The current release is 0.9.1.0 (released on October 28, 2019) and can be downloaded as

opensim-0.9.1.0.tar.gz or

opensim-0.9.1.0.zip

[2] https://opensource.org/licenses/BSD-3-Clause
[3] https://secondlife.com/?lang=ja-JP
[4] https://secondlife.com/?lang=ja-JP
[5] http://demura.net/ode
[6] https://www.ossl.wiki/
[7] https://www.yumpu.com/en/document/view/43918605/opensim-scripting-languages-lsl-and-ossl-justinccs-/3
[8] http://opensim.stanford.edu/
[9] https://secondlife.com/?lang=ja-JP
[10] http://e-words.jp/w/VoIP.html

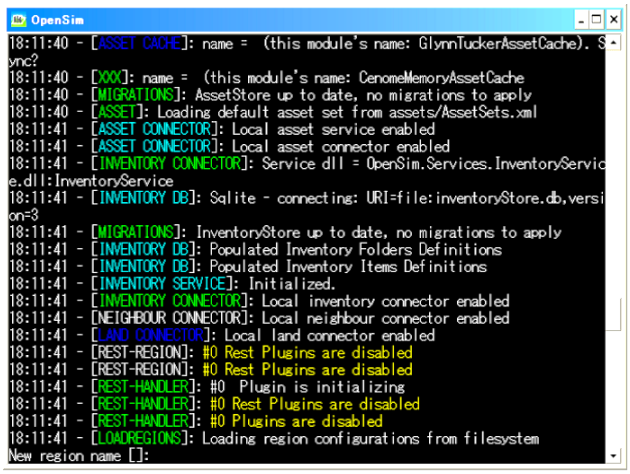


Fig. 1. OpenSimulator Description.

Fig. 2. Example Display is made by the Dedicated viewer of OpenSimulator.

Fig. 1 shows the OpenSimulator description. Meanwhile, an example of OpenSimulator environment is shown in Fig. 2. The example display is made by the dedicated viewer of OpenSimulator.

C. Software used

1) XAMPP: XAMPP is a package of free software necessary for running web applications, which is provided by apachefriends.org. Although it is mainly for development or learning, XAMPP is sometimes used as an actual operation environment on an intranet or the like. Apache (Web server), MySQL (SQL database server) and PHP which is the Web programming language and Perl[11] used for same purpose, management tool such as phpMyAdmi[12], and several auxiliary software such as SQLite Library modules are included. Originally, Linux was the only supported OS, it was called LAMPP[13] with its initial letter L, but since it corresponded to multiple OS later, changed L to X and became XAMPP.

2) PHP: In this system, PHP was used for development of the Web interface, cooperation with the database, and creation of the program. PHP is a script language that is used by Hypertext Preprocessor, which is one of extended functions of a Web server that dynamically generates Web pages. It has features such as excellent XML support and cooperation with various databases, being an interpreter type scripting language, and being grammar similar to C / C ++, Java, so it is easier to understand.

3) MySQL: Avatar e-learning MySQL was used as a database for storing data when learning content. MySQL is an open source relational database management system (RDBMS) developed by TCX DataKonsultAB[14] and others. It is possible to perform retrieval processing and update processing at high speed and keeping the speed inferior to that of commercial DBMS. MySQL has a feature that it has high affinity with the Web, and it has a reputation as a service on the Internet and a back end system of the EC site. In addition, it is compatible with many platforms such as Windows and various UNIX type OS, and interfaces for various programming languages are prepared.

4) phpMyAdmin: phpMyAdmin is free software written in PHP that can manage MySQL via the Web. It is possible to perform various operations such as database, table, field, index, authority management, etc. on MySQL database system without describing SQL statement. Also, the user can write arbitrary SQL statement and execute it.

III. Proposed e-Learning System

A. Design Concept

In this system, the avatar himself accumulates a large amount of data, extracts appropriate data to the learner from the data, and presents it. For that purpose, we developed and studied with the following algorithm. First, learners do learning. Save the result data in the avatar. (Save to Database) This allows the avatar to understand what kind of person the learner is. Next, the avatar collects the information (data) of the e-learning site and saves it. This is the knowledge of avatar. Then, using the learning history of the learner, data suitable for the learner is extracted and presented. The result data is newly stored and accumulated. The basic process flow is shown in Fig. 3.

B. Database Structure

In this research, it is necessary to accumulate data in the database as accumulation of avatar knowledge. The structure of the database was constructed as shown in Fig. 4.

Meanwhile, the detailed database structure and the contents of the database are shown in Fig. 5(a) and (b), respectively. On the other hand, an example of the form screens is shown in Fig. 6. Designed as a form screen on the web browser as shown above. Create a button with the value you are not good at, and select it. Once, learners had a lesson, then their week subjects is analyzed by the proposed e-learning system through avatar gathers their past lesson leant and achievement test scores.

[11] https://www.activestate.com/activeperl/downloads
[12] https://www.phpmyadmin.net/
[13] http://seesaawiki.jp/w/tororo66/d/LAMPP
[14] https://www.easycounter.com/report/tcx.se

Then, the week subjects are provided from the avatar as shown in Fig. 7.

The proposed e-learning system extract subjects which are not good from the correct answer rate of the learning history and present the problem of the subject to the learner. The learner inputs the answer in the text box and inputs the answer button. The response results are shown in Fig. 8. It says “Your problem is” “1+1=?” Then “Correct”.

Compares the input value with the answer, presents the result to the learner, calculates the correct answer rate including past results on this question, and presents it to the learner. Save the result as the learning history and return to the form screen. Perform the same operation as before. When a certain percentage of correct answers are exceeded, a new problem is presented. (In this research, we set the correct answer rate at 80% or more.) Again, do the same operation as before.

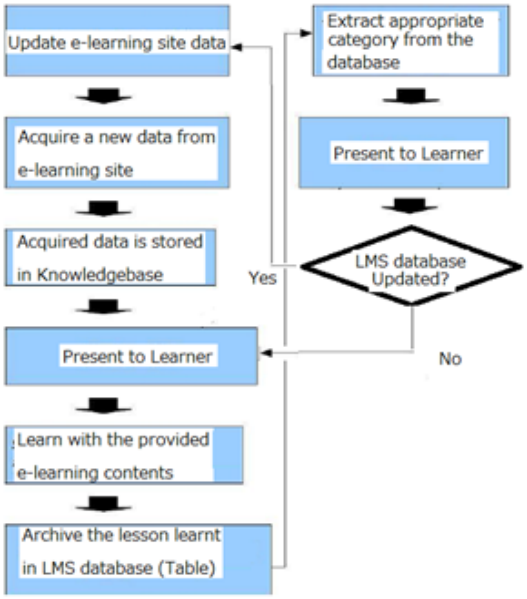


Fig. 3. Fundamental Process Flow of the Proposed e-Learning System.

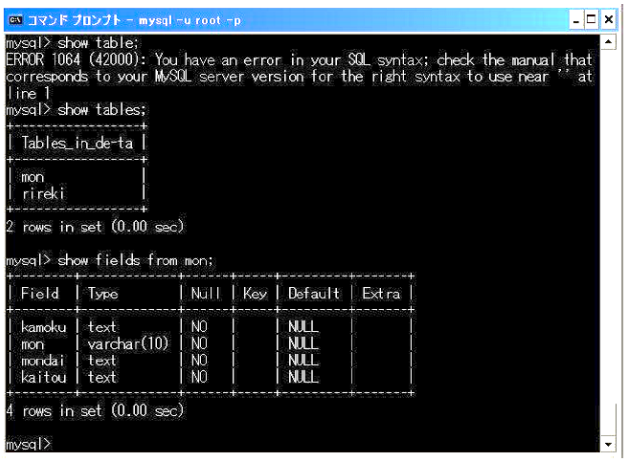
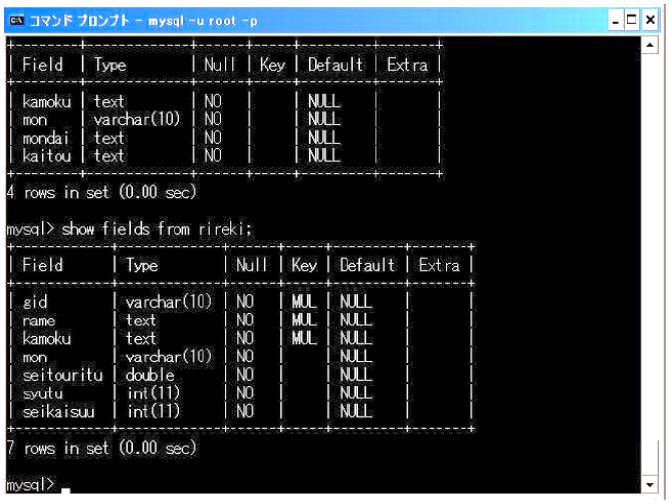
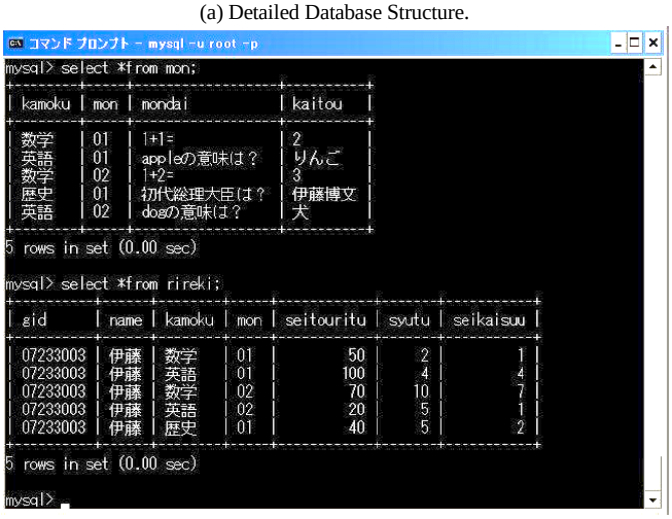


Fig. 4. Database Structure.



(a) Detailed Database Structure.



(b) Content.

Fig. 5. Contents in the Database.

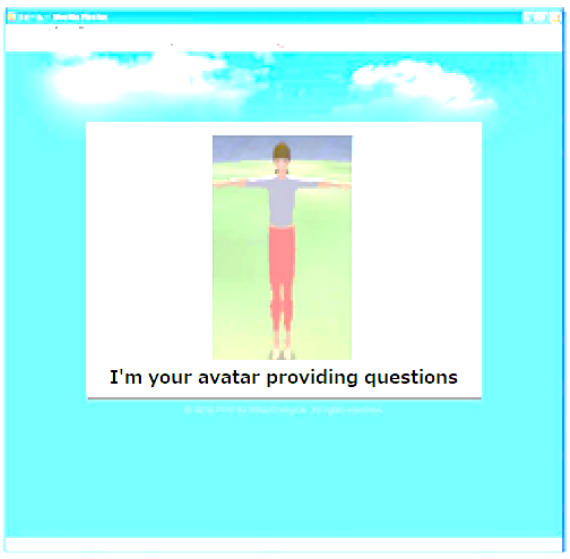


Fig. 6. Example of the form Screen.

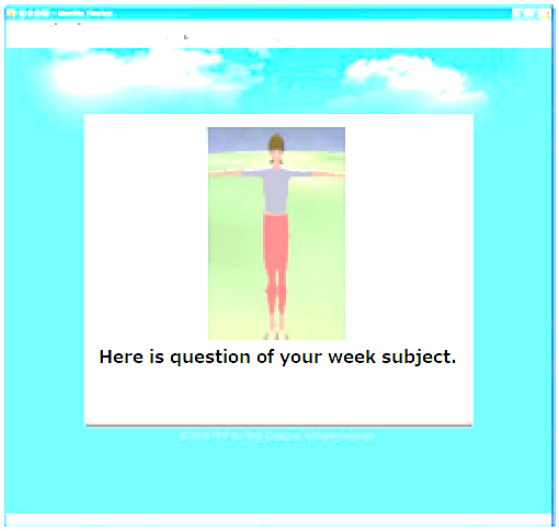


Fig. 7. Weak Subject Display.

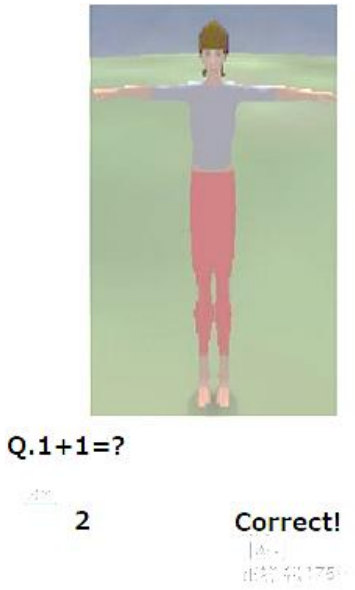


Fig. 8. Response Result.

If you answer incorrect values for the questions you have made, display as incorrect and display the correct answer. Furthermore, the correct answer rate including past results on this problem is calculated and presented to the learner as shown in Fig. 9. It says " The first Japanese prime minister?" "Incorrect Answer" "Correct answer is Hirofumi Itoh".

In this way, the proposed e-learning system save the learning history of the learner as data and present the problem suitable for the learner. Avatar accumulates learning history of learners, accumulates knowledge of avatars by acquiring data from e-learning site and accumulating data. And since avatar can extract and present mistakes learners caused in the past based on the data, learning can be done more efficiently than ever. In this time, although it was possible to extract data suitable for the learner, acquisition of data is incomplete.

Fig. 9. Correct Answer Rate Including past Results on this Problem is Calculated and Presented to the Learner.

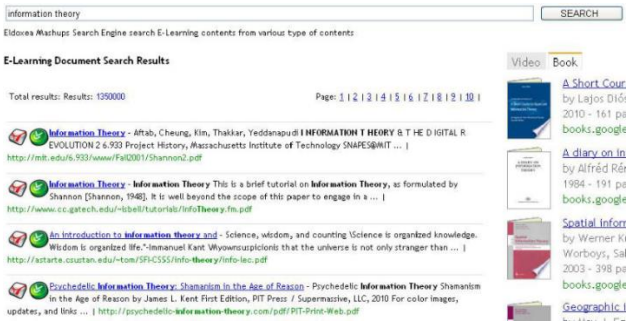


Fig. 10. Screenshot of an Example of e-Learning Content List Gathered with Mashup (Keyword = Information Theory).

C. Mashup Contents

A web mashup is a new type of web application that uses data and services from one or more external sources (usually from the Internet) to build entirely new and different web applications. Mashup web services is not a web portal. Combining Web service technologies with fresh content, collaborative approaches (such as Web 2.0[15] technologies), and possibly Web data management and semantic technologies (RSS, RDF[16]).

Mashup is a new technology based on Web 2.0 and web service technology. Mashup create a new web application with combining data from two or more web services. Therefore, a variety of e-learning contents can be provided for in particular over coming week subject by using mashup technology. Fig. 10 shows screenshot of an example of e-learning content list gathered with mashup (keyword = Information Theory).

IV. Experiment

The proposed e-learning system is used for 10 of the under graduated students in the information science department of Saga University. They learnt about mathematics (in particular, linear algebra), Physics (in particular Newtonian equation),

[15] https://en.wikipedia.org/wiki/Web_2.0
[16] https://ja.wikipedia.org/wiki/RSS

common social issues (in particular politics, economy) using the proposed e-learning system with their avatar for about 5 hours.

Their impressions are as follows:

1) My avatar tells my unknown week subject appropriately (the students may find their week points through learning processes),

2) Mashup e-learning contents are useful for overcoming week subjects (Avatar provides the questions of the student's week subjects)

3) The proposed e-learning system is much more attractive than the others because avatar is getting smarter through their learning processes (the students can enjoy the communications with their avatar).

V. Conclusion

Mashup based e-learning content collection and provide system with avatar in OpenSimulator is proposed. Through experiments with under graduated students, it is found that the proposed e-learning system is useful for the under graduated students. The students may find their week points through learning processes. Avatar provides the questions of the student's week subjects. The students can enjoy the communications with their avatar. Therefore, the students have the most appropriate e-learning content which provided by the mashup based information retrievals and have lessons with their own avatar attractively using the proposed e-learning system.

Further study is required for the followings,

1) Rethink the data acquisition algorithm again.
2) Implementation of data storage in the database.
3) Adapting this system to Open Simulator.

Acknowledgment

The author would like to tank Dr. Herman Tole of Brawijaya University, Indonesia and Mr. Kazuma Itoh of former student of the Saga University Japan for his effort to conduct the experiments.

References

[1] Kohei Arai, Anik Nur Handayani, E-learning system utilizing learners' characteristics recognized through learning process with Open Simulator, International Journal of Advanced Research in Artificial Intelligence, 2, 4, 8-12, 2013.

[2] Kohei Arai, A.N.Hardiyani, Avatar utilized Q/A system of e-learning content designed with Open Simulator, Proceedings of the International Conference on Convergence Content 2012, 149-150, 2012.

[3] 254. Kohei Arai and Tolle Herman, Module based content adaptation of composite e-learning content for delivering to mobile learners, International Journal of Computer Theory and Engineering, 3, 3, 382-387, 2011.

[4] Kohei Arai, Ronny Mardiyanto, Eye-based human-computer interaction allowing phoning, reading e-book/e-comic/e-learning, Internet browsing and TV information extraction, International Journal of Advanced Computer Science and Applications, 2, 12, 26-32, 2011.

[5] 295. Kohei Arai, Herman Tolle, Efficiency improvements of e-learning document search engine for mobile browser, International Journal of Research and Reviews on Computer Science, 2, 6, 1287-1291, 2011.

[6] Kohei Arai, T.Herman, E-learning document search method with supplemental keywords derived from keywords in meta-tag and descriptions which are included in the header of the first search result, International Journal of Advanced Computer Science and Applications, 3, 4, 99-104, 2012.

[7] Kohei Arai, Method for leaning efficiency improvements based on gaze location notifications on e-learning content screen display, International Journal of Advanced Research in Artificial Intelligence, 1, 3, 1-6, 2012.

[8] 319. Kohei Arai, T.Herman, Video searching optimization with supplemental semantic keyword for e-learning video searching, International Journal of Research and Review on Computer Science, 3, 3, 1640-1644, 2012.

[9] Kohei Arai, E-learning system which allows students' confidence level evaluation with their voice when they answer to the questions during achievement tests, International Journal of Advanced Computer Science and Applications, 3, 9, 80-84, 2012.

[10] Kohei Arai, Anik Nur Handayani, E-learning system utilizing learners' characteristics recognized through learning process with Open Simulator, International Journal of Advanced Research in Artificial Intelligence, 2, 4, 8-12, 2013.

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

[12] Kohei Arai, Lecture's e-Table (server terminal) which allows monitoring the location at which each student is looking during lessons with e-learning contents through client terminals, International Journal of Advanced Research in Artificial Intelligence, 2, 6, 40-45, 2013.

[13] Kotaro Taguchi, Mariko Oda, Ou Kouno Seio Oda, Kohei Arai, Development of learning support software with CG animations for intellectually disabled children, Journal of Education System and Information Society of Japan, 30,1, 48-56, 2014.

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

[15] Kohei Arai, Free Open Source Software: FOSS based e-learning system together with blended learning system, International Journal of Advanced Research in Artificial Intelligence, 2, 11, 9-16, 2013.

[16] Kohei Arai, Anik Nur Handayani, Predicting quality of answer in collaborative question answer learning, International Journal of Advanced Research in Artificial Intelligence, 3, 2, 23-26, 2014.

[17] Kohei Arai, Yahoo! Search and web API utilized mashup based e-learning content search engine for mobile learning, International Journal of Advanced Research on Artificial Intelligence, 4, 6, 1-7, 2015.

[18] Kohei Arai, Taiki Ishigaki, Mariko Oda, Experimental Study of Spatial recognition Capability Enhancement with Building Blocks Learning Contents for Disable Children, International Journal of Advanced Computer Science and Applications: IJACSA, 9, 6, 83-89, 2018.

Author's 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 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 37 books and published 570 journal papers. He received 30 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

The Impact of Social Networks on Students' Academic Achievement in Practical Programming Labs

Ayat Al Ahmad[1]
Department of Computer Science and its Application
Hashemite University, Zarqa, Jordan

Randa Obeidallah[2]
Department of Computer Information System
Hashemite University, Zarqa, Jordan

***Abstract*—Internet and Communication Technology (ICT) is being applied extensively in education to allow students to obtain information at anytime from anywhere. Social media is a new approach that is used to enhance and improve the delivery of education. Facebook groups and YouTube channels are the most operated networks that manage to deliver information. In this paper, we will study the impact of applying Facebook groups and YouTube videos on students' academic achievements in computer programming labs especially in object-oriented programming 2 (OOP2) lab at the Hashemite University. The practical programming lab plays an important role in understanding the theoretical programming concepts, for this reason programming lab is chosen as a case study. The proposed methodology embeds the social media networks as new major dimension in the teaching process of OOP2 Lab side by side with traditional lectures and e-learning tool such as Moodle. In this research: three surveys are utilized respectively to inspect the reasons of the weakness in OOP2 lab, evaluate the course learning outcomes (CLOs) by students before and after applying the proposed methodology and to investigate the opinions of students toward using social media networks within learning process. The results showed that operating the social media sites used by students on a daily basis creates a friendly and close educational environment as well as enhancing the academic results of students.**

Keywords*—*e-Learning; Facebook; YouTube; social network; programming lab; course learning outcome

I. Introduction

In traditional classrooms, many types of multimedia content such as images, audios, and videos are being used to deliver knowledge to students. The revolution of internet technology makes all media contents available via the internet and accessible at any time. Today, social media sites are popularly increased because they allow users to share collaborate and interact globally, in addition to access information directly and easily.

At present, the number of users who use Facebook and YouTube is dramatically growing. According to 2019 statistics published by biographon website[1], the total number of people who use YouTube are approximately 1,300,000,000 users. Regarding Facebook, the statistics of statista website[2] showed that Facebook has 2.41 billion monthly active users as of the second quarter of 2019.

Now-a-days, social media sites are being utilized in education. Studies showed that social media had improved the communication between instructors and students [1]. This helps to ease understanding and accessing information as fast as possible. As a result, studies showed that social media sites are effective in improving students' academic work [2]. However, it is recommended not to use social media during the lecture time [1]. On the other side, some studies showed that social media has a negative impact on students' academic performance because students spend more time in communication rather than studying [3]. The study of [4] revealed that students use social media to chat rather than to study. Time management is a major factor to determine a positive or negative impact of social media in education [5].

In Jordan, YouTube and Facebook websites are the most used social media sites. In this paper, we will study the impact of using such websites in computer programming practical lab courses. We will consider the course of object-oriented programming 2 (OOP2) lab in the Hashemite University (HU). OOP2 lab course teaches students to use JAVA as a programming language to understand object-oriented concepts.

It is important to mention that this research considers the following facts: Firstly, many HU students own phone mobiles connected to the internet. Secondly, most of the students have a Facebook account and finally, the YouTube website is accessible in the HU campus but not Facebook.

This paper is structured as follows: Section 2 introduces the related experiences that studied the impact of social media on the academic performance of student, Section 3 presents the object-oriented programming 2 lab course as a case study, Section 4 clarify the proposed teaching methodology, Section 5 discusses the results of applying suggested methodology and finally Section 6 concludes and summarize the study.

II. Literature Review

Studying computer programming demands gaining theoretical understanding and practical learning to develop programs. Many studies proposed approaches for learning programming using a computer laboratory context. Author in [6] presented a new approach for analyzing data, the authors suggested an approach that lights the interaction between learning by practice and learning of theory when students

[1]https://biographon.com/youtube-stats/[visited 13-9-2019].
[2]https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/). [visited 13-9-2019].

work programming within a laboratory. Author in [7] investigated the problematic relation between the learning of theory and learning of practice for novice programming student, the study discussed students' learning of practice, concepts, and specifically how these relate by utilizing phenomenography and variation theory. Author in [8] compared two approaches to producing structured, meaningful and useful descriptions of students' learning in labs.

There is no doubt that social media sites gain wider acceptability and usability in the whole world especially in third world countries that are controlled by variant values and restrictions. Our literature review focus on studying the impact of using social media sites on the academic performance of students in different countries that have miscellaneous cultures. This will be profitable for discovering the trends of diverse people toward embedding social media within learning.

The first selected research from Spain [9]. A sample of 1960 students who were studying one of two courses at the level of undergraduate and postgraduate level respectively. The study invested mixed qualitative (content analysis) and quantitative (descriptive analysis and ANOVA) design. 411 students of the sample carried out an activity based on social media participation and the remaining students carried out an activity based on traditional learning activity. The study showed that students who participated in social media-based activity had better performance than those who did not participate in the social media-based activity. Social media activity exerted a positive influence on the academic performance of students.

The second picked study from Saudi Arabia [10]. The article selected 60 Business Administration and management information students who are actively using social media. The study used quantitative as well as qualitative methods of research, the primary data acquired from 60 students. The variables of the study were: the frequency of using social media sites, the respondent's access to the internet and the usage and perception of social media. The paradigm of the study adopted Shikawa model. The results presented the dual impact of social media on the performance of students and it is necessary to approach adolescents' use of the social network with ultimate responsibility. The recommendation of the study was the institutions should focus on useful-prompting of social networks as a tool not only for communication and entertainment but also for learning.

From Pakistan [11] investigated the relationship between positive and negative characteristics of social media and the learning attitude of university students for sustainable education. The social Gratification theory was applied to examine students' behavior during social media usage. The study acclimatized the cluster sampling method and it identified 18 adversarial and constructive factors of social media from previous literatures. The positive factors were: creates awareness, lecture sharing, easy communicating, helps to maintain contacts, reduce the cost of purchasing books, improves confidence, improves social and communication skills, increases knowledge and reduce stress. The negative factors were lack of critical thinking, waste of time, disrupting writing skills, leads to break up on study connectivity, increase cyber-bullying, creates laziness in students, and creates depression, problematic communications and health hazards in student's life. The researchers distributed 1013 questionnaires among bachelor's and master's degree students from five regions. The finding revealed that social media in Pakistan has a negative influence on a students' behavior as compared to positive aspects.

In Turkey, [12] studied the effectiveness of social learning networks. The influence of the Edmodo was investigated. The independent variables of the study were: cumulative grade point average, type of school, gender and the frequency of use of the internet and social networks. The dependent variables of the study were: achievement score and attitude towards online learning. The sample of study composed of 79 learners taking the courses of special methods in the department of computer education and instructional technologies in the education faculty. In addition to face to face lectures, the proposed method mixed group works, individual assignments and discussions under the guidance of teacher via Edmodo. The results presented the students who used Edmodo were successful than those who did not, Edmodo had positive contributions to learners' academic achievements.

The fifth research is a case study from India. Author in [13] investigated the academic use of social media applications by university students. The survey method was employed to investigate the sample comprised of 482 students. The results indicated that students have little exposure to use social media in their academic life. The recommendation of the study was social media needs to be given a new lease of life by the educators through separate curricula thus avoiding the fear.

The sixth study is from Malaysia [14]. The authors achieved a pilot study of undergraduate and postgraduate students at the university of Tekologi Malaysia (UTM). The variables observed in the study were: interactive with peers, interact with teachers, engagement, perceived ease of use and perceived usefulness, influence students' satisfaction, and academic performance of students. The research used the survey method, the prepared survey had distributed on 120 students. The analysis of data performed using SPSS application. The results showed that social media affects positivity and significantly collaborated learning with interaction with peers, interaction with supervisor, engagement, perceived ease of use, and perceived usefulness.

The seventh study is from Nigeria [15]. They surveyed students' social media networking sites usage and how it affects them. 539 students were completely filled the questioner. The hypothesis of the study is the frequent use of social networking sites by students has no effect on their studies. The collected data analyzed using frequencies; percentage and graph representation while the hypothesis was tested using chi-square. The study revealed no effect on the student who uses social media sites on their studies. The recommendation of the study to learners is to give importance to time management skills during using social media sites.

III. Object Oriented Programming 2 Lab Course Case Study

Object-Oriented Programming 2 (OOP2) Lab is a mandatory course for all information technology students who enrolled in the Hashemite University. The lab is a synchronized course with the OOP2 course. In this course, object-oriented programming concepts are introduced using JAVA programming language, including inheritance, polymorphism, abstract classes, interfaces, API classes, event handling techniques, and exception handling concepts to handle run-time errors.

The OOP2 lab course is taught in every academic semester with three hours per week. Sections are distributed either Sunday, Tuesday and Thursday or Monday and Wednesday. The average number of students who register in the OOP2 lab during each semester is approximately 100 students.

A lab manual is prepared carefully by the professors of the OOP2 theoretical course and OOP2 lab instructors to accomplish course requirements. The OOP2 lab manual is composed of sixteen assignments and six appendices. Assignments must be completed by the students using the Net Beans IDE. Based on the guidance of the lab instructor, some assignments must be solved individually and others must be solved as workgroups. The appendices present general instructions that may be required by students during semesters such as read a file and write a file in java, syllabus, the hierarchy of GUI classes in java, etc. OOP2 lab the manual is continuously updated.

As a student register to the OOP2 lab, he/she had the ability to access the course account in Moodle. Moodle is enabled for all OOP2 lab students. Lab material, syllabus, announcements, and additional material are uploaded on the Moodle by the instructor. Students are encouraged to download lab material at anytime from anywhere.

In most semesters, it was noticeable that the academic results of students in OOP2 lab are weak and the performance of students is less efficient than the performance in other labs. Moreover, the direct assessment of students for the OOP2 lab course indicated the failure of achieving the objectives of course satisfactorily for an academic institution.

A broad discussion and analysis of results occurred between OOP2 lab course instructors to enclose the reasons for the imbalance of the academic results of students. The suggested seven reasons were concluded:

1) Poor understanding of the theoretical concepts of object-oriented programming, for example, the concepts of classes, objects, and inheritance.

2) Duration of the lecture is short and insufficient to analyze and solve the given problem.

3) Shortage of group work during lab lecture.

4) General weakness in the practical side of programming languages(labs).

5) The difficulty of Assignments.

6) Computer Labs environment is uncomfortable.

7) Finally, Java is considered a difficult programming language.

For our preliminary study, we asked fifty students who studied the OOP2 lab course in different semesters to answer the following question: “What is the main difficulty that faced you during studying object-oriented programming 2 labs?” Students must nominate one of the seven trapped reasons mentioned previously. The responses of answers to the questions are as shown in Table I.

Answers of students elucidate that the most three difficulties that faced students during lab are: Duration of the lecture is short and insufficient, Lack of group work during lab lecture and Poor understanding of the theoretical concepts of object-oriented programming, for example, the concepts of classes, objects, and inheritance. The opinions of students will help us to suggest a new teaching methodology for the OOP2 lab that will be demonstrated in the next section.

TABLE. I. Responses to Preliminary Study

ANSWER CHOICES	RESPONSES
1-Poor understanding of the theoretical concepts of object-oriented such as classes, object, inheritance...etc.	18.00% 9
2-Duration of the lecture is short and insufficient to analyze and solve the problem.	26.00% 13
3-Lack of group work during lab lecture.	20.00% 10
4-General weakness in the practical side of programming languages(Labs).	14.00% 7
5- Difficulty of Assignments.	12.00% 6
6-Computer Labs environment is uncomfortable.	10.00% 5
7-Java is a difficult Language.	0.00% 0
TOTAL	50

IV. Social Media Methodology for Teaching OOP2 Lab

In order to enhance student’s results in the OOP2 lab, a new approach and methodology are suggested to support the lab. A Facebook group for OOP 2 lab is created, lectures within labs are recorded as detailed videos during the practical real-life lectures when the lecturer and students analyze, design and solve the assignments.

Students within the OOP2 Facebook group can communicate with each other, discuss some complex programming exercise together, discuss some solutions and concepts that were examined within lectures, review some assignments before practical exams and criticize the level of exams and other phenomena occurs during lectures. The Facebook group helps in breaking the ice between lecturers and students and consolidates the cooperation between students.

The recorded lessons review briefly the important object-oriented theoretical concepts demanded by students to complete the problem. The videos contain useful practical experiences such as the most common syntax and logic errors carry out during writing code and how to deal with and correct these errors. The videos also contain experiences about reading, analyzing problems and how to deal with the

diagrams and components of the unified modeling language (UML). Several shortcuts within NetBeans IDE editor are explained within videos to speed up solving the given problems.

All videos are uploaded on the YouTube channel and on the HU Moodle. All YouTube lessons are easily inserted into Moodle to create an integrated e-Learning social media community for students. Fig. 1 represents the proposed e-learning social media community.

The availability of videos makes the students feel comfortable and creates the desire to review lessons again. OOP2 lab videos are recorded, edited and regenerated every semester using the Camtasia tool. Currently, the number of uploaded videos is 17. The average length for all videos is 36.61 minutes and the average number of views for all videos is 898. Fig. 2 represents the YouTube channel and sample of lessons.

Although Moodle enables the academic staff to upload videos, our proposed teaching strategy exploited YouTube for a major reason. The major reason is the Moodle system contains a set of limitations, including that the students who can watch the videos are only the students who registered the lab. The general aim of study is to improve the level of all students whether they are registered the lab, registered the theoretical course of lab, graduated students or any students in the world.

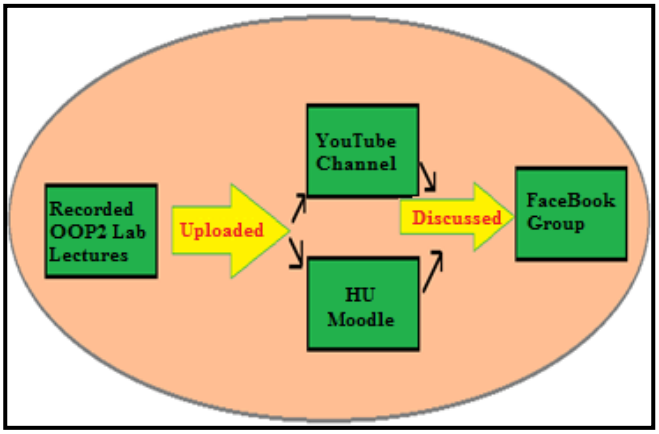


Fig. 1. Integrated e-Learning Social Media Community.

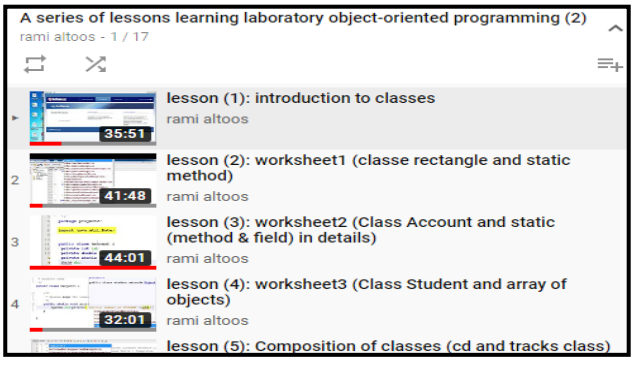


Fig. 2. YouTube Channel and the Sample of Recorded Lessons.

V. Results and Discussion

Our study focuses on measuring student's perception of their learning and how this learning is valued by different constituencies, at the same time the study examines the student's attitude toward using the YouTube channel and Facebook group in teaching OOP2 lab. In this research, a survey methodology is exploited to measure OOP2 lab course learning outcomes. Table II represents the OOP2 lab course learning outcomes (CLOs). A computerized tool is designed by the Hashemite University and it is used at the end of every course studied in information technology faculty. The tool measures the curriculum's course learning outcomes (CLOs) using a five-point Likert scale (Strongly Agree(SA), Agree(AG), Neutral(N), Disagree(DA), Strongly Disagree(SDA)). At the end of the semester and before one week of the final exam, the survey invitation is delivered to all students through the student's portal to answer the survey.

The CLO value is calculated as the following equation (1).

$$((count(SA)*5 + count(AG)*4 + count(N)*3 + count(DA)*2 + count(SDA)*1) / count(class)) / 5 \quad (1)$$

Where,

count(SA) :- number of students who answered strongly Agree.

count(AG):- number of students who answered agree.

count(N):- number of students who answered neutral.

count(DA):- number of students who answered disagree.

count(SDA):- number of students who answered strongly disagree.

count(class):- number of students within the class.

TABLE. II. CLOs of OOP2 Lab

	CLO Description	SA	AG	N	DA	SDA
CLO1	Apply object oriented programming concepts including classes, objects, inheritance, polymorphism, abstract classes, and interfaces in designing Java Applications.					
CLO2	Design Graphical User Interface using Java API classes					
CLO3	Create interactive application using event handling techniques					
CLO 4	Apply exception handling concept to handle run-time errors					

In this paper, we will compare the results of the survey before using the Facebook group and YouTube in the second semester of the academic year 2017 and after applying social media teaching approach in the academic year 2018/ first semester.

The results illustrated a remarkable increase in the percentage of achieving each objective of the course and also increasing the overall average of fulfillment for all objectives of the course. Table III represents the results before applying the social media methodology of teaching in the academic year 2017/second semester and the results of the academic year 2018/first semester after applying the social media approach in teaching. 10.49% is the percentage of enhancement.

In addition to the CLOs survey, we distributed simple survey to 50 respondents students to answer questions that are related to the new teaching approach. The simple questionnaire has been prepared to investigate whether students benefited from using YouTube Channel and Facebook groups. The questionnaire had three questions concerns about student's experience with the YouTube channel and Facebook group. The survey exploited the two-point Likert scale (YES-NO). The survey was conducted at the end of the semester after the students examine social media teaching methodology YouTube channel and Facebook group.

The results showed that 90.20% of the students agreed that OOP2 lab Facebook group contributes sufficiently to communicate with other colleagues and with the instructor. Fig. 3 represents the percentage of students who agreed that OOP2 lab Facebook group contributes sufficiently to communicate with other colleagues and with the instructor. As a result of this interaction and communication with others, students' academic results are enhanced and students are getting excited to ask, interact, and collaborate comfortably in a Facebook group rather than face to face communication.

In addition, students are willing to evaluate the teaching technique used in the classes, review the assignments questions, and discuss the level of exam, etc. 72.22% of the students agree that the Facebook group reinforces and supports criticizing the teaching technique and any other issue related to the OOP2 lab. Fig. 4 depicts the percentage of students who agree that the Facebook group support criticizing and reinforces the teaching technique and any other issue related to the OOP2 lab in the Facebook group.

TABLE. III. THE RESULTS BEFORE APPLYING THE SOCIAL MEDIA METHODOLOGY OF TEACHING IN THE ACADEMIC YEAR 2017/SECOND SEMESTER AND THE RESULTS OF THE ACADEMIC YEAR 2018/FIRST SEMESTER AFTER APPLYING THE SOCIAL MEDIA APPROACH IN TEACHING

OOP2 CLOs	year 2017 Second Semester	year 2018 First Semester
CLO1	**66.40%**	**86.15%**
CLO 2	**73.32%**	**85.13%**
CLO 3	**74.15%**	**79.66%**
CLO 4	**64.33%**	**69.28%**
Average	**69.55%**	**80.05%**

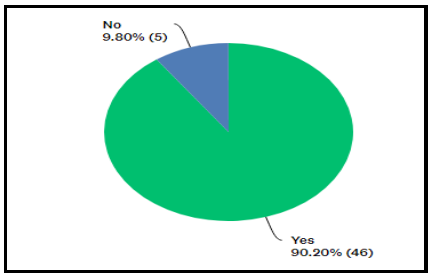


Fig. 3. The Percentage of Students Who Agree that OOP2 Lab Facebook Group Contributes Sufficiently to Communicate with other Colleagues and with the Instructor.

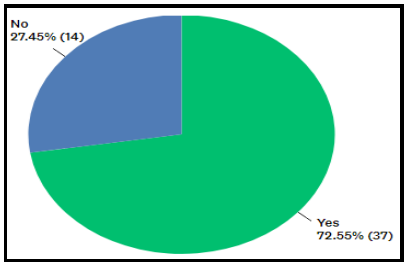


Fig. 4. The Percentage of Students Who Concur that Facebook Group Supports Criticizing the Teaching Technique and Any other Issue Related to the OOP2 Lab.

Regarding YouTube lessons, 86.27% of the students showed that the recorded videos contribute to enhance the understanding of object-oriented concepts. Fig. 5 depicts the percentage of students who showed that the recorded video contributes to enhancing the understanding of object-oriented concepts. This can be explained due to the use of the Arabic language "the native language of the students" as the explanation language of the recorded videos.

Using the Arabic language affects positively on exceeding the barrier of understanding foreign languages. YouTube can be considered as a convenient tool for mobile learning and since the wide segment of students at HU owns a mobile device, the YouTube channel is assumed as an ideal choice to develop the performance of students at the OOP2 lab. The videos are available at any time and students can view, rewind and replay the records many times as they need. The explanation language, permanent availability of videos and the ease of integrating records within the HU e-learning environment had improved the skills of taking-note, analyzing problems and focusing on the key points.

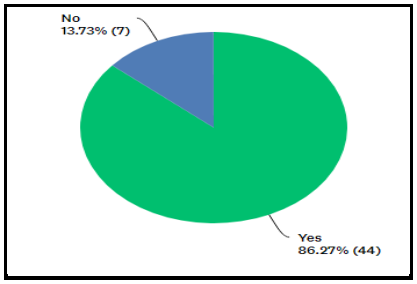


Fig. 5. The Percentage of Students Who Showed that the YouTube Recorded Video Contributes to Enhancing the understanding of Object-Oriented Concepts.

As a normal consequence of all improvements that have been performed to apply the proposed teaching methodology, the difficulty level of the exam has been increased. The difficulty level of exams, quizzes, and projects during the first semester of 2018 is increased to double compare with the exams, quizzes, and projects in the second semester of 2017. Though the direct assessment of students which measures directly the academic performance of students witness an improvement. The percentage of failure students in the second semester of 2018 dropped 0.8% compared with the percentage of failure during the second semester of 2017.

VI. Conclusion

Nowadays, social media networks are considered as another learning channel for a lot of students. Our study exploited the features included within Facebook and YouTube to enhance the academic performance of practical Labs. Our study invests YouTube channel to share video resources with students and Facebook is invested as a platform to express opinions and exchange ideas among students. The social media teaching methodology proves its success in improving the academic performance of students. In the future, our methodology will be extended to include more social media networks such as snap chat and Instagram to study their impact on the academic performance of students.

References

[1] Abousaber, I., Oueder, M. (2018). A Study on the Impact of Social Media Usage on Student Academic Performance: University of Tabuk an Example. American Scientific Research Journal for Engineering, Technology, and Sciences (ASJETS), 40(1), pp. 77-88.

[2] Amin, Z., Mansoor, A,. Hussain, S., and Hashmat,F., (2016).Impact of Social Media of Student's Academic Performance.International Journal Business and Management Invention, 5(4), pp. 22-29.

[3] Asemah, S., Okpanachi, R. (2013). Influence of social media on the academic performance of the undergraduate students of Kogi State University, Anyigba, Nigeria, Research on Humanities and Social Sciences, 3(12), pp. 90-96.

[4] Owusu-Acheaw, M., Larson, A. (2015). Use of social media and its impact on academic performance of tertiary institution students: A study of students of Koforidua Polytechnic, Ghana, Journal of Education and Practice, 6(6), pp. 94-101.

[5] Mensah, S., Nizam I. (2016). The impact of social media on students' academic performance – a case of Malaysia Tertiary Institution. International Journal of Education, Learning and Training, 1(1), pp. 14-21.

[6] Thuné, M., & Eckerdal, A. (2018). Analysis of Students' learning of computer programming in a computer laboratory context. European Journal of Engineering Education, 1-18.

[7] Eckerdal, A. (2015). Relating theory and practice in laboratory work: A variation theoretical study. Studies in Higher Education, *40*(5), 867-880.

[8] Bernhard, J., & Carstensen, A. K. (2015). Analysing and modelling engineering students' learning in the laboratory: a comparison of two methodologies. In The 6th Research in Engineering Education Symposium (REES 2015), July 13-15, 2015, Dublin, Ireland (pp. 620-628).

[9] Santoveña Casal, S. (2019). The Impact of Social Media Participation on Academic Performance in Undergraduate and Postgraduate Students. International Review of Research in Open and Distributed Learning, 20(1).

[10] Talaue, G. M., AlSaad, A., AlRushaidan, N., AlHugail, A., & AlFahhad, S.(2018).The Impact of Social Media on Academic Performance of Selected College Students. International Journal of Advanced Information Technology (IJAIT) Vol. 8, No.4/5. R.

[11] Abbas, J., Aman, J., Nurunnabi, M., & Bano, S. (2019). The Impact of Social Media on Learning Behavior for Sustainable Education: Evidence of Students from Selected Universities in Pakistan. Sustainability, 11(6), 1683.

[12] Durak, G., Cankaya, S., Yunkul, E., & Ozturk, G. (2017). The Effects of A Social Learning Network on Students'performances and Attitudes. European Journal of Education Studies.

[13] Subramani, R. (2015). The academic usage of social networking sites by the university students of Tamil Nadu. Online Journal of Communication and Media Technologies, 5(3), 162.

[14] Al-rahmi, W. M., Othman, M. S., & Musa, M. A. (2014). The improvement of students' academic performance by using social media through collaborative learning in Malaysian higher education. Asian Social Science, 10(8), 210.

[15] Camilia, O. N., Ibrahim, S. D., & Dalhatu, B. L. (2013). The effect of social networking sites usage on the studies of Nigerian students. The International Journal of Engineering and Science, 2(7), 39-46.

Performance Evaluation of Different Short Path Algorithms to Improve Oil-Gas Pipelines

Nabeel Naeem Almaalei[1], Siti Noor Asyikin Mohd Razali[2]
Department of Mathematics and Statistics
Faculty of Applied Sciences and Technology
University Tun Hussein Onn Malaysia
Pagoh Education Hub
84600 Pagoh, Johor

***Abstract*—The oil-gas pipeline is a complicated and expensive system in terms of construction, control, materials, monitoring, and maintenance which includes economic, social and environmental hazards. As a case study of Iraq, the system of pipelines is above the ground and is liable to disasters that may produce an environmental tragedy as well as the loss of life and money. Hence, this article presents a performance evaluation of different short path algorithms to improve oil-gas pipelines. The chosen algorithms in this paper were Parallel Short Path Algorithm (PSPA), Ant Colony Optimization (ACO) algorithm and Genetic Algorithm (GA). The main performance metric is the cost of the pipelines. Simulation trials were performed using the MATLAB program for the chosen algorithms. The performance comparison showed that the lowest cost of laying oil and gas pipelines was by applying the GA algorithm when the number of wells was set to 50-600. Conversely, the PSPA algorithm showed the best performance in terms of required implementation time for all scenarios. Besides, PSPA appeared to have acceptable performance in terms of the cost of the pipeline when the number of wells was arranged between50-600. Furthermore, PSPA showed the best performance for 700 and 840 wells in terms of the cost of laying the oil and gas pipelines compared to ACO and GA. It should be noted that the ACO algorithm showed medium performance in terms of the cost of laying oil and gas pipelines compared to PSPA and GA.**

***Keywords*—PSPA; ACO; GA; Oil-Gas pipeline; performance; cost; short path**

I. INTRODUCTION

Oil is a crucial source of energy. It is used in many industries, transportations and electricity supplies. Oil is transferred from the oil fields to the main stations (oil refineries) that require a network of pipes, which include valves and pipes of different diameters and pumping stations. When oil passes through the pipelines, there occurs a loss of 3% of the total oil due to the evaporation process since the oil pipelines carry large quantities of oil. The oil pipelines are used to transport (import and export) petroleum products between different cities around the world with about 17.95 million barrels per day [1]. This has led to the development of large and extensive design as well as operations of pipelines which have become more complex in recent years [2]. The main issue discussed in this proposal is in applying the proposed algorithm to find the shortest path that links the oil wells with the gathering facilities, hence obtaining the least cost for the work of the oil-gas pipeline network.

The optimization network of pipelines for gathering and transporting oil and gas should include the location, parameter, and topology. The topology structure mentioned here is the multilevel star (MS), multistage star-ring (MSR) and the multistage star-tree (MST). In the past few years, some optimization algorithms have been used to improve the oil-gas pipeline network such as the ant colony optimization (ACO) algorithm, the particle swarm (PS) algorithm, artificial neural network (ANN) algorithm, and the Genetic algorithm (GA). Many researchers studied the improvement of pipeline networks since the early 1970s. For example, Gabriele (1977) [3] proposed the Steiner algorithm, whereas Edgar (1978) used a generalized reduced gradient method for the first time. Also, Simposon et al. (1990) [4] provided a GA to the optimization design, while Shuwen (1998) proposed the Kruskal algorithm as well as the Prime algorithm. On the other hand, in prior work [5], the authors used a genetic algorithm to supply the natural gas distribution network, while in the same year, *Jiancheng* in [6], who was the first researcher to use GA in the oil-gas pipeline network, used ANN to determine the multilevel star network. Furthermore, *Jianjun* combined the Simulated Annealing algorithm (SAN) and GA to solve the star network topology. In a previous work [7], the paper presented an optimization algorithm fitted for the oil-gas gathering network which considered several obstacles.

Fig. 1 shows the map of numerous oil fields in Iraq. Many oil fields are scattered around southern and northern Iraq, where the city of Basra is one of the largest cities in Iraq containing oil fields. It consists of 530 geological formations which indicate the existence of a large quantity of oil. One hundred and fifteen oil fields were drilled, including 27 fields proven to contain a large oil reserve estimated at 65 billion barrels. Fifty-nine percent of the total oil reserves were distributed to other fields including the North Rumaila, South Rumaila, Majnoon, Zubair, River Omar, West Qurna, Al-Tuba, Al-Sabah, Al-Halafiyya, and Abu-Gharb and Al-Bazarkan fields [9].

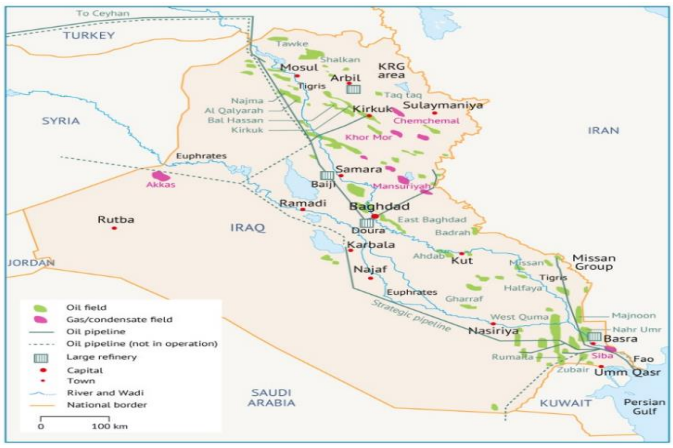


Fig. 1. Map of Numerous oil Fields in Iraq. [8].

II. Literature Review

A. Oil-Gas Pipeline Network

The transportation of solid, liquid and gaseous substances is crucial and necessary for humans. The ancient Chinese began this by using the stems of Bamboo plants to transport liquids and water up to three inches in diameter, using sloping passages to transport liquids. The idea of pipelines came from here.

Pipelines are a reliable and safe way to transport products necessary for the national economy of many countries in the world such as water, oil, and gas [10]. Dandy et al. in [11] applied the genetic algorithm and developed a technique to improve the water pipeline networks to find the minimum cost of construction. Concerning oil and gas pipeline networks, they transported and distributed large quantities of oil and gas worldwide. It was considered the safest way to transport because of its limited number of failures. In 1879, in Pennsylvania, the United States built the first oil pipeline with a length of 109 miles and a diameter of 6 inches [12]. Now, there are more than 60 countries with pipeline networks, with a length of more than 2000 km. The United States has the longest pipeline network in the world followed by Russia [13]. Many researchers had discussed the improvement of oil and gas pipeline networks, specifically in the 1970s. Some optimization algorithms such as Prime and Kruskal were proposed by Shuwen in [14] to improve the natural gas networks. The Prime algorithm is a greedy algorithm that finds the minimum extension tree for an unweighted graph. This implies that it finds a subset of the edges that make up a tree including each top of the head every time and adds the least possible connection from the tree to another's top. This algorithm was developed by the computer scientist Edsger W. Dijkstra in 1959 [15]. Therefore, it is also sometimes known as the Dijkstra algorithm. The Kruskal algorithm was also used to solve the same problem.

In 1996, the genetic algorithm technique was developed to improve pipeline networks, and its formulation was continuously improved to solve the problem of spending in New York. The genetic algorithm was considered better than traditional methods of improvement such as dynamic programming methods as well as linear and nonlinear programming [11]. This technique was powerful and was able to create the minimum cost of pipeline network constructions compared to the size of the search area. It was used to improve the water pipeline networks, taking into account the length and size of the pipe in the design to evaluate the target function [4].

Multi-objective optimization methods for the transportation of gas differ fundamentally from individual target improvement methods. On this basis, the technique of improving the ant colony of multipurpose was developed to improve the transportation of gas in pipelines by reducing the consumption of the compressor fuel as well as increasing the production to maximum [16]. The design of the gas pipeline networks was complex where it consisted of pipes and element controls such as pumps and valves. Control of the gas pipeline network required a suitable environment for all elements without violating the physical and operational limitations during the process of transferring gas from the moment the network enters to the moment it exits. In 2011, Ermin et al. [17] presented the article provided a procedure to calculate the appropriate network accessories and optimal cost based on rigorous mathematical programming methods. In 1978, Edgar [18] presented a study of a calculation algorithm to improve the design of the gas pipeline networks by determining the number of compressors, the length and diameter of the pipes, as well as the factors affecting the operation of the storage compressors to reduce the operating costs and capital using two solution techniques. The first method was applied to cases where the capital costs came from a non-zero initial fixed cost plus some horsepower output function, whereas the second method was known as the low gradient method, which is a nonlinear programming algorithm used directly in cases where the capital cost of compressors is a function of the horsepower output but does not contain any initial fixed cost [18].

Marine pipelines transporting oil and gas are exposed to erosion and degradation. It is, therefore, necessary to predict and monitor those pipelines to minimize accidents that may occur as well as optimize the operation. Models for the forecasting and evaluation of the state of oil and gas pipelines were developed using artificial neural network technology (ANN) based on the data provided for three oil and gas pipelines in Qatar. These models were able to predict the state of pipelines by a success rate of 97% [13].

Crude oil with high viscosity, freezing point, and wax content is usually transported through hot pipelines, known as the hot oil pipeline (HOP), equipped with pumping and heating stations. The cost of fuel and energy for pumping stations and heating ranges from 1-3% of total energy consumption. On this basis, a model was designed to reduce the cost of the operating power of HOP with the temperature of the output of each heating station and the operation of each pump as variables for improvement. The differential evolution (DE) algorithm was combined with the particle swarm optimization (PSO) algorithm to solve this model, where the optimum operating model saved 17.95% of the cost of energy needed to transport crude oil to a distance of 2640 m3/h [5].

B. Oil-Gas Parallel Shortest Path Algorithm

The proposed algorithm in [19] was based on finding the distances between the wells by giving the site (Cartesian coordinates) of each well and then choosing the shortest path that connects the two wells. We take into consideration the obstacles that may be located on the track which leads to finding an alternative path and finally reaching the shortest path connecting all the wells with the main station.

Parallel Shortest Path Algorithm (PSPA)

Inputs : WLT (X, Y) ∈ $R^{2\times n}$// *Location Table for all wells*
Output: $ShortPathTable$
Begin:
Calculate $D_{Table} \in R^{n\times n}$
For $i = 1{:}\,n$ **Do** // i=1, 2,n ;
For $j = 1{:}\,n$ **Do**
If **Obstacles**(i,j) == 1 **THEN**
$D(i,j) = \infty$
ELSE
D(i,j)= $SQR\left[(X_i - X_j)^2 + (y_i - y_j)^2\right]$
End IF
Next
Next
For $Well.Id = 2{:}\,n$ **Do**
Determine *Short path for Well (*$Well.Id$*).*
SET $X \leftarrow Well.Id$
S_ID = $ShortPath(Well.Id, \{D\})$, S_ID $\in \mathbf{R^{1\times 1}}$,
SET Well($Well.Id$).Links= S_ID
SET $D(Well.Id,$ S_ID$) = \infty$
SET $Y \leftarrow$ S_ID
Z = $ShortPath(Y, \{D\})$
SET $D(X, Z) = \infty, D(Z, X) = \infty$
$ShortPathTable(Well.Id)$ = Well($Well.Id$).Links
NEXT
END Algorithm

C. Formulation of ACO Algorithm in a Mathematical Model

Fig. 2 shows the ACO algorithm flowchart. We have the following mathematical model as given by:

$$p_{ij}^{k} = \frac{[\tau_{ij}]^{\alpha}[\zeta_{ij}]^{\beta}}{\sum_{z\in allowed\ y}[\tau_{ij}]^{\alpha}[\zeta_{ij}]^{\beta}}$$

where

k: the ant.

p_{ij}^{k}: The ant moving from i to j.

τ_{ij}: The amount of pheromone along the transition from node i to j.

$\propto \geq 0$: The parameter to control the influence of τ_{ij}.

ζ_{ij}: The desirability of node transition ij (a priori knowledge, typically $1/d_{ij}$, where d is the distance).

$\beta \geq 1$: is a parameter to control the influence of ζ_{ij}.

τ_{ij}, ζ_{ij}: represents the attractiveness and trail level for the other possible node transitions.

- Pheromone update

When all the ants have completed a solution, the trails are updated by [20]:

$$\tau_{ij} \leftarrow (1-\rho)\,\tau_{ij} + \sum_{k} \Delta\tau_{ij}^{k}$$

and

$$\Delta\tau_{ij}^{k} = \{\frac{Q}{L_k}\ if\ ant\ uses\ curve\ ij\ in\ its\ tour\ 0 \quad otherwise$$

where

L_k: the cost of the K_{th} ant's tour.

Q: is a constant

- Discussion variable

The discussion variable is presented as follows:

V= $\{v_0, v_1, v_2, \ldots v_n\}$. The set of wells (nods) and their number n

E= $\{(i\ , j\ ) :\ i, j \in V, i \neq j\ \}$.

$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\ \forall (i,j) \in E.$

d_{ij} = The matrix of distant between wells.

E: The set of ribs that reaches between the wells (nodes).

v_0: The wells from which the path starts.

C_{ij}: The cost matrix for moving from the well from i to$j$, (i.e., distance d_{ij}*cost).

$t_{ij.}$ time of arrival between wells i and$j$, (t_{ij} >0).

t_i: time of arrival to well i.

// ACO for Oil-Gas Pipeline Wells Problem //

- Input: $Location\ (x_i, y_i), x\ and\ y\ \in R^{1\times n}$, $i = 1,2,3 \ldots., n$
- Output : $Final_{cost}$
- Apply ***ACO ShortPath*** (Data) // for all well locations
- $Min_{cost} \leftarrow \boldsymbol{ACOShortPath}\ (Data)$ // Estimate the minimum cost / shortest path for all wells locations
- Estimate the final cost based on the Wells problem as the follows
  $$Final_{cost} = Min_{cost} - LW_{cost}$$
  where LW_{cost} is the cost of the last well . //

D. GA Algorithm

It is known that the genetic algorithm is a smart algorithm to choose the best solution among a large number of solutions and to make interventions and modifications between these solutions to create a better solution. In general, the genetic algorithm contains several basic steps to resolve various issues that differ in the method of formulation and implementation according to the issue and scope of application. The following is an explanation of the proposed genetic algorithm steps to solve the issue of the shortest path linking the oil-gas wells with the main station.

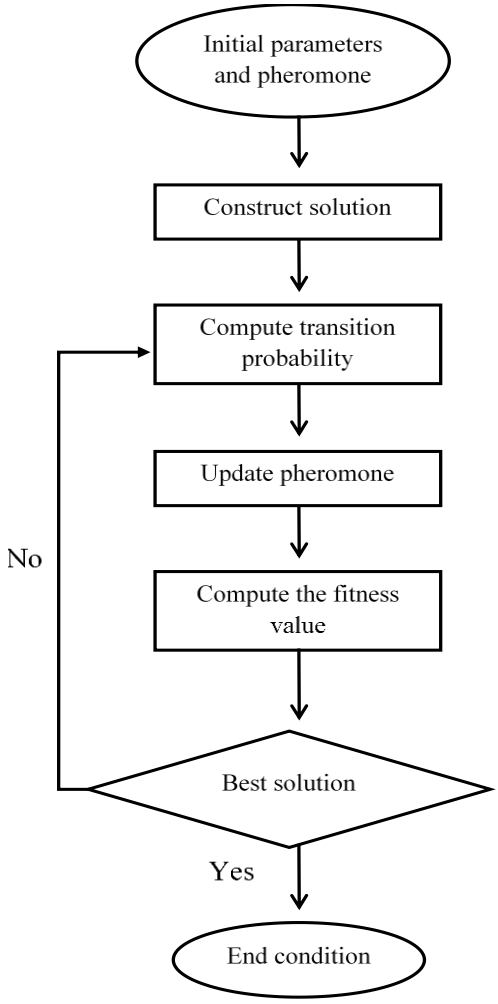


Fig. 2. ACO Algorithm Flowchart [21].

Step 1: (Initial data): This is a reading of the variables' values for the issue of linking oil-gas wells and is a cost or distance matrix. It represents the cost or the distance between the wells.

Step 2: (Generation initial population): The first point is the process of building the primary generation in a random way where several initial chromosomes are created so that each chromosome represents an integrated path. The length of the path is the number of points to pass where each gene of the chromosome represents the location of a well. The figure below shows a chromosome of length L.

Step 3: (Objective function): In this step, the objective function of the issue and each of the generation segments, which represent the sum of the costs or distances, are evaluated.

Step 4: (Selection): All chromosomes of the primary generation are selected as parents. Each chromosome mates with all chromosomes in the primary generation where each mating process produces two new individuals.

Step 5: (Generation): After selecting parent segments, subsequent generations are generated, and the number is specified by the value entered at the beginning of the execution. Fig. 3 illustrates the flow chart for generations.

Step 6: (Crossover): The overlapping process is the selection of two segments. Then, the process of transition is performed to obtain the offspring. The process of creative interference in the issue of the traveling salesman problem differs from other issues because each gene represents the sequence of a city and the city cannot be repeated within the chromosome.

Step 7: (Mutation): In this algorithm, the mutation occurs on the same gene. After configuring the generation segments, the target function for each section of the new generation is calculated in the same way that the initial generation segments were calculated. After configuring the specified number of generations, the execution of the function stops, and the results are evaluated to observe the closeness to decide whether to continue generational formation or to stop if the results are appropriate. The type of proposal is in the method of formulating and applying each step of the proposed algorithm. The chromosome is a complete pathway (each gene represents a city sequence).

Gene1	Gene2	Gene3	…	…	Gene L

Fig. 3. Generations Flowchart.

// GA for Oil-Gas Pipeline Wells Problem //

- Input: $Location\ (x_i, y_i), x\ and\ y\ \in R^{1\times n}, i = 1,2,3 \dots, n$
- Output : $Final_{cost}$
- Apply ***GA ShortPath*** (Data) // for all well locations
- $Min_{cost} \leftarrow \boldsymbol{GAShortPath}\ (Data)$ // Estimate the minimum cost / shortest path for all wells locations
- Estimate the final cost based on the Wells problem as the follows

$$Final_{cost} = Min_{cost} - LW_{cost}$$

where LW_{cost} is the cost of the last well . //

III. Simulation and Results

The scope of this study was around the northern Rumaila field, known as the largest oil field in Basra. It was discovered in 1953 with a length of 80 km and a width of 4 km, where the number of wells found was between 600 - 840 oil and gas wells [19]. All simulation experiments reported in this paper applied the model attributes as shown in Table I.

TABLE. I. Simulation Model Attributes and Parameters Value

Parameter	Value	Note
No. wells	50-840	For all algorithms (PSPA, GA , and ACO)
Pop.Size	10-50	For GA and ACO [23];[25]
numIter	1e5	For GA
maxIts	300	For ACO
numAnts	10-50	For ACO [22];[24]
beta	1	For ACO
Q	1	For ACO
rho	0.4817	For ACO
q_0	0.2770	For ACO

where
numIter : number of iteration for GA Algorithm.
maxIts : number of iteration for ACO Algorithm.
numAnts : number of ants for ACO Algorithm.
beta : A parameter which determines the weight of the heuristic function eta
Q : constant that determines how to update tau.
rho : the pheromone evaporation rate.
q_0 : A parameter determining how often we take the route with the best.

TABLE. II. COMPARISON OF DIFFERENT METHODS

No. of wells	PSPA algorithm		ACO algorithm		GA algorithm	
	Cost	Run Time (sec.)	Cost	Run Time (sec.)	Cost	Run Time (sec.)
50	74626.56	1.08	69389.00	18.45	63026.76	792.10
100	123310.83	1.45	122160.47	31.65	108717.03	850.27
200	365176.59	2.86	316753.14	59.07	300203.50	921.23
300	489842.80	4.65	383333.61	113.59	360297.20	1013.17
400	526711.57	7.42	484269.53	149.67	439977.22	1068.95
500	586227.15	9.58	547907.20	200.69	539163.68	1146.49
600	635362.30	12.39	634612.61	272.91	619332.93	1259.24
700	693048.53	15.84	695105.33	359.73	745123.53	1279.89
840	740728.59	21.74	775026.42	459.05	903172.01	1461.05

Table II and Fig. 4 show the comparison results of different methods. The studied algorithms (PSPA, ACO, and GA algorithms) were applied to determine the short path for various scenarios, where the number of wells was set to 50, 100, 200, 400, 500, 600, 700 and 840. From the results, it was clear that the lowest cost of laying oil and gas pipelines was by applying the GA algorithm when the number of wells was set to 50-600. Conversely, the PSPA algorithm showed the best performance in terms of required implementation time for all scenarios. Besides, PSPA appeared to have acceptable performance in terms of the cost of the pipeline when the number of wells was arranged between 50 and 840. Furthermore, PSPA showed the best performance for 700 and 840 wells in terms of the cost of laying oil and gas pipelines compared to ACO and GA. It should be noted that the ACO algorithm showed medium performance in terms of the cost of laying oil and gas pipelines compared to PSPA and GA.

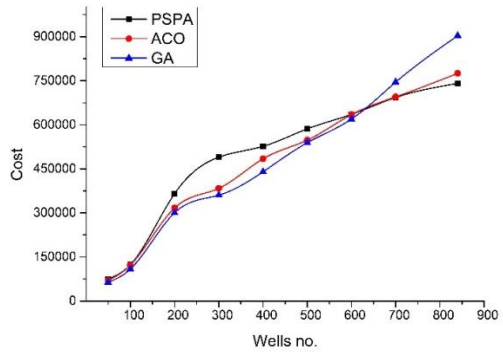


Fig. 4. The Comparison of different Methods in Terms of Cost.

IV. CONCLUSIONS

This paper presents a performance evaluation of different short path algorithms to improve oil-gas pipelines. The chosen algorithms in this paper are PSPA, ACO, and GA algorithms where the main performance metric is the cost of pipelines. The performance comparison showed that the lowest cost of laying oil and gas pipelines was by applying the GA algorithm when the number of wells was set to 50-840. Conversely, the PSPA algorithm showed the best performance in terms of required implementation time for all scenarios. Besides, PSPA appeared to have acceptable performance in terms of the cost of the pipeline when the number of wells was between50-840. Furthermore, the PSPA algorithm showed better performance for 700 and 840 wells in terms of the cost of laying oil and gas pipelines compared to ACO and GA. Note that the ACO algorithm showed middle performance in terms of the cost of laying oil and gas pipelines compared to PSPA and GA.

ACKNOWLEDGMENT

Authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for kindly providing us with the internal funding. This work is also supported by the Research Management Centre (RMC), Universiti Tun Hussein Onn (UTHM) under the TIER 1 grant number H075.

REFERENCES

[1] BP, “Statistical Review of World Energy 2014,” Br. Pet., no. June, pp. 1–45, 2013.

[2] H. R. Zhang, Y. T. Liang, Q. Xiao, M. Y. Wu, and Q. Shao, “Supply-based optimal scheduling of oil product pipelines,” Pet. Sci., vol. 13, no. 2, pp. 355–367, 2016.

[3] G. A. Gabriele, “The Generalized Reduced Gradient Method : A Reliable Tool for Optimal Design,” no. May 1977, pp. 394–400, 1977.

[4] A.R. Simposon, L. Murphy, and G. Dandy, “Pipe Network Optimisation Using Genetic Algorithms by University of Adelaide South Australia 5001 Paper to be presented at the ASCE Water Resources Planning and Management Conference Seattle , Washington Using Genetic Algorithms ." Proc, ASCE. Water Resources Planning and Management Division,” no. May, 1993.

[5] J. Zhou, G. Liang, T. Deng, and J. Gong, “Route Optimization of Pipeline in Gas-Liquid Two-Phase Flow Based on Genetic Algorithm,” Int. J. Chem. Eng., vol. 2017, 2017.

[6] Jiancheng, “Jiancheng, L., 2001. The Topology Optimization Design of Oil-Gas Gathering and Transportation [J]. Petroleum Planning & Engineering, 6,” p. 2001, 2002.

[7] Zhang.Z, The Optimization and Programming of Oilfield Surface Gathering Network [D], China University of Petroleum (Esat China), 2011, pp.1-15.

[8] “A comparison of burial, maturity and temperature histories of selected wells in southern Iraq,” no. October, 2017.

[9] Basrah oil co., “boc.oil.gov.iq.pdf,” 2016. [Online]. Available: http://www.boc.oil.gov.iq/index.php?name=Pages&op=page&pid=108. [Accessed: 08-Mar-2016].

[10] W. Dawotola, P. H. A. J. van Gelder, and J. Vrijling, “Risk Assessment of Petroleum Pipelines using a combined Analytical Hierarchy Process - Fault Tree analysis,” Proc. 7th Int. Probabilistic Work. Fac. Civ. Eng. Geosci. Delft Univ. Technol. Stevinweg 1, 2628CN Delft, Netherlands, vol. 13, no. 2006, pp. 491–501, 2009.

[11] G. C. Dandy, A. R. Simpson, and L. J. Murphy, “An improved genetic algorithm for pipe network optimization,” Water Resour. Res., vol. 32, no. 2, pp. 449–458, 1996.

[12] R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” MHS’95. Proc. Sixth Int. Symp. Micro Mach. Hum. Sci., pp. 39–43.

[13] M. S. El-Abbasy, A. Senouci, T. Zayed, F. Mirahadi, and L. Parvizsedghy, “Artificial neural network models for predicting condition of offshore oil and gas pipelines,” Autom. Constr., vol. 45, pp. 50–65, 2014.

[14] Shuwen, Zhang and R.Meng, Pileline layout Optimization of Large Gas Field Gathering [J], Petroleum Planning & Design, 1998.

[15] E. W. Dijkstra, “Dijkstra.Pptx,” vol. 271, pp. 269–271, 1959.

[16] A. K. Arya and S. Honwad, “Multiobjective optimization of a gas pipeline network: an ant colony approach,” J. Pet. Explor. Prod. Technol., no. 123456789, 2017.

[17] Ermin et al., “Gas Network Topology Optimization for Upcoming Market Requirements 1,” vol. 9, no. March, 2011.

[18] Edgar, T.F.Himmelblau D.M, Bickel T.C, Optimal Design of Gas Transmission Networks[J].Society of Petroleum Engineering Journal.SPE Number, 4:6031,1978.

[19] Almaalei, Nabeel Naeem Hasan, Siti Noor Asyikin Mohd Razali, and Nayef Abdulwahab Mohammed Alduais. "An efficient algorithm to improve oil-gas pipelines." International Journal of Engineering & Technology 7, no. 4 (2018): 5412-5418.

[20] Ning, J., Zhang, Q., Zhang, C., & Zhang, B. (2018). A best-path-updating information-guided ant colony optimization algorithm. Information Sciences, 433–434, 142–162. https://doi.org/10.1016/j.ins.2017.12.047

[21] Yoo, K. S., & Han, S. Y. (2013). A modified ant colony optimization algorithm for dynamic topology optimization. Computers and Structures, 123, 68–78. https://doi.org/10.1016/j.compstruc.2013.04.012.

[22] David Renfrew. (2009). Traffic Signal Control with Ant Colony Optimization.

[23] Gonen, B. (2011). Genetic Algorithm Finding the Shortest Path in Networks. International Conference on Genetic and Evolutionary Methods, (January 2011), 1–4.

[24] Li, J., Sun, S., Huang, Y., & Wang, N. (2010). Research into self-adaptive hybrid ant colony algorithm based on flow control. Proceedings - 2010 2nd International Workshop on Intelligent Systems and Applications, ISA 2010, (978). https://doi.org/10.1109/IWISA.2010.5473446.

[25] Scrucca, L. (2017). On some extensions to GA package: Hybrid optimisation, parallelisation and islands evolution. R Journal, 9(1), 187–206. https://doi.org/10.32614/rj-2017-008.

A New Model to Detect 2D Hand based on Multi-feature Skin Model

Abdullah Shawan Alotaibi[1]*
Computer Science Department
Shaqra University, Shaqra
Saudi Arabia

***Abstract*—Recognition of hand gesture is one of Human PCs most growing interfaces. In most vision-based signal recognition system, the initial phase is hand detection and separation. Because the hands are linked to a variety of day by day, local work experiences both extraordinary changes in the illumination and the innate unbroken appearance of the hand. In order to address these issues, we suggest another 2D hand position software that can be seen as a combination of multi-feature hand proposal generation and cascading neural system network characterization (CCNN). When considering various luminances we select color, Gabor, Hoard and Filter to separate the skin and produce a hand proposal. Therefore, we are selling a cascaded CNN that holds the deep setting information between the proposals. A mix of some datasets, including a few Oxford Hands Datasets, VIVA Hand Recognition, and Egohands Datasets, is tested as the positive example and image patch Net 2012, FDDEB dataset as a bad example; the proposed Multi-Feature Directed Cascaded CNN (MFS-CCNN) strategy. Aggressive results are achieved by the technique proposed. Our average sample dataset accuracy is considerably inferior to DPM. With an average of 43.55 and 51.78 percent accuracy, our CCNN and MFS-CCNN model perform DPM. Average accuracy of the CCNN model in a combined test set is 9.16% higher than the SSD model. Still, our model is faster than a DPM based on the statistical performance.**

Keywords—Hand detection; feature modeling; convolutional neural networks

I. INTRODUCTION

With the advances of astute processing lately, there has been a developing enthusiasm for growing increasingly instinctive and proficient methods for collaboration among human and PCs [1]. As hand signals structure a broad piece of characteristic human correspondence, vision-based motion acknowledgment is an engaging option in contrast to conventional presently utilized gadgets, for example, console and mouse [2]. As of late, countless examines [3, 4] about human hand exercises has increased exceptional consideration. 2D vison-based hand motion acknowledgment is conceivably a minimal effort data preparing instrument for human – PC interface.

Human hands are associated with a variety of tasks day after day with a lot of segmenting data. Hand identification is the biggest development in semantically awareness of hand operation. For instance [5-8] the color, region, Hoard [9], surface filter [10] or the district family CNN such as RCNN, Fast RCNN, Quicker R CNNN [11—13] and R-FCN [14] etc. are the key dependent features of usable recognition scans. The assignment of the recognition of human hands consistently experiences the multifaceted nature of the foundation which in addition causes rapid change in the condition of enlightenment [15].

Secondly, when (for example, gesture based communication or a command when associating with computers) or controlling objects constantly changes the form of human hands. Such changes in presentation impact the exploration and the understanding of hand exercises [6]. Thirdly, a specification for a perfect hand-location should be able to continue to work, similar to the steps for detection of hand motion. Extraordinary consideration is given to speed in business applications [16].

They recommend a MFS-CCNN approach to manual positioning and sorting in order to successfully supervise these problems. We could produce hand-area proposals to monitor a Cascaded CNN model by means of the viability of multi-feature registration. In the meantime, we use the cascade CNN system to keep information missing for classification beyond what others find feasible. We have built our CNN cascaded template with a few free datasets such as Oxford Hand Dataset[5], VIVA hands [18], and image patches Net 2012[19], FDDB[20].

The fundamental commitments of this paper can be condensed as pursues.

- To understand skin district cover and construct hand offering jumping boxes we use a multi feature skin model.
- We found that cascading CNN structure which aggregates attempts to improve the detection by setting the data for bouncing box regression.
- They demonstrated that mixing handcraft and CNN approaches can produce accurate results for hand recognition tasks.

This paper is organized as follows. In Section 2, we review state-to-art researches on hand detection and the average and limitations of feature based and region-based CNN work. Section 3 will address the methodology suggested for hand identification and classification. The findings of the test were summarized in Section 4. We will eventually reach a conclusion in Section 5 and prepare for the future.

*Corresponding Author

II. Related Work

In the individual PC community, improved fact and command driver behavior and so on, recognition and following of human hands are commonly used. In this article we concentrate on differentiating hands in multiple situations. Mittal et al. [5] exhibited a successful two-way approach to human hands. Three fundamental finders were used to introduce hand bouncing containers, which could be used to train a classifier for the final determination. Nevertheless, their skin-based features were profoundly respected in different conditions of enlightenment. The base pixels using the surface and limiting features were demonstrated by Fathi et al. [7]. We distinguish hands from various items using color histograms from the missing facial region pixels. Li and Kitani [6] have therefore prepared a pixel level indicator for objects with slowly realistic ego-centered characteristics, for example, lighting gestures and lighting switch. Their strategy combines super-pixels with color + surface features and invariance descriptors. While their methods have achieved a focused outcome, a few disappointments exist while hands are on bland premises. Also, as shown in the Fig. 1, it separates hands and other divisions of body, such as the lower arms and face.

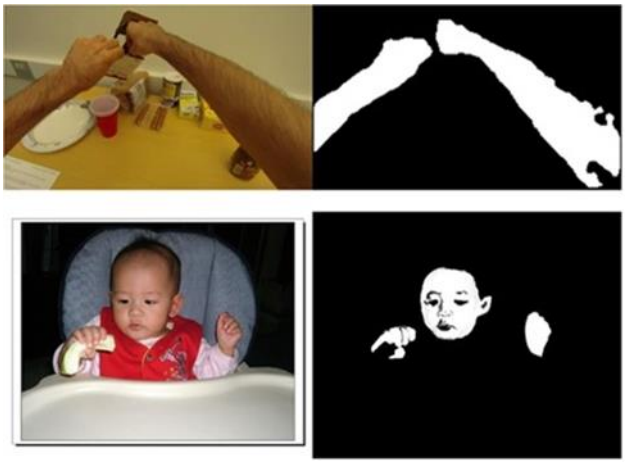


Fig. 1. The Hand Detector Pixel Stage not only Detects Hands but also other Field Skin Masks.

In [21] the two-arrange hand indicators of the color (RGB, HSV, LAB) and edge (HOG, GIST) features of Betancourt [24] were also proposed, using a classification (SVM, random forest, choice trees). As recently as this, Bambach et al.[23] introduced another egocentric dataset called EgoHands. This collection contains documents where a few people wear camera glasses each other while playing a pre-game. Their approach pipeline is like R-CNN, but they send the district a probabilistic plan and split it into a pixel point. Large estimated information requires the viability of profound learning approaches [31] to work consistently in difficulties. MS-FRCNN [24] was shown in THN le et al. to disable several features so that sincere hand seeking in vehicles can be done. They extend the Faster-RCNN system in the region Proposal Network (RPN) and organize discovery with a diverse scale of deep element extraction. The technologies referred to above are nevertheless centered on fixed vision hand discovery alongside their picked data set (egocentric or in a vehicle). Specific conditions cannot be checked. This problem can be dealt with by entering datasets which include more scenarios. The problem of datasets can be caused by the combination of datasets methodology. Yin, H. proposed [23, 32] included technical determinations to address this problem. In a number of applications, this principle could be used.

The problem of data malfunction in the RCNN family reduces the number of articles. For example, in RCNN's last item guide, a 32×32 Article stays 2×2. Hypernet [25] presented Hyper, which contains different levels of characteristic maps and has them packed into a single space. Although the Hyper Feature fuses overall information, the skip layer system also decreases data that is not associated with the Hyper Feature.

We are suggesting a new method that uses skin that demonstrates the generation of hand proposal and cascading CNN for the identification of hands in several situations in this paper. Little research apparently has sought to strengthen the display of the skin and the CNN approach for hand recognition. Given the pace of the planned technique implementation, they have altered the SSD [26] with a cascaded overall structure to retain in-depth settings. As positive example and photo patch Net 2012 [19], FDDB [20] data as bad case, we have built and are testing the proposed Oxford Hand Dataset [5], VIVA Hand Discovery [17] and EgoHands [18] mixed strategies. Our technique accomplishes aggressive hand identification results.

III. Proposed Method

For hand location, we present a multi-function controlled cascaded CNN structure. Fig. 2 shows the general flow. Next, by color region and histogram (to be multi-included specific), we focus the skin district A series of hand guidelines from the regional skin areas will be established at this stage. Eventually, the manual target is characterized by a cascaded CNN system. Skin Region Detection

For skin recognition task, Color is a generation included [27]. In the test [6], the low surface area can be isolated between hands and other similar hued items. We have used HSV and LAB color areas in conjunction with [6] to test the identification of the skin region. When color-based skin conditions are addressed, items giving equal color to skin are abused. In order to increase the bias of the skin color surface, they incorporate the response of 32 Gabor channels (4 directions and 4 scales). Therefore, the work for the skin display relies instead of a single pixel assessment on nearby pixel data. Since spatially modified neighbourhood-inclining histogram characteristics are capable of productively capturing the invariable characteristics of the area, we have chosen the Hoard [9] and Filter [10] descriptors to capture neighbouring hands [22] forms. Like [6], we train a gathering of regressors listed by a worldwide color histogram.

The posterior appropriation of a pixel i given a nearby appearance feature l and a global appearance include g, is registered by minimizing over various scenes s

$$p(i|\, l, g) = \sum_s p(i|\, l, s)\ p(s|g) \quad (1)$$

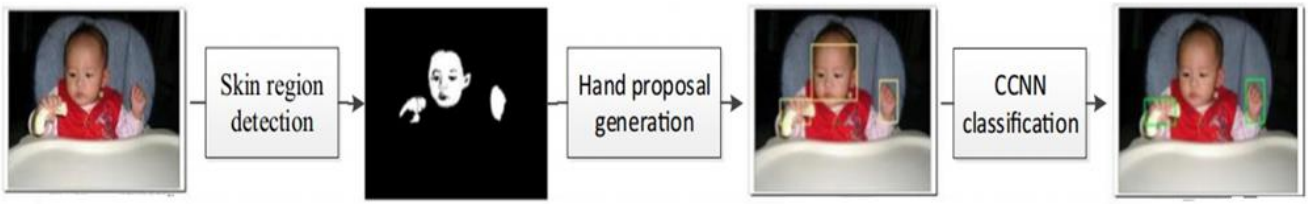


Fig. 2. Presents the Proposed Method Pipeline.

Where $p(i|\, l, s)$ is the yield of a discriminative worldwide appearance-explicit regressor and $p(s|\, g)$ is a contingent conveyance of a scene s given a worldwide appearance feature g. We use k-implies the bunch of different objects to construct global appearance models on the HSV histogram and each community is taught about various arbitrary tree regressors. Through scene unit codes both the look and the light of the scene by a histogram across each of the three HSV color streams. At the end, the skin saw will share a comparative transport appearance in the feature area. The restrictive $p(s|\, g)$ is approximated utilizing a uniform dispersion over the n closest models (in our examinations, $n = 5$) learned at preparing.

A. Hand Proposed Generation

In view of the skin mask, we determined the base encasing square shape (r_m) of each associated area (R). Likewise, equivalent to the explanation organization of article location task, The Hand proposal box (B) is defined as the most severe x and y directions provided by xmax, xmin, ymax and ymin to potentially useful objects. To generate a hand proposal from the hands that contribute to the arm area, we decide whether the length is over multiple times than the width of r_m, this associated district presumably speaks to an arm locale and the hand item is toward the finish of this area. If there should arise an occurrence of misdetection of multi-feature model (particularly limit pixel misfortune), we calculate a hand proposition box with t times the width of rm at each finish of this district. In our test, $t = 1.5$ created the best outcomes. The hand proposition generation is exhibited in algorithm 1. The hand proposal generation represented in Table I.

B. Hand Recognition with Cascaded Feature Aggregation CNN

We create a cascaded feature aggregation CNN to identify the hand area under the control of a hand proposition generated by multiple apps. As previously mentioned, the hand goal position has different hand photos in real situations. CNN's craft-based policies such as Quicker RCNN [13] and SSD [26] have unbelievable results in the discovery of the items on the PASCAL VOC dataset. However, the subsequent aims of the last layer feature guide, because it is much literal than the information picture (during the use of the pooling layer), generally do not contain the subtleties of a small article and setting of data. In the genius presented hand location CNN strategy, we pick SSD [26] as fundamental design in light of identifying speed and a cascaded feature conglomeration CNN structure is utilized to lessen the loss of setting data.

Assuming i and O represent the input and output of SSD network respectively. F_n is the feature maps of $n - th$ layer, $\theta_n(.)$represents the non-linear layers between the $(n - 1) - th$ layer and $n - th$ layer including convolutional layers, pooling layers and ReLU layers, and so on., $Y_n(.)$ is the capacity to change the $n - th$ layer feature maps to the recognition results for a specific scale range. f is the last operation to complete all the intermediate outcomes and produce the new recognition is therefore defined by multi-scale SSD feature maps:

$$F_n = \theta_n(F_{n-1}) = \theta_n\theta_{n-1}\,(\dots\theta_1(I))), \tag{2}$$

$$O = f\big(\gamma_n(F_n), \dots, \gamma_{n-i}(F_{n-i})\big), n > i > 0, \tag{3}$$

As per Eq. (3), it seems to rely strongly on a solid hypothesis of success. Since the feature charts in each layer are solely responsible for the performance of their size, each F alone is believed to be sufficiently advanced to support correct identification and restriction. The detailed map of component F must contain appropriate data that can heartfully capture the careful area of the low, shielding, obstructed or dark objects, including hands [13, 26]. We present a gradually sensitive element reproduction that takes account of the data setting in a cascaded system structure. Fig. 3 shows the nature of the cascaded structure. The function is characterized as follows:

$$O = \hat{f}\left(\gamma_n\left(\hat{F}_n(S)\right), \gamma_{n-1}\left(\hat{F}_{n-1}(S)\right), \dots, \gamma_{n-i}\left(\hat{F}_{n-i}(S)\right)\right), \tag{4}$$

$$S = \{F_n, F_{n-1}, \dots, F_{n-i}\}, n > i > 0,$$

$$size\,(F_{n-i}) = size\,\left(\hat{F}_{n-i}(S)\right), for\ all\ i \tag{5}$$

TABLE. I. SHOWS THE HAND PROPOSAL GENERATION

Hand proposal generation
Input: skin region mask set ***R***
Output: hand proposal box set ***B***.
For $i = 1: num(R)$
1. Calculating of minimum enclosing rectangle r_m along with the width and length of r_m.
2. For each r_m
If $\frac{1}{3}\ width \le length \le 3\ width$
Then ***B*** = xmax, xmin, ymax, ymin of R
3. Else
B_1 =xmin+1.5 width, xmin, ymax, ymax-1.5width of ***R***
B_2 =xmax, xmax-1.5 width, ymin+ 1.5 width, ymin of ***R***
End for
4. Return ***B***
End

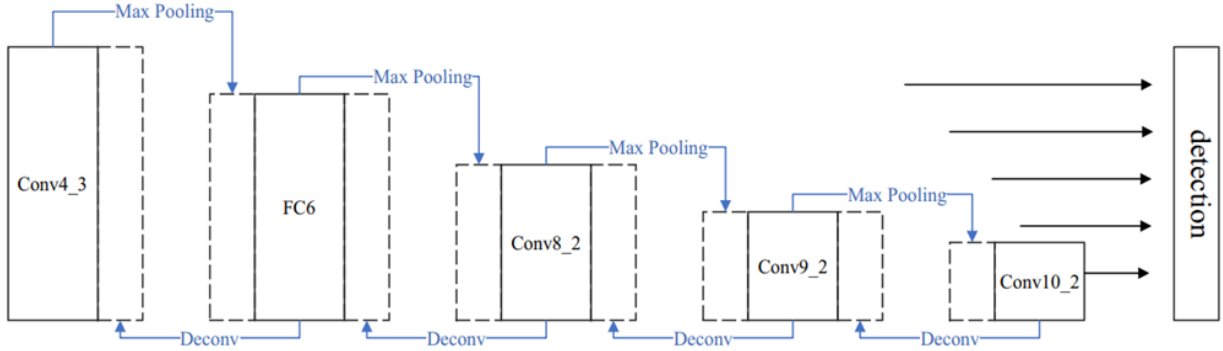


Fig. 3. Structure of the CCNN Network. The Function Aggregation in Reduced VGG-16 is shown in this figure. The Strong Boxes Represent SSD Characteristics Such as Conv4_3, FC6, Conv8_2, Conv9_2 and Conv10_2. The Dotted boxes Selected by Arrows were Added to the Next Feature Maps. Inspired by Hypernet [25], we are Integrating Max Pooling and Deconvolution Operations, as shown by Arrows, to Add up and Down Capabilities.

Where S is a lot of all the feature maps contribute to the detection function $f(\bullet)$. Dissimilar to in eq. (3), $\hat{F}_n(.)$ is a function that considers all the contributing feature maps and yields another include portrayal of a similar dimensionality to F_n.

An immediate mapping from S to $\hat{F}_n(S)$ need to be depend on an impressive size deep network with various layers of non-linearity. This will cost a lot of algorithm and it will be difficult to organize one generation. The alternative is to design an iterative approach where each move advances a little but substantially and consistently. Enlivened by [25], the map maps from conv1, conv3 and conv5 are used for Max pooling and Deconvolution. We use a similar activity in CCNN to add the neighbouring layer feature maps (as shown in Fig. 3). The following is a mathematical overview of this approach,

$$\widehat{F_p^{t+1}} = M\left(\widehat{F_p^t}, \widehat{F_{p-1}^t}, \widehat{F_{p+1}^t}; W\right), t > 0, \quad (6)$$

$$\widehat{F_n^t} = F_n, for\ all\ n \text{ where } t = 1, \quad (7)$$

Where M is a function maps only $\widehat{F_n^t}$ and its adjacent (higher and lower level) counterparts at step t to a new $\hat{F}_p$ at step $t + 1$. The variable M has a weight parameter W. The learning goal failure feature is similar to SSD with the lack of position and confidence. Assuming $x_{ij}^k = \{1,0\}$is an indicator for matching the $i - th$ default box to the $j - th$ ground truth box of category k:

$$L(x, c, h, g) = \frac{1}{N}(L_{conf}(x, y) + \alpha L_{loc}(x, h, g)) \quad (8)$$

Where, N is the organized standard box number. If N= 0 is omitted, the loss is set to 0. The loss of restrictions is a Smooth L_1 loss between the box (h) and the parameters for the bottom truth box (g). The Softmax loss for many classes of confidences is the loss of certainty (c). Through cross-approval, the weight variable α is set at 1.

C. Summary of Proposed Method

As indicated by the past description, the algorithm flow of MFS-CCNN can be summarized by Fig. 4. We extract the skin mask of input image patch by Multi-feature model. In light of skin mask, the hand proposition bouncing boxes could be produced by algorithm 1. At last, CCNN could characterize and relapse the hand protests alongside exact position. The assessment of MFS-CCNN will be introduced in Section 4.

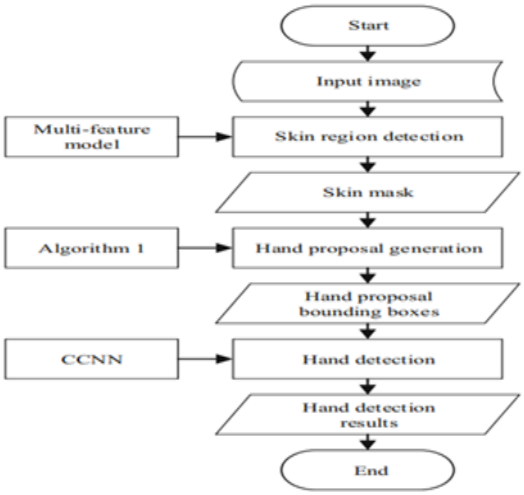


Fig. 4. MFS-CCNN Structure Graph.

IV. Experimental Results

In multi-scenes, this work aims to recognize hands. Our skin recognition model uses a sparse combination of HSV, LAB Gabor, Hoard and filter features. We created the template using the GTEA dataset and the skin district with a 9×9 image patch update, as suggested by [6]. The hand district plan is based on the skin template over-division that is appropriate for multi-situations with different foundations. The CCNN base system is a 2012 object patch network pre-trained VGG16 [28]. Our models were developed and tested using a mixture of Oxford Hand Dataset [5], VIVA hand location 17] and EgoHands [18] and neutral examples were also used to distinguish hands from various grades [19] and faces [20]. Remember that this research is just a set of commands, so EgoHands' "left own," "wrong say," "left side" and "other right" are listed as one category – hand. Table II, which contains 14,151 image patches for preparation and approval and another 7283 photographs for testing, shows how subtle the joined hand dataset is. For evaluation of the proposed technique we receive a 0.5 IoU Sift. The Stochastic Slope Plummet network has been optimized to maximize 60,000 simulations, with an energy of 0.9. We have reduced weight and the underlying study rate is 0.0005. The sample sequence of the analysis is shown in Fig. 5. The main section is the first selected image from the hand datasets of EgoHands, VIVA and

Oxford. We use the multi-function skin recognition model right outside the bat to focus skin cover as shown in the following section. We establish the proposition of the hand item as shown in the third column in light of the cover on body. Ultimately, the hand exploration findings are the last segment. The implementation of the DPM-based strategy [5], SSD [26], CCNN and MFS-CCNN is outlined in Table III. The details of the combined hand dataset are shown in Table II; it contains 14,151 images for training and validation, and another 7283 images for testing. A 64-bit Ubuntu 14.04 PC with the CPU Intel (R) Core(TM) i7-5960X, 320 G memory and TITAN Xp GPU will assess the method proposed, but the results showed that it can improve the recognition with skin model proposal and the conglomeration of information within the CNN. The proposed technology is also evaluated. Our normal precisions on the test datasets are altogether superior to [5]. Our normal data set accuracy is totally greater than [5]. With standard precision of 43.55 % and 51.78 %, our CCNN models alone and MFS-CCNN beat [5]. In consolidated test set, the CCNN model achieved normal accuracy of 9.16 % higher than the SSD model. In the combined sample collection Oxford Hand Dataset and Egohands, MFS-CCNN obtained the most notable performances. However, the VIVA test results show that the best possible performance for the CCNN model is created (0.86 per cent above CCNN-MFS). In addition, the consequences were assessed using the precision check bends outlined in the Fig. 6 for MFS-CCNN, CCNN and SSD. For red, green and blue respectively, the after-effects of MFS-CCNN, CCNN and SSD are spoken of. We use 11 accuracy points to create the plot and check information. As shown in line two Fig. 5, the right position is in a low luminances state, the whole locale of this hand can hardly be differentiated by our skin template, so that CCNN refuses to recognize it. Fig. 7 demonstrates the efficiency of the minor hand on the right side by CCNN. While the skin position model in low luminance is shaky, it has improved hand discovery (the MFS-CCNN model has averaged CCNN by 8.23%).

We also contrasted the execution time of DPM [5], SSD [26], CCNN and MFS CCNN versions. They also compared the time. It's much faster than our methodology [5]. The conglomerate structure with cascaded components and the skin template with many inclusive improved calculations in comparison to SSD. For instance, Squeezenet [29] and MobileNets [30], we accept that our approach can be time efficient by using a lightweight CNN as the basis system.

TABLE. II. DISTRIBUTION OF EXPERIMENT DATASET

Datasets	Training	validation	testing
Oxford Hand Dataset	3031	1780	823
VIVA hand detection	3465	2035	5500
EgoHands	2458	1382	960
Total	8954	5197	7283

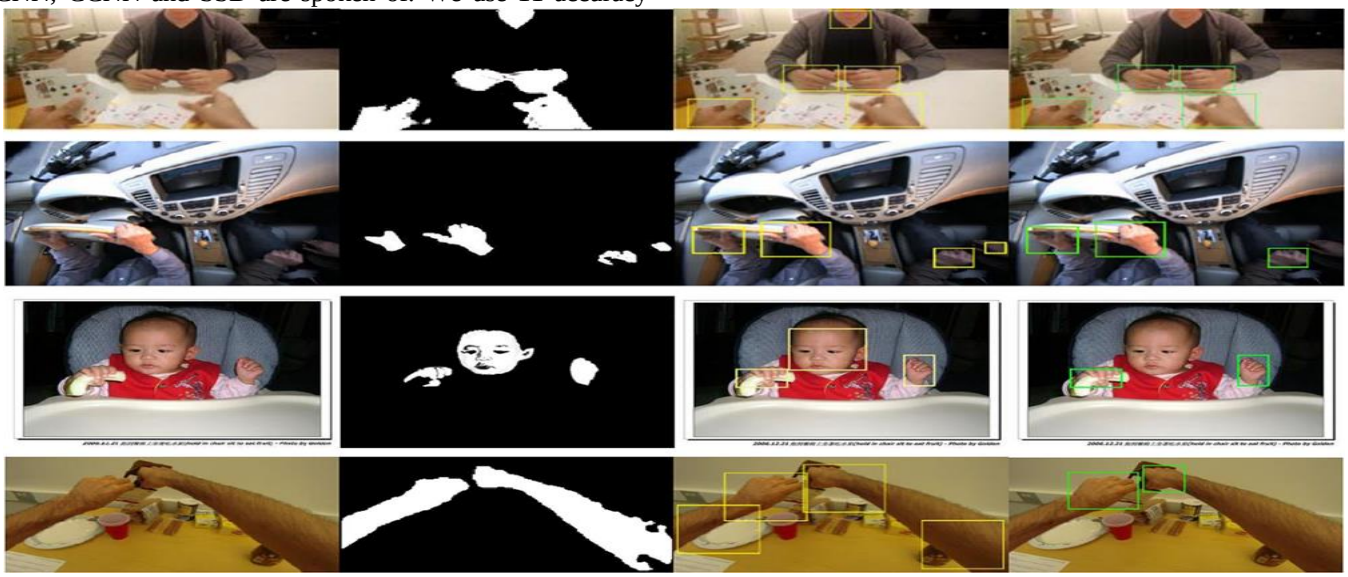


Fig. 5. The Original Patch of the Selected Datasets is Column 1. Column 2 is the Mask Produced by our Skin Model for the Skin Region. Column 3 is a Skin Mask Hand Proposal. The Detection Results are in Column 4.

TABLE. III. RESULTS ON THE TEST SET

Methods	Combined test set	Oxford Hand Dataset	VIVA hand detection	EgoHands	Running time
DPM based [5]	46.62%	48.20%	40.15%	47.13%	26 s
SSD [26]	79.43%	79.78%	82.49%	77.96%	0.018 s
CCNN model (ours)	89.67%	85.65%	92.17%	88.81%	0.046 s
MFS-CCNN model (ours)	91.22%	89.26%	91.31%	91.76%	0.057 s

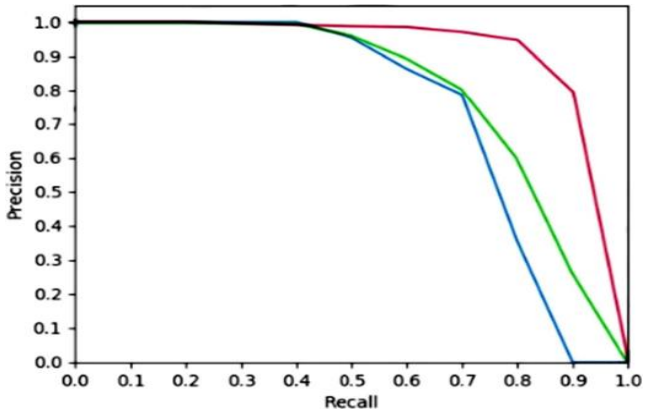


Fig. 6. Curve Notification Accuracy for Hand-Detection Correlation test. The Tests of MFSCCNN, CCNN and SSD are shown by a Black, Green and Blue Curve.

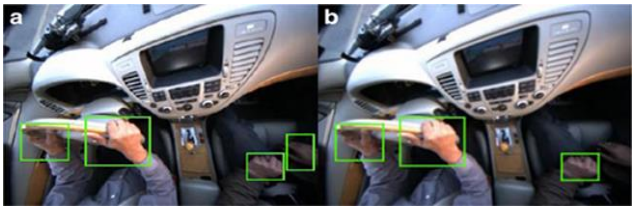


Fig. 7. Comparison between CCNN and MFS-CCNN for Weak Luminance Detection Tests. The CCNN and MFS-CCNN Results are Respectively (a) and (b).

V. Conclusion

In this paper we propose an approach for hand detection based on the multi-feature cascaded CNN model for the skin. We first use a HSV, LAB, Gabor, Hoard and Filter skin model for the identification of the skin area. Secondly, hand suggestion boxes depending on skin masks are produced. Finally, we propose to classify hand objects by cascading the CNN feature aggregation. The test shows that both the skin model and CCNN are able to enhance recognition and achieve state-of-the-art findings on mixed datasets. The technology proposed achieves aggressive results. Our average data collection accuracy is substantially better than DPM. With an overall accuracy of 43.55 and 51.78 % respectively, our MFS-CCNN system and CCNN CCNN alone outperform. The average accuracy of CCNN model in the combined test set was 9.16 per cent better than that of the SSD. Furthermore, the numerical results show our model is faster than DPM. Future work will focus on timely assessment. The current evaluation and recognition of operation will also be updated.

References

[1] Stergiopoulou, K. Sgouropoulos, N. Nikolaou, N. Papamarkos, and N. Mitianoudis, "Real time hand detection in a complex background," Engineering Applications of Artificial Intelligence, vol. 35, pp. 54-70, 2014.

[2] A. Ebert, N. D. Gershon, and G. C. van der Veer, " Human-computer interaction: introduction and overview," Künstliche Intelligenz, vol. 26, no. 2, pp. 121-126, 2012.

[3] J. Zariffa, and M. R. Popovic, "Hand contour detection in wearable camera video using an adaptive histogram region of interest," J NeuroEng Rehab), vol. 10, no. 1, pp. 114–114, 2013.

[4] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D image patches," IEEE International Conference on Computer Vision, vol. 22, pp. 3889–3897, 2015.

[5] A. Mittal, A. Zisserman, and P. Torr, "Hand detection using multiple proposals," British Machine Vision Conference, vol. 40, pp. 1–11, 2011.

[6] C. Li, and K. M. Kitani, "Pixel-level hand detection in ego- centric videos. Computer Vision and Pattern Recognition," vol. 9, pp. 3570–3577, 2013.

[7] A. Fathi, and J. M. Rehg," Learning to recognize objects in egocentric activities," IEEE Conference on Computer Vision and Pattern Recognition, vol. 42, pp.3281-3288, 2011.

[8] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara," Hand segmentation for gesture recognition in EGO-vision," ACM International Workshop on Interactive Multimedia on Mobile & Portable Devices, vol. 24, pp. 31–36, 2013.

[9] N. Dalal, and Triggs,"Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 1, pp. 886–893.

[10] D. G. Lowe, "Distinctive image patch features from scale-invariant key points," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region- based convolutional networks for accurate object detection and segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp.142–158, 2015.

[12] R. Girshick, "Fast R-CNN," IEEE International Conference on Computer Vision, pp.1440-1448, 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun," Faster R-CNN: Towards real-time object detection with region proposal networks," International Conference on Neural Information Processing Systems, vol. 39, pp. 91–99, 2015.

[14] W. Dai, Y. Li, K. He, and J.Sun, " R-fcn: Object detection via region-based fully convolutional networks," 2016.

[15] A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, and X. Twombly,"Vision based hand pose estimation: A review," Computer Vision & image patch Understanding, vol. 108, no. 1, pp. 52–73. 2007.

[16] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan," Vision-based hand-gesture applications," Communications of the ACM, vol. 54, no. 2, pp. 60– 71. 2011.

[17] The Vision for Intelligent Vehicles and Applications (VIVA) Challenge, Laboratory for Intelligent and Safe Automobiles, UCSD. http://cvrr.ucsd.edu/vivachallenge/, accessed on (9 feb 2019).

[18] S. Bambach, S. Lee, D.J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," IEEE International Conference on Computer Vision, pp.1949-1957, 2016.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei," image patch net large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[20] V. Jain, and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," UMass Amherst Technical Report, 2010.

[21] A. Betancourt, "A sequential classifier for hand detection in the framework of egocentric vision,"Computer Vision and Pattern Recognition Workshops, pp.600-605, 2014.

[22] Q. Wang, and G. Zhang, "Ore image patch edge detection using hog-index dictionary learning approach," Journal of Engineering, vol. 1, no. 1, 2017.

[23] H. Yin, and K. Gai," An empirical study on preprocessing high-dimensional class-imbalanced data for classification," International Conference on High PERFORMANCE Computing and Communications, pp.1314-1319.

[24] T. H. N. Le, C. Zhu, Y. Zheng, K. Luu, and M. Savvides," Robust hand detection in vehicles," International Conference on Pattern Recognition, pp.573-578, 2017.

[25] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," Computer Vision and Pattern Recognition, pp.845-853, 2016.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.Y. Fu,"SSD: Single shot MultiBox detector," European Conference on Computer Vision, pp.21-37, 2016.

[27] P. Kakumanu, S. Makrogiannis, N. Bourbakis,"A survey of skin-color modeling and detection methods," Pattern Recognition, vol. 40, no. 3, pp. 1106–1122, 2007.

[28] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image patch recognition," arXiv preprint arXiv:1409.1556, 2014.

[29] F. N. Iandola, S. Han, M. W. Mickiewicz, K. Ashraf, W.J. Dally, and K. Keutzer," Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," arXiv preprint arXiv:1602.07360, 2016.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, and T. Weyand," Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[31] K. Gai, M. Qiu, and X. Sun," A survey on fintech," Journal of Network & Computer Applications, vol. 103, pp. 262-273, 2018.

[32] H. Yin, K. Gai, and Z. Wang," A classification algorithm based on ensemble feature selections for imbalanced-class dataset, "International Conference on Big Data Security on Cloud (pp.245-249), 2016.

The Computational Efficiency of Monte Carlo Breakage of Articles using Serial and Parallel Processing: A Comparison

Jherna Devi[1]
Institute of Technology for Nanostructures (NST) and Center for Nano Integration Duisburg-Essen (CENIDE)
University Duisburg-Essen, Duisburg, D-47057, Germany
Department of Information Technology, Quaid-e-Awam University of Engineering Science & Technology (QUEST) Nawabshah, 67480, Sindh, Pakistan

Jagdesh Kumar[2]
School of Technology and Innovations
University of Vaasa, Finland
Department of Electrical Engineering, Quaid-e-Awam University of Engineering Science & Technology (QUEST) Nawabshah, 67480, Sindh, Pakistan

***Abstract*—This paper presents a GPU-based parallelized and a CPU-based serial Monte-Carlo method for breakage of a particle. We compare the efficiency of the graphic card's graphics processing unit (GPU) and the general-purpose central processing unit (CPU), in a simulation using Monte Carlo (MC) methods for processing the particle breakage. Three applications are used to compare the computational performance times, clock cycles and speedup factors, to find which platform is faster under which conditions. The architecture of the GPU is becoming increasingly programmable; it represents a potential speedup for many applications compared to the modern CPU. The objective of the paper is to compare the performance of the GPU and Intel Core i7-4790 multicore CPU. The implementation for the CPU was written in the C programming language, and the GPU implemented the kernel using Nvidia's CUDA (Compute Unified Device Architecture). This paper compares the computational times, clock cycles and the speedup factor for a GPU and a CPU, with various simulation settings such as the number of simulation entries (SEs), for a better understanding of the GPU and CPU computational efficiency. It has been found that the number of SEs directly affects the speedup factor.**

***Keywords*—Breakage of particles; Central Processing Unit (CPU); Graphics Processing Unit (GPU); CUDA; computational time; clock cycle; speedup factor**

I. INTRODUCTION

The breakage of particles is of interest in various fields of engineering and scientific research, including chemical engineering, aerosols, agriculture and medicine [1–3]. The population balance equation (PBE) provides a platform to develop the distributed phase. The PBE includes all the processes, such as nucleation, coagulation and breakage, that produce fragments and break simulation entries or parent particles from the population. Breakage is of major importance for understanding the behavior of, and dealing with, particle systems.

The particle breakage-population balance equation (BP-PBE), which characterizes the breakage dynamics in terms of the time evolution of the particle size distribution (PSD), is shown in Equation :

$$\frac{dn(v)}{dt} = -S(v) \cdot n(v) + \int_v^\infty n(v')S(v')b(v,v')dv' \quad (1)$$

where $n(v)dv$ is the concentration of particle sizes and $n(v)$ is the particle size distribution. The size range between v and $v + dv$ per unit volume, in time t, is defined by the BP-PBE Equation1.

where$S(v)$ is the rate at which a particle of size v breaks and the breakage function $b(v, v')$ describes the number fragment particles of size v resulting from the breakage of one parent particle of sizev'. The death term (first term on the RHS) of Equation (1) represents the deletion of particles of size v due to breakage into smaller fragments [4]. The birth term (second term on the RHS) of Equation (1) defines the addition of fragment particles with volume v due to breakage of particles with volume v', where $v < v'$. The breakage rate $S(v)$ and the breakage kernel $b(v, v')$ can be obtained by modelling and simulation or via experiments [1, 2].

A variety of methods can provide a solution for the PBE [5], such as the sectional method, the method of moments and Monte Carlo methods [1–3]. In this paper, the time-driven Monte Carlo (MC) approach has been implemented for solving the PBE for breakage on the CPU [3, 6] which uses serial processing, and on the GPU [1, 3], which facilitates parallel processing.

The modelling of breakage [6] of the particle is the process of making a model, which is a depiction of the structure and working of the particle breakage process [7]. A model is simple, but it represents similar to the real process. The main objective of the model of a process, for an analyst, is to estimate the outcome of variations in the process. Simulations of processes help to meet the specifications of a particular system, to reduce the chance of failure, to remove unexpected bottlenecks, to avoid overconsumption or underconsumption of resources and to optimize the performance of the process [7]. This paper focuses on: (1) Modelling and simulation of the breakage process serially and parallelly. (2) The differences between the GPU and CPU computational times and the speedup factor for a given breakage rate and breakage function.

The importance of this reported investigation is to provide awareness about parallel processing can save time and cost.

The remainder of this paper is organised as follows. In Section II we have presented the literature review of the Modelling and simulation of the Particle breakage using Monte Carlo methods, the serial and parallel processing differences are briefly summarized, according to different the hardware, software, architecture and the compilation of the algorithm. The discussion and the results are covered in Section III. The paper is concluded in Section IV and the future work highlighted in Section V.

II. Literature Review

A. Monte Carlo Simulation for Particle Breakage

MC methods (MC experiments) consist of a variety of computationally efficient algorithms that rely on recurring random sampling to obtain a numerical output. The MC method is widely used for modelling population balances (MC-PBs) [7]. The MC method has a discrete and stochastic nature [8, 9] and is suited to particle dynamics.

The MC simulation comprises of more than 1,000 iterations [10] or recalculations. During an MC simulation, the input probability distribution values are tested at random [11]. Every set of tests is called a recalculation or iteration, and the result of that test set is stored. Usually, Monte Carlo methods are classified by the time discretization system into event-driven Monte Carlo methods and time-driven Monte Carlo methods.

Event-driven MC [12] first estimates awaiting for time or time interval Δt among two consecutive events and then select the event randomly that occurs during time interval Δt.

Time-driven Monte Carlo [12] calculates a time step Δt, also known as a pre-specified time step, and considers all possible events that may happen within that time step.

The two most widely used sampling strategies for simulating breakage in Monte Carlo methods are the acceptance-rejection (AR) strategy [2, 10] and the inverse strategy [1, 12].

These sampling strategies have different ways of choosing the particle or SE to break randomly as shown in Fig. 1

Acceptance-rejection sampling is a more straightforward way of selecting the required particle: it randomly chooses a particle, calculates the breakage probability and checks to see whether or not to accept it. The Monte Carlo method is used to handle the particle breakage. Many attempts are often needed to find a suitable particle for breakage.

On the other hand, inverse sampling chooses a desirable particle by generating a random number between 0 and 1 and comparing it to a normalized value (R in [2, 3]) for the considered part of the PSD. Using the inverse scheme, one can always find a particle after a finite number of attempts (up to the total number of simulation particles (SEs)).

The MC method using AR sampling features inherent parallelism [2][13]. This is because the choices of a random particle (to be broken) via AR sampling are uncorrelated, and hence can be made in parallel.

This inherent feature of Monte Carlo AR sampling is easy to implement on a parallel architecture, such as the GPU [1, 2, 14]. GPUs can execute millions of lightweight threads in parallel and simultaneously.

B. Serial and Parallel Processing

Population balance equation processes, such as coagulation, breakage and nucleation, have been modelled on CPUs [8, 11] and GPUs [1, 2, 13]. It is worth emphasizing that an ideal grouping with high efficiency is important for PB-MC, because the increase in the number of simulation entries leads to an increase in the accuracy of the implemented process or algorithm, while the computing efficiency decreases. Computational power has increased over the last 10 years, but it is also important to increase computational efficiency for estimating particle dynamics. The MC simulation can be accelerated in two ways [15] as follows.

1) CPU parallel processing, via OpenMP (open multi-processing) and MPI (message passing interface) [15].

2) GPU parallel processing: processing via the GPU and CUDA [16].

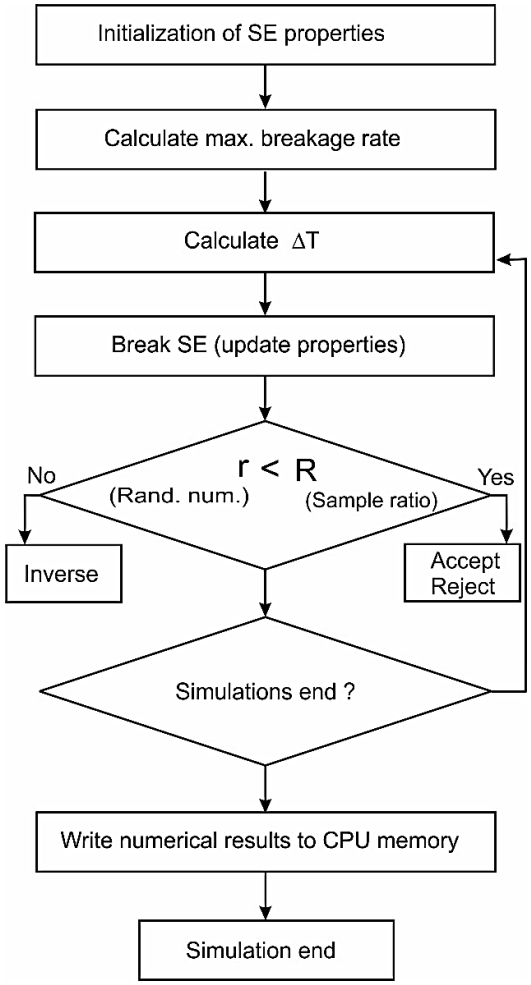


Fig. 1. Simulation Model.

In fact, the parallel computing of MC simulations utilizes more computer resources concurrently to decrease computational time. The CPU time increases linearly with computational complexity O(N) [1]. Monte Carlo acceptance-rejection sampling [16] has a significant effect on computational performance.

Recently, fast inverse and AR MC sampling on the GPU have been proposed [1, 2], to enhance the performance of MC methods for particle breakage.

Several nanotechnology applications have high inherent parallelism. The GPU is capable of high performance, as it supports a large number of cores, [16], and in many applications, GPU gets high performance. The clock speed of the serial processing [17] has driven the attention of the researchers to the parallel architecture,[16, 18–22] which is capable of providing tremendous computational efficiency [1, 2]. Usually, MC simulations require looping during the simulation of particle breakage events on CPU at the end of the simulation the computational cost increases [10]. The GPU and its advanced capabilities used for rendering of 3D games basically timeframe [23]. Now the GPU competencies are being coupled to accelerate computational workload by modelling of the complex and computationally expensive processes in different fields of scientific research.

1) Hardware: With regard to the hardware used for simulation, the CPU by Intel with four cores maximum performance, while the Nvidia GPU with 2,304 cores [24] can deliver 4,156 Gigaflops. The use of GPUs in high-performance computing also affects the computational cost. The GPU has two main characteristics [25] as follows: Table I. Allowing contact with the GPU's cores using CUDA programming. CUDA by Nvidia [23] is embedded in the standard C language. The function of a GPU is known as a kernel distribution function and is executed on the GPU cores [25]. 2. Achieving the distribution of the task via a kernel grid, where the kernel grid is split into blocks, with each block subdivided into several threads. The GPU has off-chip global memory and on-chip shared memory [26, 27]. This enables the programmer to select the type of memory used according to the requirement of the kernel [28].

2) Architecture: GPUs and CPUs were developed using different theories. CPUs can provide a prompt response time for individual tasks [29], whereas GPUs are built specifically for graphical applications and for rendering [27]. As stated above, the CPU used in modelling the breakage of a particle consists of four cores and a great deal of cache memory as shown in Fig. 2, which can deal with some application threads simultaneously.

The CPU memory is also known as host memory. The GPU, on the other hand, consists of hundreds of cores [14] and is capable of handling thousands of threads in parallel. The Nvidia graphics processing units are composed of streaming multiprocessors (SMs) [29].

Every SM contains a couple of cores. The GPUs have three types of memory: very fast and high-latency global memory, the on-chip low-latency shared the memory of each SM and the local private memory of each thread [29, 30]. The computational efficiency increases with more appropriate use of GPU memory. This parallel processing of the GPU can accelerate the algorithm by 60-70 times, compared with the serial processing of the CPU. The GPU is also more cost-efficient than an ordinary CPU. The general-purpose central processing unit is able to run several applications.

The graphics card cannot work alone, without CPU support. The GPU processes the task in parallel and the CPU controls it. The CPU invokes the task on the GPU, and the execution of the kernel is handled by the CUDA library, which runs on the central processing unit, while the GPU executes the kernel (functions). Both the CPU and the GPU work in parallel [29]. Here, the Peripheral Component Interconnect Express (PCIe) enables CPU-GPU data communication. The PCIe speed also limits the CPU-GPU data transfer time.

TABLE. I. HARDWARE USED FOR IMPLEMENTATION AND COMPRESSION OF THE BREAKAGE ALGORITHM WITH CUDA 8.0

Simulation Hardware		
Simulation Tool	CPU	GPU
Generation	4th	Kepler
Model	Intel Core i7-4790	GeForce GTX 780
Core	4	2,304
Clock Rate	3.60 GHz	1.006 GHz
Flops	44.24 Giga	4,156Giga

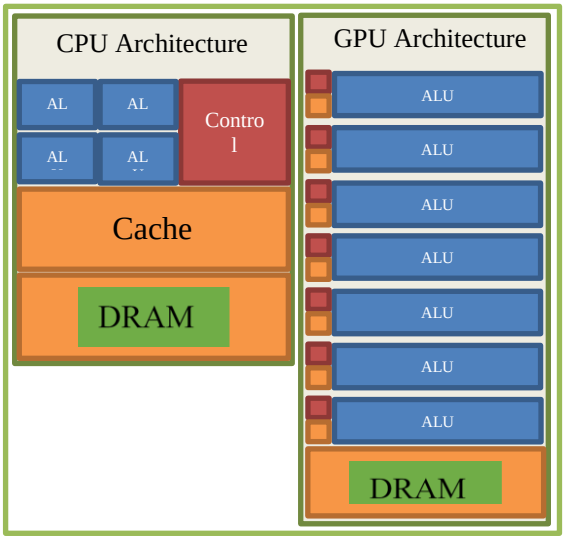


Fig. 2. The Architecture of the CPU and GPU.

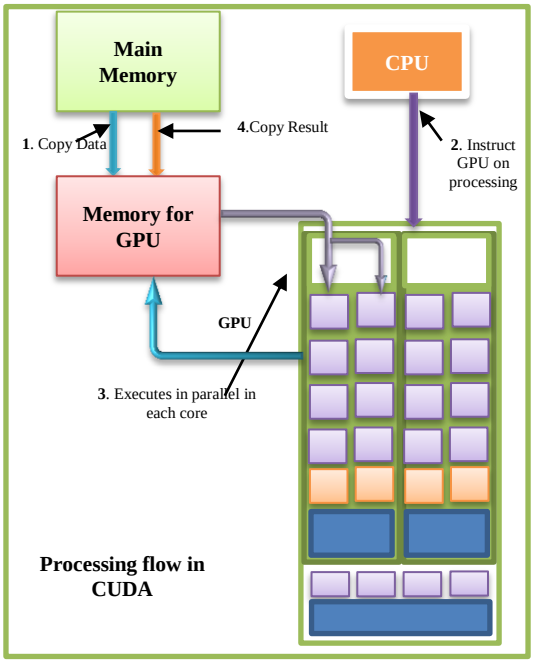


Fig. 3. The CUDA Processing Flow Copies Data from the main Memory to the GPU Memory and the CPU Instructs the GPU on the Process. The GPU Executes it with Each Core in Parallel, then Copies the Result from the GPU Memory to the main Memory.

3) CUDA and C programming: CUDA is a piece of software that deals with the GPU device, which can execute the computation in parallel. A GPU is a known device and it can execute lightweight threads in parallel through the CUDA programming as can be seen in Fig. 3. The GPU kernel, written in C, represents an extended programming language, compiling the device code and working on large quantities of data in parallel.

4) Compilation of the GPU and CPU code UDA: There are a few steps to compile the GPU CUDA program. Frist CUDA front end cudafe (cudafe.exe) splits the CUDA program into two parts: host and device code. The host code consists of a C/C++ program, and the device code uses GPU CUDA kernels. The host code part is compiled by the standard C compiler GCC and the device code is compiled using the CUDA compiler NVCC. The CUDA compiler NVCC converts intermediate code as in the type of assembly programming known as parallel thread execution (PTX). PTX is a low-level programming language used for GPUs by Nvidia as in Fig. 4. Furthermore, PTX code is interpreted into the binary code of the graphics processing unit cubin, which uses the ptxas compiler [30]. The compilation time also depends on the number of instructions and programs per kernel.

The increase in compilation and speed evaluation is directly proportional to the number of instructions [31] and programs per kernel.

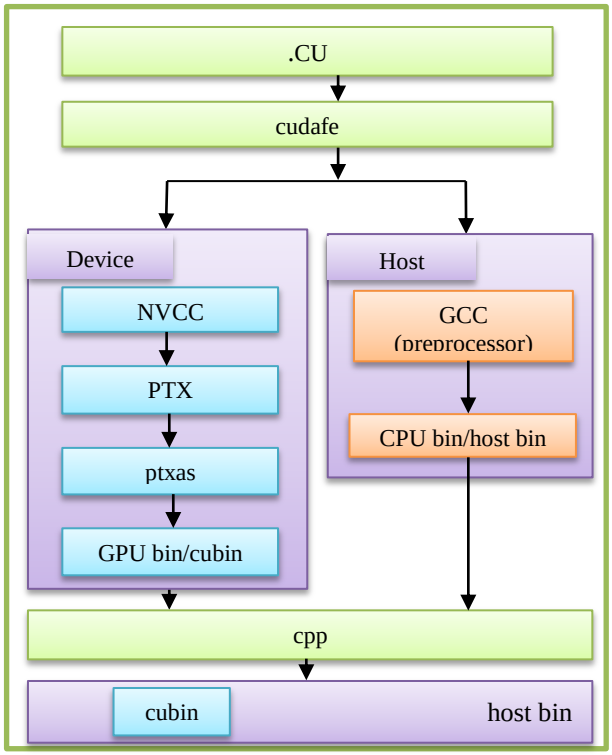


Fig. 4. CUDA Program Compilation.

III. Simulation Results and Discussion

A. Discussion

The process of breakage of particles has been simulated. Considering the performance on the GPU GeForce GTX 780, the most obvious and substantial outcome is the speedup factor of the breakage algorithm, compared to the performance on the CPU Intel Core i7-4790. The efficiency difference is greater than the difference between floating-point rates for the CPU and the GPU.

In this study, the particles are represented as simulation entries (SEs). For comparison, the results are obtained from calculations using the CPU and the GPU. A well-parallelized algorithm produced the same results as the sequential version.

The simulations were performed for various SEs on a 3.60-GHz i7 CPU and a GeForce GTX-780 GPU.

This detailed comparison of the computational efficiencies of the GPU and CPU can be used as guidance for choosing a hardware tool for the MC simulations. The results show the following.

The computational performance of GPU rendering is superior to that of CPU rendering for a single-cell scenario.

The acceptance-rejection sampling procedure is simple and easy to program. The GPU shows improved computational efficiency over the CPU.

The speedup ratio for the CPU and GPU demonstrates the difference between them. The speedup factor increases as the number of simulation entries increases. The GPU accelerates the rendering of complex processes, but the GPU acceleration power is limited by the quantity of input data allowed by the memory of the graphics card. However, this could be more than millions of SEs. For ideal performance, at least 100,000 particles are needed, with 500 simulations (iterations). The GPU requires full utilization in order to take advantage of the hardware's latency. It is not optimal to use the GPU for a low number of simulation entries, because most of the GPU time is spent on initializations for the calculations.

The computations are much faster on a GPU GeForce GTX 780 compared to a CPU. The GPU-based algorithm runs more than 65 times faster than our portable C implementation. As a result, there is an average speedup ratio of 65 compared to the scalar C++ code.

B. Results

1) Comparison of computational efficiency on CPU (serail) and GPU (Parallel): The computational efficiency of CPU- and GPU-based code has been compared. A difference in computing time can clearly be seen between the GPU and CPU, for the same settings. The GPU accelerates the rendering, as shown in Fig. 5. The computational time efficiency of the simulations depends on the number of SE settings.

The execution time depends on the number of simulations (iterations) for different SE settings and is calculated on the GPU. The GPU program performs the simulation for the breakage of a SE (particle), for various numbers of iterations. As shown in Fig. 5, the time required to simulate the process and compute the results for 500 simulations, for 10^5, 10^4, 10^3 and 10^2 SEs, is around 41, 4.1, 0.6 and 0.2s, respectively. If the number of simulations increases by 10 times, the computational time also increases by a factor of approximately 10, as predicted. The plot shows that for a small number of SEs, very less time is required to render and compute the results on the GPU. The computational time is directly proportional to the number of SEs and the number of simulations. For 150, 250, 350, 450 and 500 simulations, the time taken to render 10^5 SEs was 12.6, 20.8, 29, 37.3 and 41.4 seconds, respectively.

The computational time for the CPU-based algorithm executing the breakage process for various SE settings repeated for different numbers of simulations, is shown in Fig. 5. The time required to simulate and compute the results for 500 simulations, for 10^5, 10^4, 10^3 and 10^2 SEs, was around 2,600, 230, 22 and 2 s respectively. This clearly shows that the computational time increases by a factor of 10 with an increase in SEs by the same factor, as expected. The serial algorithm for a low number of SEs needs much less computational time to render and compute the results on the CPU and the computational time is directly proportional to the number of SEs.

Fig. 7 shows the computational time plotted logarithmically for different synchronization points of the simulations, for 10^5 SEs and 500 simulations (iterations or repetitions of a piece of code for statistical purposes). Each synchronization point (data export time point: after a certain time, the simulation computes the results and stores them in the CPU memory) is used to calculate the statistical results. Simple observation shows that the serial code (CPU) and parallel code (GPU) for the maximum number of SEs (10^5) take 2,600s and 41s, respectively.

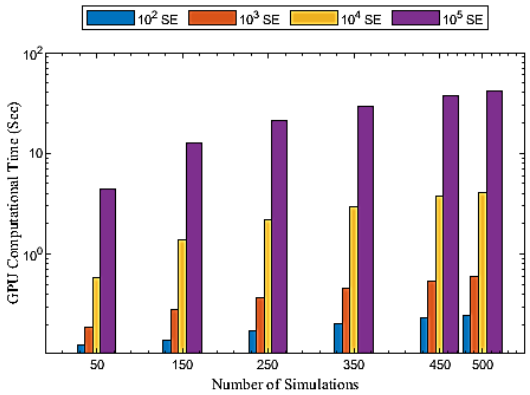


Fig. 5. GPU Program Computational Time for different SE Settings and different Numbers of Simulations.

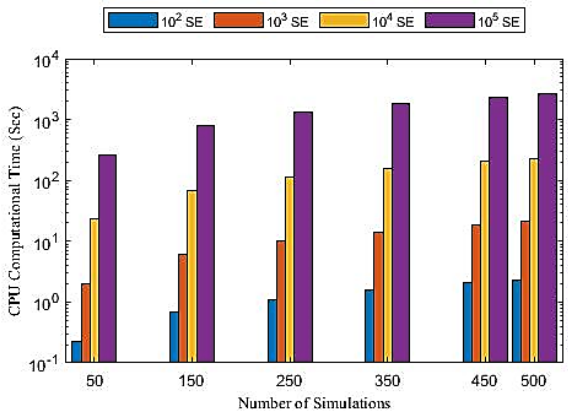


Fig. 6. CPU Program Computational Time for different SE Settings and different Numbers of Simulations.

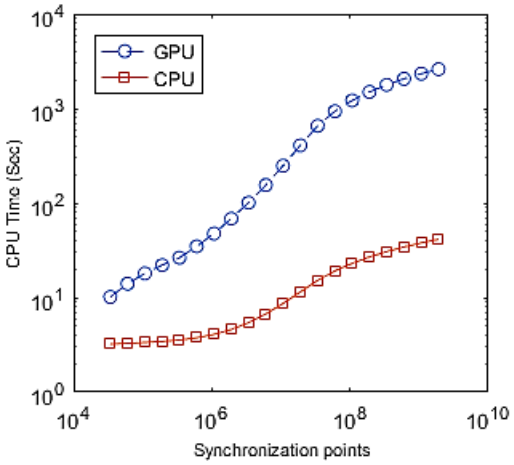


Fig. 7. Computational Time Synchronization Points Defined in the Simulations.

The clock register (clock ()) is used to obtain the time measurements as in equation (2):

$$\text{Computational time} = (\text{end\_clock}() - \text{start\_clock}()) \quad (2)$$

Simple observation shows that the serial code and parallel code for the maximum number of SEs (10^5) take 2,600 s and 41 s respectively. The performance difference between the CPU and GPU computational efficiencies is better illustrated in the speedup plot. To determine the time in seconds (s), the function clock () is divided by the CLOCKS_PER_SEC macro.

$$\text{Computational time} = \frac{(\text{end\_clock}() - \text{start\_clock}())}{\text{CLOCKS\_PER\_SEC}} \quad (3)$$

The macro CLOCKS_PER_SEC represents the number of clock ticks per second calculated as in equation (3).

2) Speedup factor: The speedup factor accelerating the particle dynamics simulation shows that the speed is higher than for the algorithm implemented in serial programming (CPU). The most common method for assessing the advantage of using the GPU is to calculate the computational times for the GPU and CPU [30, 32]. Speedup (S) is the computational time of the serial code (T_{CPU}) divided by the computational time of the parallel code (T_{GPU}), where both are computing the same result, i.e.,

$$S = \frac{T_{CPU}}{T_{GPU}} \quad (4)$$

To make a comparison, Fig. 5 and Fig. 6 show the CPU time for the CPU and GPU as a function of the simulation time. Simple examination shows that the CPU simulation is much slower than the GPU simulation. Usually, the speedup factor is calculated as in equation (4) and is plotted on the y-axis and the SEs or numbers of iterations (simulations) on the x-axis, as shown in Fig. 8. The plot shows that the GPU processing was 65 times faster than CPU processing. The GPU operated up to its maximum processing limit and achieved a speedup factor of 65. The GPU cannot operate any faster than this, for the considered particle breakage algorithm [2]. It does not matter if the number of simulation entries increases. For lower numbers of SEs, the GPU process is much faster than the CPU computations. To use the maximum power of the GPU, we used the maximum number of SEs (10^5), with 500 simulations. Fig. 8 shows the usage of the maximum computational power of the GPU. It is clear from Fig. 9 that the GPU reaches a certain limit, and the speedup factor is approximately constant after that limit. The number of SEs and the number of simulations no longer affects the efficiency of the GPU.

The speedup factor shows that GPU processing is much faster than for a normal CPU. We do, in fact, achieve a considerable speedup over the CPU case, although we should expect significant speedups (30 to 65 times), according to the GPU acceleration computation.

Fig. 8 shows the real speedup achieved using the graphics processing unit. There is a significant difference in the speedup factor with variations in the number of SEs. It increases to certain limit, depending on the memory of the graphics card. The speedup ratio, compared to the CPU for the GTX card, clearly shows the performance change between the CPU and the GPU.

The speedup plot depends on the SEs. There are no differences in computational efficiencies. If we consider only 1,000 SEs and 10 simulations (iterations) the total number of threads to be launched on the GPU is 10,000 (SEs x iterations). Alternatively, or we take 10000 SEs and 1 simulation. The computational time is calculated and found as the same time consumed by both settings. For the most significant number of SEs and simulations, a speedup factor of up to 65 was obtained.

3) Clock cycles: The clock cycle is the speed of the processor. It is the time taken to complete a full process [33].One can predict this from the speed of the processor [34]. The i7 CPU with a speed of 3.6 GHz is used as a hardware tool in this study. The speed of a processor is measured in Gigahertz (GHz).

This means that this CPU processor can perform 3,600,000,000 clock cycles per second. Fig. 9 shows that for all input SEs, GPUs show their advantages over CPUs, leveraging their parallel computing abilities. Specifically, for 10^5 input SEs and 500 simulations, 4.16×10^{10} cycles are needed by the GPU, as shown in Fig. 9, compared to 9.39×10^{12} cycles for the CPU. The major overhead for GPUs and CPUs arises from executing instructions, which depend on memory access. The serial and parallel algorithm clock cycles are summarized in Table II.

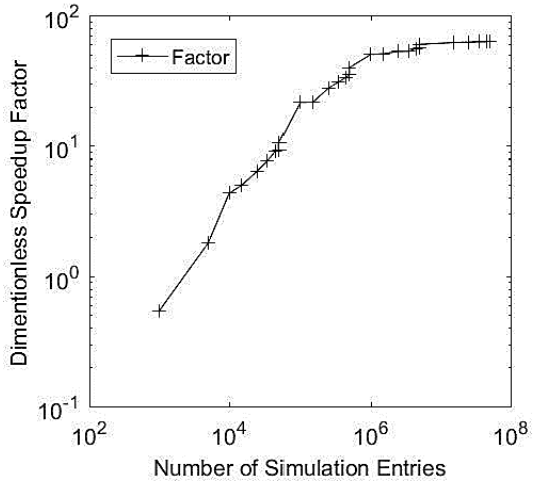


Fig. 8. The Dimensionless Speedup Factor for the Total SEs used for the Simulations. The Total Number of Simulation Entries was Simulated only once. No Iterations were Performed. Dimensionless Speedup Factor for the different SE and Iteration Settings. The Speedup is Plotted on the y-Axis and the x-Axis shows the Number of Iterations for the SEs.

TABLE. II. CLOCK CYCLES OF THE SERIAL AND PARALLEL IMPLEMENTED ALGORITHM

Iterations	CPU Clock Cycles				GPU Clock Cycles			
	10^2 SEs	10^3 SE	10^4 SEs	10^5 SEs	10^2 SEs	10^3 SEs	10^4 SEs	10^5 SEs
	10^{12}				10^{10}			
50	0.0008	0.0071	0.0823	0.9472	0.0126	0.0190	0.0578	0.4415
150	0.0025	0.0217	0.2467	2.8080	0.0139	0.0282	0.1379	1.2716
250	0.0039	0.0362	0.4115	4.6980	0.0173	0.0366	0.2180	2.0995
350	0.0056	0.0508	0.5638	6.5376	0.0206	0.0460	0.2977	2.9244
450	0.0075	0.0653	0.7513	8.3916	0.0231	0.0544	0.3768	3.7594
500	0.0081	0.0757	0.8273	9.3924	0.0243	0.0604	0.4137	4.1628

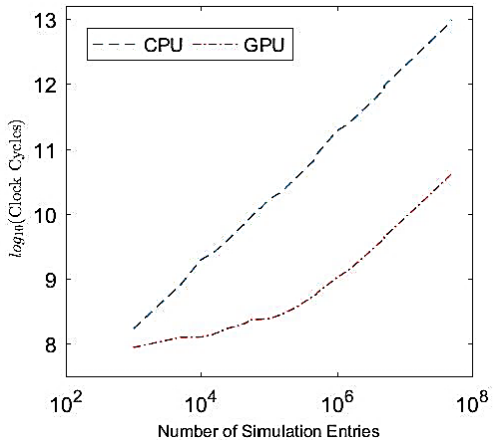


Fig. 9. The Clock Cycle for the Total SEs (SEs * NoSim = Number of Simulation Entries) used to Simulate the Breakage Process. Execution Cycles of the Two Versions of Hardware Tools used. The y-Axis Shows Clock Cycles for the GPU and CPU for Various SEs and the x-Axis shows the Number of Simulations. A Log Scale is used.

IV. CONCLUSIONS

We provided here a new background for modelling particle breakage using a graphics processing unit as a parallel environment and a CPU for serial processing. We performed simulations of the particle breakage process with different input SEs and simulation settings on a 3.60-GHz Intel Core i7 CPU and a GeForce GTX-780 GPU.

It was found that the GPU-based rendering is much faster than the CPU- based rendering. GPU saves time and produces fast results. Significant speedup was achieved for the graphics processor over the central processing unit, for the reported breakage algorithm (BP-PBE).

In the case of a stochastic particle model, the typical speedup is around 60-65 times, depending on the particle numbers (SEs) considered. Comparisons of CPU and GPU simulation results show that there are differences due to the different computational platforms. Significant differences between GPU and CPU results can be observed for the case of a large number of simulation entries. However, the monitored differences in the computational time for the simulation of particles on the GPU and CPU are of the order of the characteristic CPU time. This demonstrates that the GPU is a proficient and efficient tool for rendering parallel processes, such as modelling the breakage of particles.

The objective of this paper is to show the advantages of using the GPU for rendering the breakage of particles. A fast and efficient breakage algorithm is developed, and the performance improvement of the parallel algorithm on the GPU is observed and compared with the efficiency of CPU implementations. This study examined the computational efficiency of single-cell implementations on both the GPU and the CPU.

V. FUTURE WORK

In future, the authors are interested to implement breakage-coagulation algorithms to evaluate the efficiencies of various algorithms. Moreover, multiple-cell implementations could also be optimized to improve the performance of the breakage process. Each cell will have a different breakage rate, decreasing with time. Small particles take more time to break, and it can easily be predicted that if the breakage rate is reduced, then the computational time will increase, and the speedup factor will also be affected.

ACKNOWLEDGMENT

The support of the Quaid-E-Awam University, Nawabshah Sindh (Pakistan) by a scholarship under the Faculty Development Program is greatly acknowledged.

REFERENCES

[1] J. Devi and F. E. Kruis, “A fast Monte Carlo GPU based algorithm for particle breakage,” IEEE-CODIT 17, pp. 784–789, 2017.

[2] G. Kotalczyk, J. Devi, and F. E. Kruis, “A time-driven constant-number Monte Carlo method for the GPU-simulation of particle breakage based on weighted simulation particles,” Powder Technology, vol. 317, pp. 417–429, 2017.

[3] J. Devi, G. Kotalczyk, and F. E. Kruis, “Accuracy control in Monte Carlo simulations of particle breakage,” IJMIC, vol. 31, no. 3, p. 278, 2019.

[4] P. J. Hill and K. M. Ng, “New discretization procedure for the breakage equation,” AIChE J., vol. 41 No. 5, pp. 1204–1216, 1995.

[5] J. Kumar, G. Warnecke, M. Peglow, and S. Heinrich, “Comparison of numerical methods for solving population balance equations incorporating aggregation and breakage,” Powder Technology, vol. 189, no. 2, pp. 218–229, 2009.

[6] K. F. Lee, R. I.A. Patterson, W. Wagner, and M. Kraft, “Stochastic weighted particle methods for population balance equations with coagulation, fragmentation and spatial inhomogeneity,” Journal of Computational Physics, vol. 303, pp. 1–18, 2015.

[7] W. K. Chan et al., Eds., Modeling as the practice of representation: Proceedings of the 2017 Winter Simulation Conference, December 3-6, 2017, Las Vegas, NV. Piscataway, NJ, Madison, WI: IEEE; Omnipress, 2017.

[8] H. Zhao, A. Maisels, T. Matsoukas, and C. Zheng, "Analysis of four Monte Carlo methods for the solution of population balances in dispersed systems," Powder Technology, vol. 173, no. 1, pp. 38–50, 2007.

[9] F. E. Kruis, J. Wei, T. van der Zwaag, and S. Haep, "Computational fluid dynamics based stochastic aerosol modeling: Combination of a cell-based weighted random walk method and a constant-number Monte-Carlo method for aerosol dynamics," Chemical Engineering Science, vol. 70, pp. 109–120, 2012.

[10] J. Wei and F. E. Kruis, "A GPU-based parallelized Monte-Carlo method for particle coagulation using an acceptance–rejection strategy," Chemical Engineering Science, vol. 104, pp. 451–459, 2013.

[11] Z. Xu, H. Zhao, and C. Zheng, "Fast Monte Carlo simulation for particle coagulation in population balance," Journal of Aerosol Science, vol. 74, pp. 11–25, 2014.

[12] G. Kotalczyk and F. E. Kruis, "Fractional Monte Carlo time steps for the simulation of coagulation for parallelized flowsheet simulations," Chemical Engineering Research and Design, vol. 136, pp. 71–82, 2018.

[13] G. Kotalczyk and F. E. Kruis, "A Monte Carlo method for the simulation of coagulation and nucleation based on weighted particles and the concepts of stochastic resolution and merging," Journal of Computational Physics, vol. 340, pp. 276–296, 2017.

[14] NVIDIA Corporation (2013), "NVIDIA GeForce GTX 780 Specifications," https://www.geforce.com/hardware/desktop-gpus/geforce-gtx-780/specifications, 2013.

[15] Kruis, F.E: Zhao, H, J. Wei, and C. & Zheng, "A parallelized population balance-Monte Carlo method for diffusion and coagulation of nanoparticles," pp. 26–29, 2010.

[16] J. Wei and F. E. Kruis, "GPU-accelerated Monte Carlo simulation of particle coagulation based on the inverse method," Journal of Computational Physics, vol. 249, pp. 67–79, 2013.

[17] M. Woźniak, K. Kuźnik, M. Paszyński, V. M. Calo, and D. Pardo, "Computational cost estimates for parallel shared memory isogeometric multi-frontal solvers," Computers & Mathematics with Applications, vol. 67, no. 10, pp. 1864–1883, 2014.

[18] J. Wei, J. Wang, Q. H. Wu, J. Chen, and N. Jia, "Multisegment pulverised coal mill model and online implementation for condition monitoring," International Journal of Modelling, Identification and Control, vol. 1, no. 3, pp. 206–214, 2006.

[19] J. Wei, "A parallel Monte Carlo method for population balance modeling of particulate processes using bookkeeping strategy," Physica A: Statistical Mechanics and its Applications, vol. 402, pp. 186–197, 2014.

[20] Z. Xu, H. Zhao, and C. Zheng, "Accelerating population balance-Monte Carlo simulation for coagulation dynamics from the Markov jump model, stochastic algorithm and GPU parallel computing," Journal of Computational Physics, vol. 281, pp. 844–863, 2015.

[21] S. Hong, T. Oguntebi, and K. Olukotun, Eds., Efficient Parallel Graph Exploration on Multi-Core CPU and GPU: IEEE, Oct. 2011.

[22] S. Ryoo et al., Optimization principles and application performance evaluation of a multithreaded GPU using CUDA. New York, NY: ACM, 2008.

[23] C. Gregg and K. Hazelwood, "Where is the data? Why you cannot debate CPU vs. GPU performance without the answer," Performance Analysis of Systems and Software (ISPASS), 2011 IEEE International Symposium ,Austin, TX, USA, pp. 134–144, 2011.

[24] J. P. Harvey, GPU acceleration of object classification algorithms using NVIDIA CUDA, 2009.

[25] V. W. Lee et al., "Debunking the 100X GPU vs. CPU myth: An Evaluation of Throughput Computing on CPU and GPU," SIGARCH Comput. Archit. News, vol. 38, no. 3, p. 451, 2010.

[26] X. Mei and X. Chu, "Dissecting GPU Memory Hierarchy Through Microbenchmarking," IEEE Trans. Parallel Distrib. Syst., vol. 28, no. 1, pp. 72–86, 2017.

[27] J. Nickolls and W. J. Dally, "The GPU Computing Era," IEEE Micro, vol. 30, no. 2, pp. 56–69, 2010.

[28] J. Wei, "A Fast Monte Carlo Method Based on an Acceptance-Rejection Scheme for Particle Coagulation," Aerosol Air Qual. Res., 2013.

[29] F. Molnár, T. Szakály, R. Mészáros, and I. Lagzi, "Air pollution modelling using a Graphics Processing Unit with CUDA," Computer Physics Communications, vol. 181, no. 1, pp. 105–112, 2010.

[30] C. P. Da Silva, D. M. Dias, C. Bentes, M. A. Pacheco, and L. F. Cupertino, "Evolving GPU machine code," Journal of Machine Learning Research, vol. 16, pp. 673–712, 2015.

[31] T. E. Lewis and G. D. Magoulas, Eds., Identifying similarities in TMBL programs with alignment to quicken their compilation for GPUs. New York, NY, USA ©2011: ACM, 2011.

[32] M. S. Friedrichs et al., "Accelerating molecular dynamic simulation on graphics processing units," (eng), Journal of computational chemistry, vol. 30, no. 6, pp. 864–872, 2009.

[33] H. McVeigh, "Factors influencing the utilisation of e-learning in post-registration nursing students," (eng), Nurse education today, vol. 29, no. 1, pp. 91–99, 2009.

[34] S. Abdel-Hafeez and A. Gordon-Ross, "An Efficient O(N) Comparison-Free Sorting Algorithm," IEEE Trans. VLSI Syst., vol. 25, no. 6, pp. 1930–1942, 2017.

Developing Lexicon-based Algorithms and Sentiment Lexicon for Sentiment Analysis of Saudi Dialect Tweets

Waleed Al-Ghaith
Department of Information Systems
Shaqra University
Imam Muhammad Ibn Saud Islamic University
Riyadh, Saudi Arabia

***Abstract*—Majority of studies on sentiment analysis field, specifically Arabic lexicon-based approach, are focused on doing preprocessing methods on targeted dataset text or collected textual data from Twitter (Twitter dataset) rather than dealing with lexicon itself. This study proposes a new method, we constraint firstly on building a new sentiment lexicon with reasonable number of words and then doing adequate preprocessing methods on the lexicon's words in addition to the (Twitter dataset). The study presents Saudi Dialect Sentiment lexicon called SaudiSentiPlus contains 7139 words which mostly generated from Saudi tweets and other dictionaries. Moreover, this study also presents two lexicon-based algorithms for Saudi dialect to deal with (prefixes and suffixes) letters in order to increase performance of proposed Saudi dialect lexicon. The experiment which has been conducted in this study to evaluate the performance of SaudiSentiPlus comprises four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). The results, show that accuracy of SaudiSentiPlus with the two lexicon-based algorithms reached to 81%.**

Keywords*—Sentiment analysis; opinion mining; lexicon-based; Arabic text mining; Saudi Arabia

I. Introduction

A Social Network Site (SNS) is a platform enables people to share their opinions on any issue and to build social relations with individuals within and beyond their social circle [1].

Twitter as a one of the most popular SNSs that has been growing rapidly in recent years. Twitter's users increased by more than 500% since 2009 [1]. Twitter's users express their feeling, opinions or spreads news or facts about 200 billion times annually via their tweets, 500 million of them per day, 350,000 per minute, and 6000 per second [2].

In 2014, total number of active Twitter users in the Arab world reached 5,797,500 users and the country with the highest number of active Twitter users in the Arab region is Saudi Arabia with 2.4 million users, accounting for over 40% of all active Twitter users in the Arab region. The estimated number of tweets produced by Twitter users in the Arab world in March 2014 was 533,165,900 tweets, an average of 17,198,900 tweets per day [3] (see Fig. 1).

Currently, in 2019, Saudi Arabia was ranked the fourth in the world with around 10 million active users after the United States, Japan and the United Kingdom (see Fig. 2) [4].

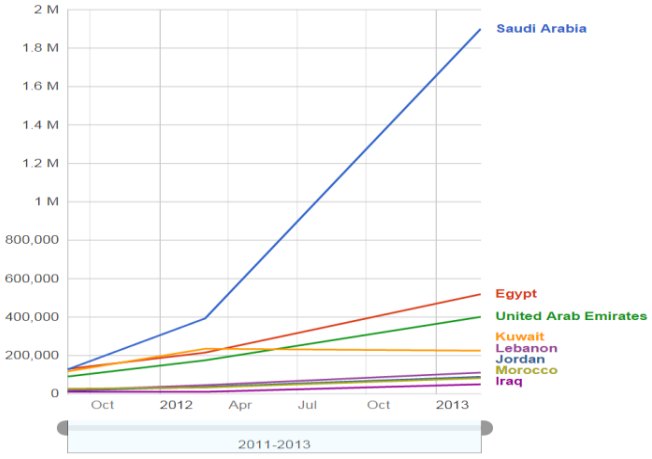


Fig. 1. Number of Active Twitter users in the Middle East [3].

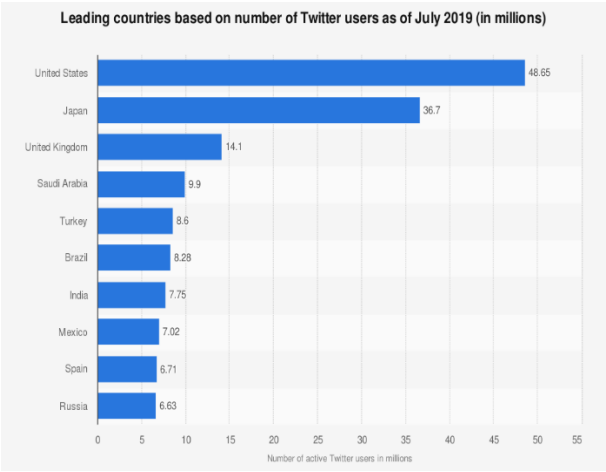


Fig. 2. Number of Active Twitter users in the World [4]..

This means that a massive amount of contents and opinions of Saudis toward phenomenon, topic, institution or individuals can be obtained and studied via twitter. This content can be either objective contents (e.g. news, facts) or subjective

contents such as (opinions or sentiments about entities). Opinions mining or sometimes alternatively mentioned as sentiment analysis is the research discipline which aims to analyze individuals' sentiments or opinions toward entities such as topics, people, issues, organizations or events [5] and classifying them as negative, positive or neutral opinions.

The Saudis speak and write in Arabic language and few of them are fluent in English. Arabic language is the fastest growing language on the web (8,917.3 %), it is ranked the fourth among languages on the web as illustrated in Table I [6].

Arabic language has many variants, however we can categorized it to three categories. The first is the Qur'an language which is classical Arabic; the second is Modern Standard Arabic (MSA) which it used in formal speech and writing. The third is informal or dialectical Arabic. Dialectical Arabic refers to all oral diversities spoken in daily communication for 27 Arabic countries and from one area of the same country to another [7].

According to Darwish and Magdy [8] Arabic social media's users tend to use Arabic dialects online rather than MSA. Likewise, Saudis use their colloquial language in social media and in Twitter in particular, which makes study their opinions or doing sentiment analysis based on their tweets a challenging task. In social media, Arabic colloquial or dialect are changeable and has word elongations with nonstandard spellings. Consequently, doing sentiment analysis based on standard formal-dependent lexicon is inefficient since that it will be unable to capture colloquial or dialect language in social media text. Thus, there is a need to develop another efficient method considering create a dialect-dependent lexicon for sentiment analysis of social media.

In this study, we think beyond of the box, we constraint firstly on building a new sentiment lexicon with reasonable number of words and phrases, and then conducting adequate preprocessing methods on the lexicon's words and phrases in addition to the (Twitter dataset). To the best of our knowledge, no such effort (doing preprocessing methods on the lexicon and collected dataset) has been made in prior studies. This study presents Saudi Dialect Sentiment lexicon (SaudiSentiPlus) contains 7139 words and can be used for sentiment analysis of Saudi dialect tweets. Moreover, in this paper we propose a new method based on presenting two lexicon based algorithms to deal with (prefixes and suffixes) letters of lexicon's words and phrases. This new method has a positive significant effect on increasing the performance or accuracy of (SaudiSentiPlus) lexicon. We evaluated the performance of SaudiSentiPlus through four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral) as presented in evaluation section.

Next section presents the proposed methodology in details. Followed by evaluation, results and discussion, and then we conclude this study in Section 5.

TABLE. I. TOP TEN LANGUAGES USED IN THE WEB

Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	**World Population for this Language (2019 Estimate)**	**Internet Users by Language**	**Internet Penetration (% Population)**	**Internet Users Growth (2000 - 2019)**	**Internet Users % of World (Participation)**
English	1,485,300,217	1,105,919,154	74.5 %	685.7 %	25.2 %
Chinese	1,457,821,239	863,230,794	59.2 %	2,572.3 %	19.3 %
Spanish	520,777,464	344,448,932	66.1 %	1,425.8 %	7.9 %
Arabic	**444,016,517**	**226,595,470**	**51.0 %**	**8,917.3 %**	**5.2 %**
Portuguese	289,923,583	171,583,004	59.2 %	2,164.8 %	3.9 %
Indonesian / Malaysian	302,430,273	169,685,798	56.1 %	2,861.4 %	3.9 %
French	422,308,112	144,695,288	34.3 %	1,106.0 %	3.3 %
Japanese	126,854,745	118,626,672	93.5 %	152.0 %	2.7 %
Russian	143,895,551	109,552,842	76.1 %	3,434.0 %	2.5 %
German	97,025,201	92,304,792	95.1 %	235.4 %	2.1 %
TOP 10 LANGUAGES	5,193,327,701	3,346,642,747	64.4 %	1,123.0 %	76.3 %
Rest of the Languages	2,522,895,508	1,039,842,794	41.2 %	1,090.4 %	23.7 %
WORLD TOTAL	7,716,223,209	4,386,485,541	56.8 %	1,115.1 %	100.0 %

Source: [6]

II. Proposed Methodology

Variant methodologies were wildly adapted among researchers to produce sentiment lexicons [9, 10], however, three approaches of them are broadly accepted in Arabic lexicon construction process. The first approach is building sentiment lexicon by taking words from Arabic dictionaries or from other sentiment lexicon and divided these words based on their polarities (see [11], [12]). The second approach is based on translation of English lexicons (see [13]). The third approach, based on selecting seed sentiment words and then finding the words that occur in conjunction with the seed words (see [14]).

In this study, we started with building the study's sentiment lexicon, firstly, by using the second approach which is automatic translation of English sentiment lexicons that already created by two prior studies [15] and [16]. Then, in order to enrich the study's sentiment lexicon with Saudi dialect words, we manually extracted all the sentiment Saudi dialect words from the twitter data (datasets). This approach partly inlines with the third approach.

Finally, more words (4431 words) have been taken from Saudi dialect sentiment lexicon (SauDiSenti) for sentiment analysis of Saudi dialect tweets [17] and next we deleted the repeated words and divided all these words based on their polarities. This approach is consistent with the first approach. To determinate the polarity, two-way classification (positive or negative) on the datasets has been adopted. We asked three annotators to classify the words manually. All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. If there is any disagreement among the three annotators, we solved it by voting. We called the study lexicon SaudiSentiPlus, it contains (7139 words).

A. Preprocessing

Applying sentiment analysis directly to collected textual data from the Twitter (Twitter dataset) could lead to inaccurate outcomes [18]; thus, collected textual data or (Twitter dataset) need to be prepared for another processing procedure. This called data preprocessing which means transform the collected textual data or dataset into a format that be more adequate to the purpose of the study. Applying preprocessing techniques before doing the sentiment analysis processes is highly recommended in many Arabic sentiment analysis studies, particularly in the Arabic dialectal dataset because it is commonly written in an unstructured shape [19, 20, 5, 11].

Thus, in this study, to enhance sentiment analysis results for the collected Twitter dataset, we applied some of preprocessing techniques which are as follows:

1) Tweet cleaning. This step to remove irrelevant data such as user names, Twitter characters, URLs and all non-Arabic letters.

2) Elimination of redundant letters. In a moment of emotion, some twitter's users tend to repeat some word's letters when they want to emphasize something such as Gooooooal, COOOOL, WOOOW, In the same way, it happens to Arab or Saudi twitters for example "أليييييييييم" or "رووووعة".

B. The Proposed Algorithms

In order to increase performance of proposed Saudi dialect lexicon (SaudiSentiPlus), we developed two lexicon based algorithms to deal with (prefixes and suffixes) letters of the lexicon's words (see Fig. 3 and Fig. 4).

Due to that Saudi dialect words originally and mostly are extracted from Arabic language words and Arabic language is a morpho¬logical language and their words might be varied depending on the presence and position of some well-known letters in a word. Moreover, some of these letters come at the beginning (prefixes) or end (suffixes) of a word. Furthermore, these letters also have different shapes depending on their word appearance in the text or context. For instance, (A) letter in Arabic (ا) which pronounce (Alif) might be drawn or written in four ways (آ ، إ ، أ ، ا). This has its effects in (The) in Arabic which is a combination of two letters "ا" and "ل" to give us in its turn (The) in four forms (ال ، أل ، إل ، آل) [21].

And because this letters is frequently repeated in most Saudi dialect words, we have counted and confined most of these repeated letters (prefixes) and (suffixes). Prefix letters contain ("ال", "آل", "إل", "أل") and Suffix letters are ("ة", "ه", "ها", "ون", "ين", "هم", "هن", "وهن", "وهم", "نهم, "نهن", "ني"). We applied the first algorithm (see Fig. 3) to remove these letters (prefixes and suffixes) from most of lexicon words and saved all these words in new lexicon with words have no (prefix or suffix) letters. Next step we applied the second algorithm (see Fig. 4) which used the new lexicon with words have no (prefix or suffix) letters. The second algorithm used words from new lexicon and added (prefix or suffix) letters (one after one) step by step on these words and compared it to the texts taken from Twitter. In other words, we applied the algorithm to each word to give us all possible options by adding firstly (prefixes) alone to the same word one by one and then compared it one by one with all Twitter dataset's text and then added (suffixes) to the same word and compared them (one by one) with all dataset's text again and then added some of (prefixes and suffixes) together one by one to the same word and compared them with the all dataset's text and this has been repeated until last word in the new SaudiSentiPlus lexicon (see Fig. 4) . For example, the word "غبي" which means "stupid for male" when we applied light stemming algorithm the chances of finding or matching this word against the text taken from Twitter is increased. Since that this algorithm, for instance, will add the letter "ة" as a suffix to the word "غبي" to give us new word "غبية" which means "stupid for female".

III. Evaluation

To evaluate the effectiveness of SaudiSentiPlus, we compared it with SauDiSenti. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral). All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. They labeled 300 tweets for each classification.

These tweets have been used to evaluate the lexicon's accuracy for SaudiSentiPlus and SauDiSenti. Four of the most

widely used accuracy measures in the literature are utilized ([17, 13]). They are precision (P), recall (R), F measure (F), and accuracy (Acc) and their mathematical equations are as follow:

Precision (P) = TP / (TP+ FP)

Recall (R) = TP / (TP + FN)

Accuracy (Acc) = (TP + TN) / (TP + FP +TN +FN)

F-Score (F) = 2TP / (2TP + FP + FN)

Where TP or True Positive indicates to number of tweets that are correctly predicted as a positive, TN or True Negative are number of tweets that are correctly predicted as a negative, FP or False Positive indicates to number of tweets that are incorrectly predicted as a positive, FN or False Negative are number of tweets that are incorrectly predicted as a negative.

IV. Results and Discussion

The purpose of this experiment is to study the effect of increasing the size of the lexicon and to find whether there is any effect when we applied the first (see Fig. 3) and the second algorithms (see Fig. 4). As aforementioned above, these two lexicon based algorithms were developed to deal with (prefixes and suffixes) letters in order to increase performance of proposed Saudi dialect lexicon.

Table II illustrates the performance results of the experiment. Better accuracy (74%) has been achieved when the lexicon size is increased (from 4431 words to 7138 words). Moreover, accuracy has been increased to reach (81%) when we applied the two algorithms with the lexicon-based approach. Table II lists the precision, recall, accuracy, and F-Score results of the experiment.

As aforementioned, the lexicon construction accomplished through four phases. In the first phase, the lexicon was at its smallest size with 4554 words taken from automatic translation of English sentiment lexicons that already created by two prior studies [15] and [16] and other more sentiment Saudi dialect words which were manually extracted from the twitter data (datasets). The lexicon (SaudiSentiPlus 1) performance or its accuracy reached 61% which is better than 54% of SauDiSenti with its 4431 words [17]. In the second phase, or (SaudiSentiPlus 2) as shown in Table II, no new words have been added to the lexicon however we applied the two lexicon based algorithms on the lexicon (SaudiSentiPlus 1) to yield better accuracy (68%).

In the third phase, more words (4431 words) have been taken from Saudi dialect sentiment lexicon (SauDiSenti) for sentiment analysis of Saudi dialect tweets [17] and next we deleted the repeated words and divided all these words based on their polarities to reach to 7139 words. In this stage (see SaudiSentiPlus 3 in Table II) the accuracy has been enhanced to reach to (74%). Finally, we noticed that accuracy has been increased to reach (81%) when we applied the two algorithms with the lexicon-based approach (see SaudiSentiPlus 4 in the Table II).

Input: lexicon L

Output: new lexicon NL with words have no (prefix or suffix) letters

1. Create array of new lexicon's words NLW
2. initialize Prefix = ("آل", "إل", "أل", "ال")
3. initialize Suffix = ("نهن", "ني", "ه", "ة", "وا", "ون", "ين", "هم", "هن", "وهن", "وهم", "تهم")
4. **for** each w ∈ L **do**
5. **if** w contains any (prefix or suffix) letters **then**
6. **Remove** (prefix or suffix) letters
7. **Add** new word with no (prefix or suffix) letters to the NLW array
8. **else**
9. **Add** w to NLW array
10. **end if**
11. **end for**
12. **Save** NLW array in NL **file**

Fig. 3. Algorithm#1 to Remove (Prefixes and Suffixes) from most of Lexicon Words.

```
Input: new lexicon NL, set of texts T
Output: Sentiment for each text t ∈ T
initialize Prefix = ("ال", "أل", "إل", "آل")
initialize Suffix = ("ه", "ة", "وا", "ون", "ين", "هم", "هن", "وهن", "وهم", "نهم", "نهن", "ني")
Create a data structure to hold the lexicon. lexicon = dict()
# In this study, we used Python diction. The key of the dictionary was the word
# and the value was the word's score.
for each w ∈ NL do
    lexicon[w] = polarity of the same word (w)
 end for
Create a data structure to hold the tweet's text. tweet = dict()
for each t ∈ T do
   initialize score to 0
   Clean t (remove any non-Arabic letters from the text)
   for each w ∈ NL do
      initialize foundMatch to false
      if w = any word in the text t then
            foundMatch is true
      else if w with any letter of Prefix = any word in the text t then
            foundMatch is true
      else if w with any letter of Suffix = any word in the text t then
            foundMatch is true
      else if w with any letter of Prefix and Suffix = any word in the text t then
            foundMatch is true
      end if
      if foundMatch then
            score = score + int(lexicon[w])
       end if
    end for
   tweet['text'] = t
   tweet['score'] = score
   if (score > 0) then
      tweet['sentiment'] = 'positive'
      num_pos += 1
   else if (score < 0) then
      tweet['sentiment'] = 'negative'
      num_neg += 1
   Else:
      tweet['sentiment'] = 'neutral'
      num_neu += 1
    End if
End for
```

Fig. 4. Algorithm#2 to Increase the Chances of Finding Saudi Dialect Words.

TABLE. II. PERFORMANCE OF THE SAUDISENTIPLUS LEXICON COMPARED WITH SAUDISENTI

Lexicon	Method	P	R	Acc	F
SauDiSenti (4431 w)	lexicon-based	55%	53%	54%	54%
SaudiSentiPlus 1 (4554 w)	lexicon-based	61%	59%	61%	60%
SaudiSentiPlus 2 (4554 w)	lexicon-based + Algorithms	67%	68%	68%	67%
SaudiSentiPlus 3 (7139 w)	lexicon-based	73%	73%	74%	74%
SaudiSentiPlus 4 (7139 w)	lexicon-based + Algorithms	81%	82%	81%	80%

V. CONCLUSION

Majority of researchers on sentiment analysis field, specifically Arabic lexicon-based approach, are focused on the dataset text preprocessing methods rather than dealing with lexicon itself.

In this study, we think beyond of the box, we constraint firstly on building a new sentiment lexicon with reasonable number of words and then doing adequate preprocessing methods on the lexicon's words in addition to the (Twitter dataset). The study presents Saudi Dialect Sentiment lexicon called SaudiSentiPlus contains 7139 words.

Due to that Saudi dialect words originally and mostly are extracted from Arabic language words and Arabic language is a morphological language and their words might be varied depending on the presence and position of some well-known letters in a word. Moreover, some of these letters come at the beginning (prefixes) or end (suffixes) of a word. Furthermore, these letters also have different shapes depending on their word appearance in the text or context.

In order to increase performance of proposed Saudi dialect lexicon (SaudiSentiPlus) we developed two lexicon based algorithms to deal with (prefixes and suffixes) letters of the lexicon's words (see Fig. 3 and Fig. 4).

The experiment which has been conducted to evaluate the performance of SaudiSentiPlus comprises four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral). All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. They labeled 300 tweets for each classification.

A comparison has been made among SauDiSenti with its 4431 words [17] and the study proposed lexicon (SaudiSentiPlus). The results, as illustrated in Table II, show that SaudiSentiPlus with the two lexicon-based algorithms achieved 81% accuracy which outperformed SauDiSenti with its 54% of accuracy.

REFERENCES

[1] Alghaith, W. (2015). Understanding Social Network Usage: Impact of Co-Presence, Intimacy, and Immediacy. International Journal of Advanced Computer Science and Applications (IJACSA), 6(8), 99-111.

[2] Alharbi, A, & Donckera, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54 (5), 50–61.

[3] Arab Social Media Report. (2014). Citizen Engagement and Public Services in the Arab World: The Potential of Social Media. Mohammed bin Rashid School of government, 1(6). Retrieved from Arab Social Media Report Website: http://www.arabsocialmediareport.com/.

[4] Statista. (2019, July). Leading countries based on number of Twitter users as of July 2019 (in millions). In Statista - The Statistics Portal. Retrieved from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ In-text Citation: (Statista, 2019).

[5] Adayel, H. & Azmi, A. (2016). Arabic tweets sentiment analysis – a hybrid scheme. Journal of Information Science, 42(6), 782–797.

[6] Internet World Stats. (2019, July). Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language). In Internet World Stats Portal. Retrieved from https://www.internetworldstats.com/stats7.htm/ In-text Citation: (Internet World Stats, 2019).

[7] Boudad, N., Faizi, R., Haj Thami, R., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9 (2018) 2479–2490.

[8] Darwish, K., & Magdy, W. (2014). Arabic Information Retrieval. Foundations and Trends in Information Retrieval, 7(4):239–342.

[9] Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., & Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. Journal of Information Science, 44(3), 345–362. https://doi.org/10.1177/0165551516683908

[10] Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science, 44(4), 491–511. https://doi.org/10.1177/0165551517703514

[11] Abdul-Mageed, M, & Diab, M. (2014). SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In LREC (pp. 1162–1169).

[12] Ibrahim, H., Abdou,S. & Gheith, G. (2015). Automatic expandable large-scale sentiment lexicon of modern standard Arabic and colloquial. In 2015 First International Conference on Arabic Computational Linguistics (ACLing) (pp. 94–99). Cairo, Egypt.

[13] Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. Journal of Information Science 44(2): 184–202.

[14] AlNegheimish, H., Alshobaili, J., AlMansour, N., Bin Shiha, R., AlTwairesh, N., & Alhumoud, S. (2017). AraSenTi-Lexicon: A different approach. In International Conference on Social Computing and Social Media (pp. 226–235). Springer, Cham.

[15] Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). Seattle, Washington: ACM DL.

[16] Liu, B., Hu, M. and Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14th International World Wide Web conference (WWW-2005). Chiba, Japan: ACM DL.

[17] Al-Thubaity, A., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. In: Proceedings of the 4th International Conference on Arabic Computational Linguistics (ACLing 2018). Dubai, UAE: Procedia Computer Science 142. 301–307.

[18] Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, 40(4), 501–513. https://doi.org/10.1177/0165551514534143

[19] Froud, H., Lachkar, A., & Ouatik, SA. (2012). A comparative study of root-based and stem-based approaches for measuring the similarity between Arabic words for Arabic text mining applications. Advanced Computing: An International Journal, 3(6), 55–67.

[20] Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014). AraNLP: a Java-based library for the processing of Arabic text. In: Proceedings of the international conference on language resources and evaluation (LREC). Reykjavik, 26–3.

[21] Al-Kabi, M.N., Alsmadi, I.M., Gigieh, A.H., Wahsheh, H.A., & Haidar, M.M. (2014), Opinion mining and analysis for Arabic language. International Journal of Advanced Computer Science and Applications (IJACSA), 5(5), 181-195.

Empirical Analysis of Object-Oriented Software Test Suite Evolution

Nada Alsolami[1]
Computer Science Department, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

Qasem Obeidat[2]
Computer Science Department
University of Bahrain
Sakheer, Kingdom of Bahrain

Mamdouh Alenezi[3]
Computer Science Department
Prince Sultan University
Riyadh, Saudi Arabia

***Abstract*—The software system is evolving over the time, thus, the test suite must be repaired according to the changing code. Updating test cases manually is a time-consuming activity, especially for large test suites, which motivate the recent development of automatically repairing test techniques. To develop an effective automatic repair technique that reduces the effort of development and the cost of evolution, the developer should understand how the test suite evolves in practice. This investigation aims to conduct a comprehensive empirical study on eight Java systems with many versions of these systems and their test suites to find out how the test suite is evolving, and to find the relationship between the change in the program and the corresponding evolution in the test suite. This study showed that the test suite size is mostly increased, where the test suite complexity is stabilized. The increase (or decrease) in the code size will mostly increase (or decrease) the test suite size. However, the increasing or decreasing in the code complexity is offset by stabilizing the test suite complexity. Moreover, the percentage of the code coverage tends to be increased more than decreased, but in the mutation coverage, the opposite is true.**

***Keywords*—Software; test; code complexity; code coverage; test evolution**

I. Introduction

Software testing is an important and essential step to identify the correctness and quality of software system. In the software testing process, the tester should write one test case or more to check each function of the system. The test case is the smallest meaningful unit of the tests. The result of each test case is either pass or fail. .If test cases are passed (i.e., the actual results = the expected results), then the functionality of a software system corresponding to these passed test cases is working correctly. The test suite is a collection of test cases to test system functionalities. Any software system (S) is divided into two parts: program (P) and test suite (T). All system test cases (Tc) are stored in (T). These test cases are used to check the correctness of all parts of P. The new version of the software system (S') should have a different program (P') and test suite (T').

Software systems evolve and change during their development and maintenance. Even a little change in the software code can affect a large number of test cases [1]. The software system upgrades are accompanied by code refactoring or code evolution. The code refactoring is a process of improving the internal structure of code without changing the external behavior or system functionality [2]. Refactoring process makes the code more readable, does not contain duplications, easier in maintenance, and increase the quality of the code. On other hand, the code evolution is adding new code to add new functionality to the system or deleting/ modifying the existing parts of the code to edit functionality in the system. Code evolution is a continuous process; it may change the system functionality and external behavior.

Software evolution is one of the essential and normal issues required for most existence software throughout their lifetime. The changes in the code make some of the test cases in the current test suite become out of date for the new version of the software. Therefore, the tester must revise all changes on the code to repair the corresponding test cases in the test suite. While the code evolution may happen frequently, it is very hard for a tester to follow all these code evolutions and make the correct decisions as create, delete, and update test cases. Also, it is time-consuming to repair test cases manually, particularly, the large test suite [1]. This motivates researchers to develop automatic test repair techniques. The basic requirement to automate test repair technique is comprehensive understanding of how test suite evolves in practice. Generally, analysis a test suite evolution can help developers to build effective automated test repair techniques. So, this paper is intended to conduct an empirical study to understand and identify how the test suite evolves during the code changes and to create or build a relationship between code changes and the corresponding changes of the test suite. The main goal of the experiment is to provide answers to the following research questions regarding test suites evolution:

RQ1. Is the test suite size increasing/ decreasing/ stabilizing during a software evolution?

RQ2. What is the relationship between source code size and test suite size during a software evolution?

RQ3. Is the test suite complexity increasing/ decreasing/ stabilizing during a software evolution?

RQ4. What is the relationship between source code complexity and test suite complexity during a software evolution?

RQ5. What is the effectiveness of the test suite during a software evolution?

II. Related Works

This section shows some previous studies on the test suite evolution and the co-evolution between code and test.

A. Test Suite Evolution

Elbaum *et al.* examined the impact of software evolution on code coverage information [3]. This examination showed that a little change in software code can give large impact on code coverage information. This impact increases as the degree of change increases and it could be difficult to predict.

One study presented a technique for studying test-suite evolution [1]. There were 88 program versions studied, 14,312 test cases, and 17,427 test changes (i.e., modification, addition, and deletion). This study provided initial insight on how test cases are added, removed, and modified in practice. It focused on test repair and investigated the characteristics of deleted and added test cases and implemented technique within a tool called TestEvol [4]. TestEvol is a tool which enables the systematic study of test-suite evolution for Java programs and JUnit test cases. In [1] and [4], the researchers showed that the test modifications tend to be complex and hard-to-automate. The most important results were, firstly: the occurring of non-repair test modification nearly four times as frequently as test repairs. In other words, repairing test is a relatively small fraction of the activities performed during test evolution. Secondly: many test cases are not really deleted and added, but rather moved or renamed.

Another approach to study test suite evolution is based on the observation that software developers follow common patterns to identify changes and adapt test cases [5] and [6]. Mirzaaghaei et al. proposed a novel approach for repairing and generating test cases automatically during software evolution. These studies defined a set of algorithms for repairing test cases commonly adopted by software developers and implemented those algorithms on TestCareAssistant (TCA) for evaluation. TCA properly repairs 90% of the compilation errors, where the TCA addressed and generated test cases that cover the same amount of instructions of state of the art techniques.

All aforementioned studies in this section examined the test suite evolution from different aspects, whether from repairing test cases based on following common patterns commonly adopted by software developers or from fixing test oracle or others. However, in this study, we studied the test suite evolution in term of size (RQ1), complexity (RQ3), and effectiveness (RQ5).

B. Co-Evolution between Code and Test

Marsavina *et al.* investigate fine-grained co-evolution patterns of production and test code [7]. This investigation analyzed five open source systems and then generates six patterns. These patterns explain the relationship between the change in code and corresponding test cases in the test suite.

Levin *et al.* have done a large scale study of 61 open source projects to study the relationship between test maintenance and production code maintenance in semantic changes [8]. The most important results were that the test maintenance is individually in each project rather than standardized.

Several researchers studied the nature of the co-evolution between code and test (i.e. synchronously or phased) [9], [10] and [11]. In [9], Lubsen *et al.* used two cases studies: open source system and industrial software system, as they used association rule mining to study the natural of co-evolution. They concluded that within an open source system the development and testing are separate activities, wherein the industrial software system, the developer used test-driven development strategy. In [10] and [11], the researchers proposed three views which are: change history view, growth history view, and test coverage evolution view to study the nature of the co-evolution. In study number [10], the researchers used two open source projects, while in [11], they used two open source projects and one industrial software project. They concluded that the nature of co-evolution depends on the development style that is used to develop a project.

Ens *et al.* create and implement the interactive visual analytics tool for analyzing co-change and co-evolution between code and test [12]. It enables managers and engineers to display 2D and 3D views. In addition, it helps in determining the intensive period of testing and development and determining the development style of the project.

III. Methodology

The main goal of this study is to understand how test suite evolves over the time. Thus, to achieve this goal, several versions of 8 open source Java systems with their test suits were used to investigate different aspects of test-suite evolution. These systems were selected according to many criteria, which are popular, system size, each system has at least 5 versions, and each version has a JUnit test suite. The 8 open source Java systems that used in our empirical study are selected from GitHub (https://github.com/). Table I lists the systems and its versions.

Most researches have studied the relationship between the code and the test suite generally and provided general information about the relationship between the code and the test suite. However, in the current investigation, we studied the relationship between code and test suite in term of size and complexity.

A. Relationship between Code and Test Suite

The test suite is changing and evolving during its lifetime according to the code changes. Therefore, the relationship between the code and its test suite must be investigated. Accordingly, this paper studied the relationship between the code and test suite in terms of size and complexity.

Several metrics are used to determine the size of production code or tests, such as the number of classes, Line of Code (LOC), number of methods, and number of packages. The software complexity focuses on how a piece of code interacts with other pieces. One of the most popular measurements of software complexity is McCabe metric or Cyclomatic complexity metric. The Cyclomatic complexity per method metric is the maximum number of linearly independent paths within method [13].

TABLE. I. SOFTWARE SYSTEMS USED IN THE EMPIRICAL STUDY

Program	Description	Number of versions
Commons-Lang[1]	It is a library which provides extra methods for manipulation of the core classes of the Java API.	11
OGNL[2]	It is an expression language for Java, which using the simpler expression than the Java language.	12
Biojava[3]	It is an open source project for manipulating and processing biological data in Java language.	12
Commons-DBCP[4]	It implements Data Base Connection Pooling service.	5
Californium[5]	It is a Java implementation of the Constrained Application Protocol (CoAP).	9
Assertj[6]	It is a Java library which provides an interface for writing rich and strongly typed assertion to improve maintainability and readability of tests [14].	11
MessagePack[7]	Is a binary serialization (pack) format. This enables a process to exchange data as simple and fast as possible.	10
Aho-Corasick[8]	It is a Java implementation of the Aho-Corasick algorithm.	8

In this paper, we used Eclipse Metrics plug-in 1.3.8 tool to measure the size and complexity metrics because it is one of the most commonly used Java tool in many research either in mobile applications or in other applications [15], [16], [17], and [18]. Moreover, this tool work with the most widely used platforms, such as Windows, Mac, and Linux.

B. Test Suite Effectiveness

The test suite effectiveness (i.e., test suite quality) can be described as the number of bugs in code detected by the test suite. It could be measured in two broad ways: code coverage and mutation testing. The code coverage metric measures the percentage of code covered by the test suite. In this study, we used Eclemma tool because it is giving more accurate results [7] and it is one of the most widely used tools for code coverage [7], [19], [20], [21], [22], and [23].

The mutation testing is a testing technique used to check if the current test cases are able to detect any fault in mutant or the change in software system code [24]. Each modified or mutated version of a program called mutant. In mutation testing, the tool generates multiple mutants of the original program and then executes the test suite on each mutant. If the outputs of the same test case in both, mutant and original program, are different, then the test case detects the fault and the mutant is called killing. However, if the fault is not detected, then the mutant is called surviving. The mutation coverage calculated as a number of detecting mutants over the total number of generating mutants. The detected mutants are killed mutants plus timeout mutants. The mutants called timeout if it causes an infinite loop. The survived mutants are equivalent mutants or not detected mutants. The equivalent mutant is a mutant that acts as original program behavior, and cannot be detected by any test case. The following example explains the equivalent mutants [25]:

Original program:

```
int index = 3;
if (index >= 2)
 return "foo";
```

Mutant program:

```
int index = 3;
if (index > 2)
 return "foo";
```

The effectiveness of the test suite can be measured by the ability of the test suite to detect most mutants. Pitclipse tool is used in this research because it is fast, also it is considered as one of the most popular tools [26]and [27], and it has been used successfully in several studies [28], [29] and [25].

IV. RESULT

In this section, the results of this study will be presented. The results were divided according to the research questions as follows:

A. Code and Test Suite Size Metrics

The size metrics, for both code and test suite, for the eight systems has been illustrated in Fig. 1 according to the number of classes for each version. The number of classes is mostly compatible with the number of line of codes (LOC), so, there is no difference between them.

These results show that the overall percentage of increase in the test suite size within all versions of all systems is equal to 78.6%. This percentage calculated as a number of increased versions over the number of changed versions (i.e. 70 versions). The percentages of stabilizing and decreasing of the test suite size in all versions of all systems are equal 18.5% and 2.9%, respectively. That means the test suite size is mostly increased during its evolution. The percentages of stabilizing and decreasing have been calculated by the same method of calculating the increasing percentage. In all eight systems, there is a harmony which based upon the number of classes' changes between the system code and the test suite as shown in Fig. 1. Here, the compatibility means that any modification to the older version of the system code in terms of increasing or decreasing in the code size is accompanied by the same effect on the test suite size.

In the current investigation, 8 systems with 78 versions have been evaluated. Where, 70 versions have been changed, where the remaining 8 versions are the initial versions of all systems. There are 65 versions out of 70 versions (92.9%) are compatible and most of them have been increased in the size. In the remaining 5 versions (7.1%), there is no compatibility between the changes in code size and test suite size.

B. Code and Test Suite Complexity Metrics

According to the average cyclomatic complexity per method for each version, the complexity metric, for both code and test suite, for the eight systems has been illustrated in Fig. 2. These results show that the overall percentage of

[1] https://github.com/apache/commons-lang
[2] https://github.com/jkuhnert/ognl
[3] https://github.com/biojava/biojava
[4] https://github.com/apache/commons-dbcp
[5] https://github.com/eclipse/californium
[6] https://github.com/joel-costigliola/assertj-core
[7] https://github.com/msgpack/msgpack-java
[8] https://github.com/robert-bor/aho-corasick

stabilizing or increasing in the test suite complexity in all versions of all systems is equal to 94.3% (91.4% stabilizing and 2.9% increasing). On other hand, the percentage of decreasing in the complexity of the test suite in all systems is equal to 5.7%, which means that the test suite complexity is mostly stabilized during software evolution. All previous percentages calculated by the same method used in the previous Section 4.1, besides complexity metric rather than size metric.

In more detail analysis, there are only four versions that increase in the code complexity, where the test suite complexity for one version is decreased and in the other three versions it was stabilized. In addition, there is only one version where the complexity of the code has decreased, but the test suite complexity has stabilized. This means that the increasing or decreasing in code complexity is offset by stabilizing in the test complexity, this relationship achieved by 80%. Furthermore, the code complexity stabilization means stabilization in test complexity, this relationship achieved by 85.7% for all versions of within systems.

C. Test Suite Effectiveness

In this empirical study, the code coverage and mutation coverage was used to predict the quality of the system's test over the time. The code coverage and mutation coverage metrics results are explained in the following subsections, 4.3.1 and 4.3.2.

a) Code Coverage: The results for the code coverage metric for all systems have been illustrated in Fig. 3. The code coverage is increased in 32 versions (45.7%), stabilized in 20 versions (28.6%), and decreased in 18 versions (25.7%). These results show that the code coverage tends to increase more than it is stabilized or decrease during the systems improvement. All previous percentages calculated by the same method used in Section 4.1, besides using code coverage metric rather than size metric.

b) Mutation Coverage: The results for the mutation coverage metric for all systems have been illustrated in Fig. 4. The mutation coverage is increased in 22 versions (31.4%), stabilized and decreased in 24 versions (34.3%). These results show that the mutation coverage tends to stabilize or decrease more than it is increased during the systems improvement. All previous percentages calculated by the same method used in the previous Section 4.1, besides mutation coverage metric rather than size metric.

V. Discussion

In this section, we will discuss the results and we will answer the research questions.

A. Size

The test suite size tends to increase or stabilize in all versions of systems, except for the second version of CommonsDBCP and sixth version of Californium. In these two versions, the test suite size is decreased, this may due to many reasons, such as remove some test cases (i.e., redundant test cases), restructuring the test cases by merge two test cases within one test case, or there are some changes should be considered on the source code. The answer to the first research question (RQ1) of this study, the test suite size is often increased over the time by adding new test cases. This addition caused by the developer who has frequently adding new functionalities to the systems and fixing new critical defects that are found.

As shown in Fig. 1, for all eight systems, there is compatibility between the changes in code size and test suite size. In other words, the increasing, decreasing, or stabilization in the code size leads to an increase, decrease, or stabilize the test suite size, respectively. This relationship achieved in 65 versions (92.9%), while the remaining 5 versions do not satisfy this relationship (7.1%). This is considered an answer to the research question (RQ2). For example, there is an incompatibility between the code size and test suite size for the third and eighth versions of Californium because the test suite may be improved by restructuring test cases or restructuring the source code, where the number of classes of the source code was decreased, while the LOC was increased.

B. Complexity

The complexity results shown in Fig. 2 can be used to answer the research questions (RQ3) and (RQ4). Where, the test suite complexity is mostly stabilized during the software evolution, even as the size of the test suite increases. Moreover, the percentage of decreasing in the test suite complexity was greater than the percentage of increase. Here, the test suite may evolve by adding more methods rather than extend the existing methods (i.e., increase nodes and edges). In most versions, the code complexity was stabilized. The increase, decrease, or stabilization in code complexity is offset by stabilizing in test suite complexity. This is because the test case just calls the code methods, which do not increase the test complexity (i.e. do not increase the number of linearly independent paths within the test method).

In general, the size of the test suite does not increase and evolve by adding new functionalities (methods or classes) to the source code only, but also by improving the current test cases or adding more test cases for the current functionalities. As the first versions of the system always need frequent improvement processes because the developer will understand the functional requirements better by the time, particularly, after the system deployment in real life. In parallel, the improvement may consider the non-functional requirements that effect on the system quality, where the codes for both system and test suite should be written in high quality and in a professional way for the latest versions and it will be more stabilized.

C. Test Suite Effectiveness

In this paper, the test suite effectiveness and quality were measured by two metrics: code coverage and mutation coverage to answer the research question (RQ5).

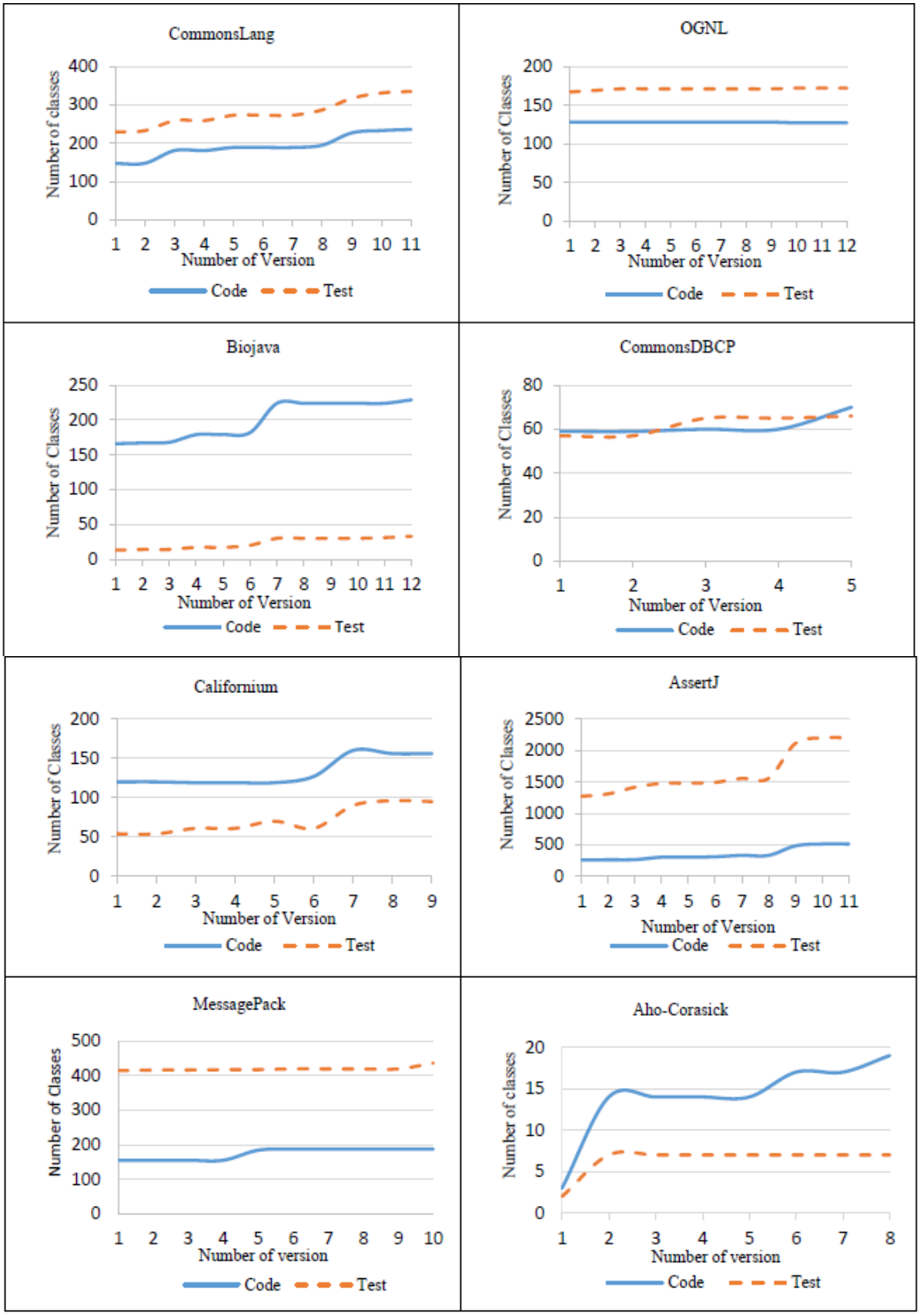


Fig. 1. The Size of the Code and Test Suite for All Systems Versions.

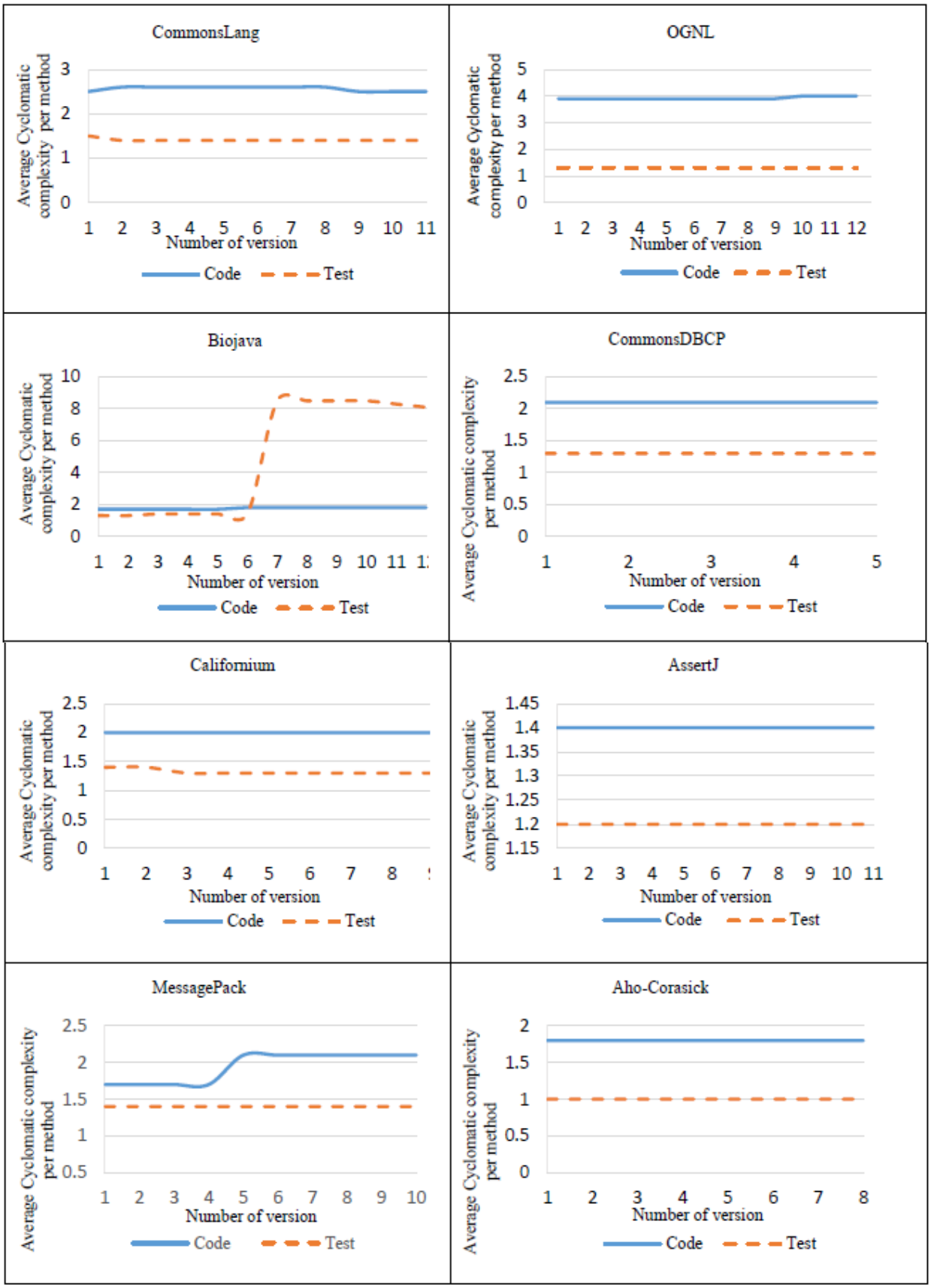


Fig. 2. The Complexity of Code and Test Suite for All Systems Versions.

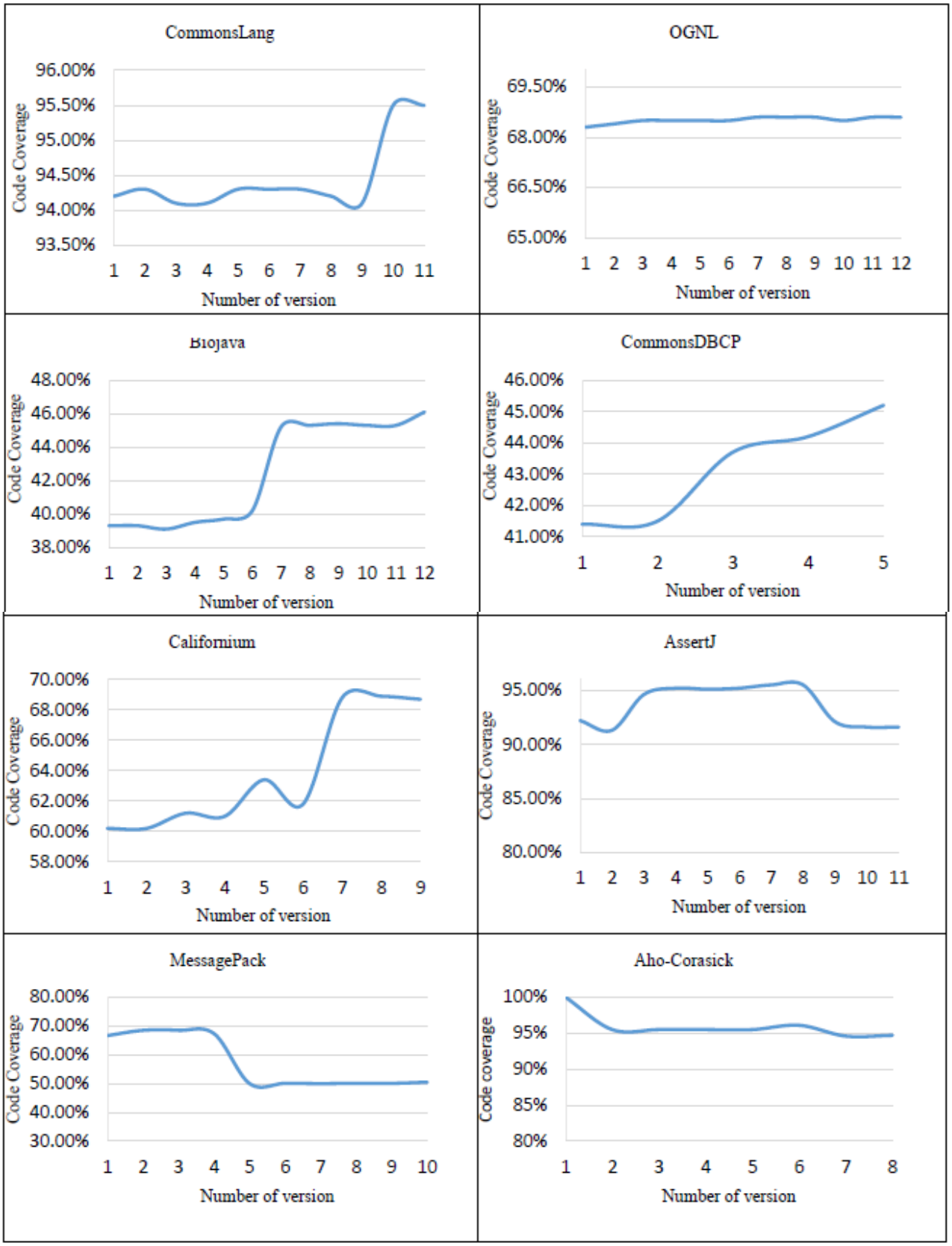


Fig. 3. The Code Coverage for All Systems Versions.

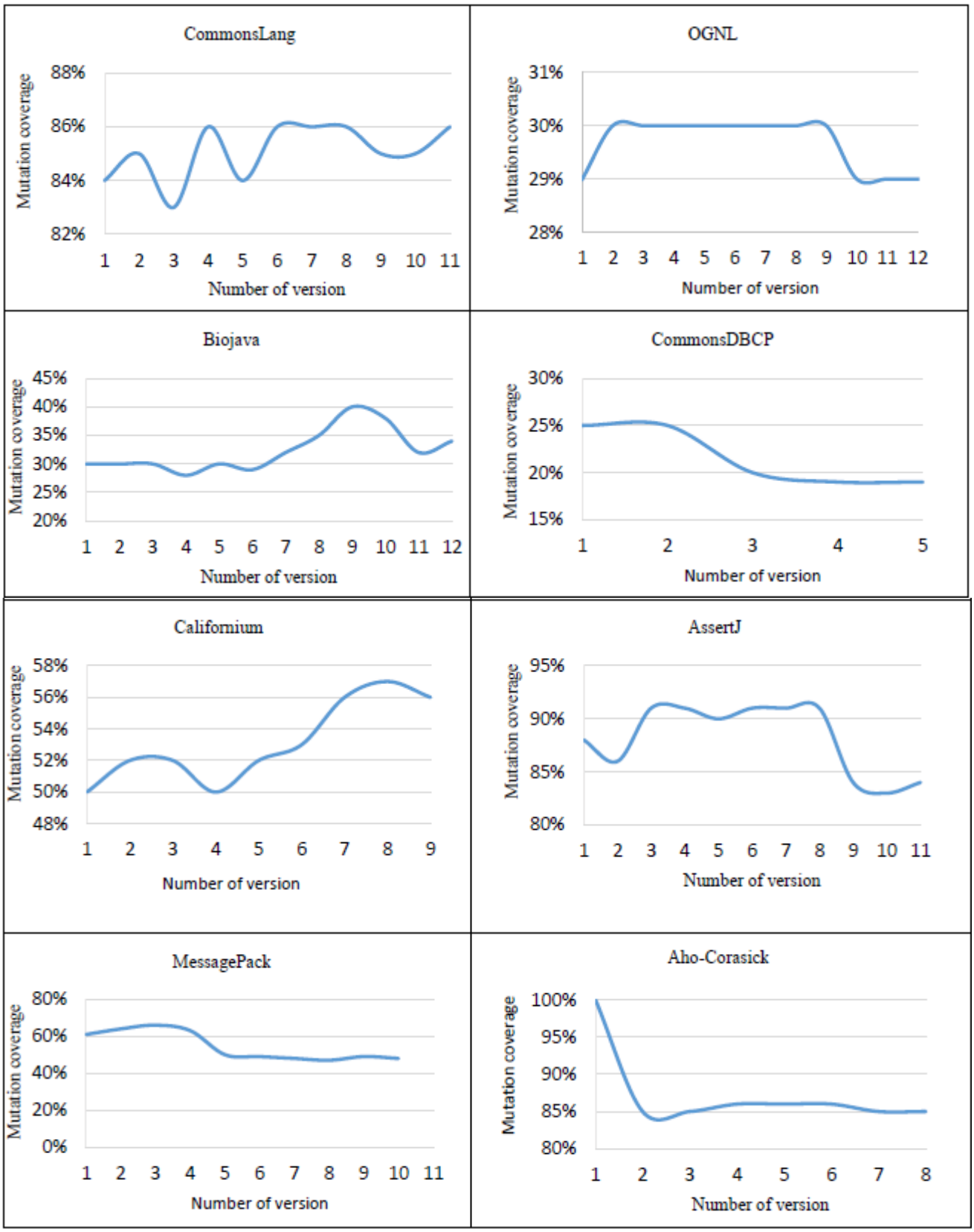


Fig. 4. The Mutation Coverage for All Systems Versions.

The average of the code coverage for all versions in all systems is 71%. This percentage calculated as total code coverage in all versions over the number of versions (i.e. 78 versions). Furthermore, the percentage of the code coverage mostly increase rather than stabilize or decrease within the system versions, as shown in Fig. 5. This indicates that the test suite improved over the time by adding more effective test cases. In a few versions, the code coverage decreased because of some reasons, such as the developer may add new code (i.e., functionality) without adding new test cases. In other words, the code size is increased while the test suite size is stabilized.

When the test suite size is increased more than the code size (i.e., number of classes), the percentage of the code coverage is increased or stabilized by 76.2%. In contrast, when the code size is increased more than the test suite size, then the percentage of the code coverage is decreased by 38.5%. In sum, 76.5% of all versions are increased or stabilized regarding the percentage of their code coverage when increasing the test suites size. This indicates that the percentage of the code coverage can be improved by adding more test cases, to test untested classes and new functionalities added to the source code. All previous percentages have been calculated by the number of versions that satisfied the relationship over the number of versions that satisfied and did not satisfy this relationship.

The average of the mutation coverage of all versions in all systems is 57%. This percentage calculated as total mutation coverage in all versions over the number of versions (i.e. 78 versions). As shown in Fig. 6, the percentage of either decreasing or stabilizing mutation coverage was (34.3%), where it was a little bit greater than the percentage of increasing (31.4%).

The mutation coverage percentage increases as timeout mutants increase, and /or as the test suite size increases to kill more mutants. On the other side, the mutation coverage decreases as survived mutants and equivalent mutants increase. In general, the percentage of increasing and stabilizing for the mutation coverage of all systems versions, at the test suits size increasing or stabilizing was about 66.2%. This percentage has been calculated by the number of versions that satisfied the relationship over the number of versions that satisfied and did not satisfy this relationship. The mutation coverage for the third, fifth, ninth, and tenth versions of CommonsLang and the fourth version of Californium was decreased. This is may be due to the equivalent mutants that affect mutation coverage and causing its decrease. In addition, in CommonsDBCP, MessagePack and Aho-corasick systems the mutation coverage percentage was high but it decreased after a while, this is maybe because the Pitclipse tool wrongly deals with the mutants as a timeout mutants (i.e. infinite loop) which causes increase in mutation percentage in the first versions [25].

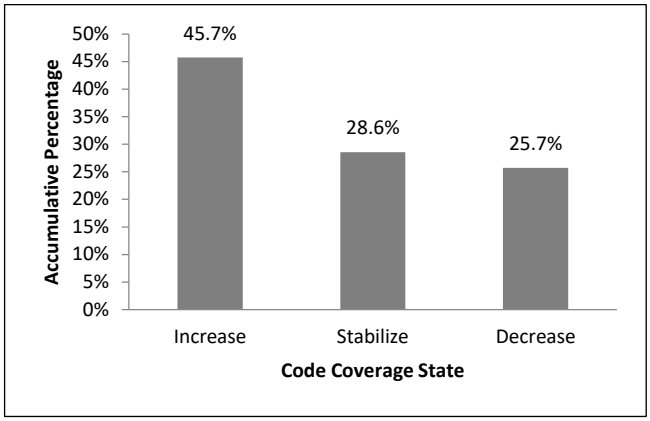


Fig. 5. The Percentage of Accumulative Code Coverage State for All Systems Versions.

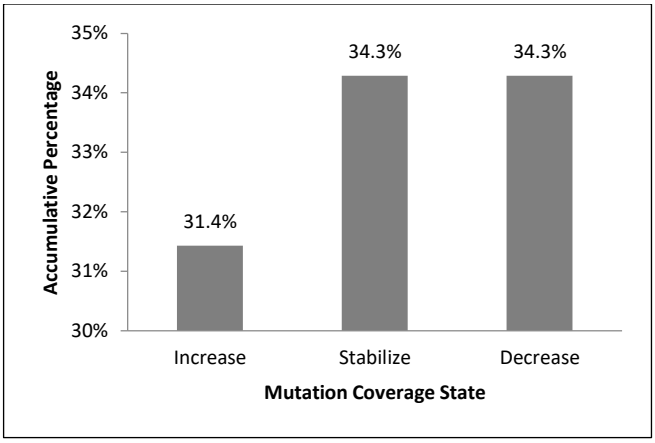


Fig. 6. The Percentage of Accumulative Mutation Coverage State for All Systems Versions.

VI. Conclusion and Future Work

The test suite size mostly increases or stabilizes in program versions. However, the test suite complexity mostly stabilizes and sometimes decreases. The change (i.e. increase or decrease) or stability in the code size is often accompanied by the same change (i.e. increase or decrease) or stability in the test suite size. Often, the complexity of the code and test is stable and does not change between the versions of the system, but in a few cases, it increase or decrease. In these cases, the increase, decrease, or stability in the complexity of the code, offset by stabilization of the complexity of the test.

The code coverage and mutation coverage metrics were measured by using Eclemma and Pitclipse tools, respectively, to evaluate the effectiveness of the test suite. In code coverage, the percentage of increasing the code coverage in program versions is more than the percentage of decreasing. However, in mutation testing, the percentage of decreasing mutation coverage is more than the percentage of increase.

We are planning to extend the current empirical study by engaging and evaluating more well-known Java open-source systems, especially the large ones. Relying on the new results we will develop and build an automated test suite repairing tool. This tool enables a software tester to update the test suite automatically by generating new test cases, deleting, or updating some existence test cases.

REFERENCES

[1] L. S. Pinto, S. Sinha, and A. Orso, "Understanding Myths and Realities of Test-suite Evolution", In Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering, p. 33. ACM, New York, NY, USA, 2012, pp. 33:1–33:11.

[2] M. Fowler, K. Beck, J. Brant, W. Opdyke, and D. Roberts, Refactoring: Improving the Design of Existing Code, Addison-Wesley Professional, 2018.

[3] S. Elbaum, D. Gable, and G. Rothermel, "The impact of software evolution on code coverage information", in Proceedings IEEE International Conference on Software Maintenance. ICSM 2001, pp. 170–179.

[4] L. S. Pinto, S. Sinha, and A. Orso, "TestEvol: A tool for analyzing test-suite evolution", 35th International Conference on Software Engineering (ICSE), 2013, pp. 1303–1306.

[5] M. Mirzaaghaei, F. Pastore, and M. Pezze, "Supporting Test Suite Evolution through Test Case Adaptation", in Verification and Validation IEEE Fifth International Conference on Software Testing, 2012, pp. 231–240.

[6] M. Mirzaaghaei, "Automatic Test Suite Evolution", in Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering, New York, NY, USA, 2011, pp. 396–399.

[7] C. Marsavina, D. Romano, and A. Zaidman, "Studying Fine-Grained Co-evolution Patterns of Production and Test Code", in IEEE 14th International Working Conference on Source Code Analysis and Manipulation, 2014, pp. 195–204.

[8] S. Levin and A. Yehudai, "The Co-evolution of Test Maintenance and Code Maintenance through the Lens of Fine-Grained Semantic Changes", in IEEE International Conference on Software Maintenance and Evolution (ICSME), 2017, pp. 35–46.

[9] Z. Lubsen, A. Zaidman, and M. Pinzger, "Using association rules to study the co-evolution of production #x00026; test code", in 6th IEEE International Working Conference on Mining Software Repositories, 2009, pp. 151–154.

[10] A. Zaidman, B. V. Rompaey, S. Demeyer, and A. v Deursen, "Mining Software Repositories to Study Co-Evolution of Production #x00026; Test Code", in and Validation 1st International Conference on Software Testing, Verification, 2008, pp. 220–229.

[11] A. Zaidman, B. Van Rompaey, A. van Deursen, and S. Demeyer, "Studying the co-evolution of production and test code in open source and industrial developer test processes through repository mining", Empir. Softw. Eng., vol. 16, no. 3, pp. 325–364, Jun. 2011.

[12] B. Ens, D. Rea, R. Shpaner, H. Hemmati, J. E. Young, and P. Irani, "ChronoTwigger: A Visual Analytics Tool for Understanding Source and Test Co-evolution", in Second IEEE Working Conference on Software Visualization, 2014, pp. 117–126.

[13] G. K. Gill and C. F. Kemerer, "Cyclomatic complexity density and software maintenance productivity", IEEE Trans. Softw. Eng., vol. 17, no. 12, 1991, pp. 1284–1288.

[14] "AssertJ". [Online]. Available: http://joel-costigliola.github.io/assertj/. [Accessed: 09-Sep-2018].

[15] D. Franke and C. Weise, "Providing a Software Quality Framework for Testing of Mobile Applications", in Verification and Validation 4th IEEE International Conference on Software Testing, 2011, pp. 431–434.

[16] Shalini and S. I. Hassan, "An empirical evaluation of the impact of aspectization of cross-cutting concerns in a Smart-phone based application", in International Conference on Computing for Sustainable Global Development (INDIACom), 2014, pp. 448–454.

[17] J. Hernandez, A. Kubo, H. Washizaki, and F. Yoshiaki, "Selection of metrics for predicting the appropriate application of design patterns", In Proceedings of the 2nd Asian Conference on Pattern Languages of Programs, p. 3. ACM, 2011.

[18] T. Pessoa, F. Brito, M. P. Monteiro, and S. Bryton, "An Eclipse Plugin to Support Code Smells Detection", arXiv preprint arXiv:1204.6492, 2012.

[19] L. Mariani and F. Pastore, "MASH: A tool for end-user plug-in composition", In Proceedings of the 34th International Conference on Software Engineering, pp. 1387-1390. IEEE Press, 2012.

[20] S. Pathy and D. S. Baboo, "Analysis of code coverage metrics using eCobertura and EclEmma: A case study for sorting programs", vol. Volume 4, no. Issue 2, 2016, pp. 121–130.

[21] N. Li, X. Meng, J. Offutt, and L. Deng, "Is bytecode instrumentation as good as source code instrumentation: An empirical study with industrial tools (Experience Report)", In IEEE 24th International Symposium on Software Reliability Engineering (ISSRE), 2013, pp. 380-389.

[22] P. Dhareula and A. Ganpati, "Open Source Code Coverage Tools for Java: A Comparative Analysis", Indian Journal of Science and Technology, Vol 9(32), 2016, pp. 1–5.

[23] A. Bergel, V. Peña-Araya, and T. Kuhn, "Controlled Experiment to Assess a Test-Coverage Visualization: Lesson Learnt", 2015.

[24] "Mutation Testing: Complete Guide". [Online]. Available: https://www.guru99.com/mutation-testing.html. [Accessed: 17-Apr-2018].

[25] O. Alfsson, "An analysis of Mutation testing and Code coverage during progress of projects", Bachelor's thesis, Umeå University, p. 22.

[26] S. Rani, B. Suri, and S. K. Khatri, "Experimental comparison of automated mutation testing tools for java", in 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015, pp. 1–6.

[27] H. Coles, T. Laurent, C. Henard, M. Papadakis, and A. Ventresque, "PIT: a practical mutation testing tool for Java (demo)", In Proceedings of the 25th International Symposium on Software Testing and Analysis, ACM, 2016, pp. 449–452.

[28] L. Inozemtseva and R. Holmes, "Coverage is not strongly correlated with test suite effectiveness", In Proceedings of the 36th International Conference on Software Engineering, ACM, 2014, pp. 435–445.

[29] J. Zhang, L. Zhang, M. Harman, D. Hao, Y. Jia, and L. Zhang, "Predictive Mutation Testing", IEEE Trans. Softw. Eng., 2018, pp. 1–1.

On Designing Bee Inspired Routing Algorithm for Device-to-Device Communication in the Internet of Things

Asmaa Mohammed Almazmoomi[1], Muhammad Mostafa Monowar[2]
Department of Information Technology, Faculty of Computing and Information Technology
King AbdulAziz University, Jeddah, 21589, Kingdom of Saudi Arabia

***Abstract*—Device-to-device communication is popular research trend that presents ubiquitous information exchange on the Internet of Things. D2D communication provides data exchange without transiting to a base station using direct communication between two devices. For such environment, successful delivery of data to the receiver is needed. In this paper, we suggest a Bee-Inspired Routing Algorithm (BIRA) for D2D communication in IoT exploits the multiple interfaces of a "thing" in IoT having different wireless standards. BIRA is on demand routing algorithm simulates the bee's foraging behavior model to find optimal path between source and destination for multi-hop communication. The performance of BIRA is assessed through extensive simulations that concludes BIRA realizes better packet delivery ratio as well as it performs lower average end-to-end delay in different traffic load compared to the conventional AODV protocol. Also, BIRA achieves least energy consumption than AODV and increases network lifetime.**

Keywords*—Device-to-device communication; internet-of-things; Bee-Inspire Algorithm; routing protocol

I. INTRODUCTION

The Internet of Things (IoT) is ubiquitous computing that introduces the concept of connecting several physical objects to generate, share and consume information with minimal human intervention. It changes traditional system to be smart and available at any time and position [1], [2]. In such environment, different type of devices can collaborate with each other to ensure quality communication and to make more efficient utilization of information. With the increasing demands for diverse applications as well as boost in micro-electro-mechanical systems, the vast proliferation has been observed in the use of these devices (also known as "things") which demands more spectrums supplementary to the radio frequency (RF) for future generation. The Device-To-Device Communication (D2D) is one of the communication models that is implemented in IoT ecosystem which applies direct connection between devices to exchange data in a distributed fashion without traversing a core network. D2D communication is an emerging research phenomenon which benefits ultra-low latency [3] in communication due to a shorter signal traversal path. D2D communication architecture consists of heterogenous computing devices with different capabilities connected together [4], [5]. In addition, due to the limited energy of wireless sender and receiver, it is required for the devices to transfer data through multihop communication.

However, devices characteristics in such network are constrained in terms of battery life and memory size [6]. Due to the device constraints, User Equipments (UEs) should be connected using different short-range wireless networks such as wireless sensor networks (WSNs), wireless fidelity (WiFi), Bluetooth, radio-frequency identification (RFID) networks, and ZigBee [7]. A routing protocol considering these limitations is required to provide efficient data delivery in D2D network.

To date, a number of researches are presented that use bio-inspired algorithm to decrease memory consumption with effective data delivery [8-21]. Biology inspired algorithm is a class of algorithm that mimic specific phenomena from nature [7], and it used to solve problems which resembles the way nature performs. D2D communication in IoT network possesses some unique challenges including autonomous and complex architecture, non-homogeneous nature of devices, resource constraints and distributed control and infrastructure. We argue that routing protocols that adopt bio-inspired algorithm can effectively address these challenges since this genre of algorithms have distinctive features as follows. First, they are appropriate for optimization processes in terms of throughput, energy consumption, and packet delivery ratio. Second, they can handle the heterogeneity of devices and communication technologies in D2D communications. Third, they support device collaboration and self-configuration. Fourth, they have the ability to repair failures. Finally, they provide efficient management for limited resources.

In [22], the authors presented the process of adopting bio-inspired engineering as shown in Fig. 1. First, analogies need to be pointed out between biological and technical systems such as Information and Communication Technology (ICT). Second, model to be created for realistic biological behavior to use it later for developing a technical solution. Finally, simplify and tune the model for the technical application.

Bio-inspired algorithms can be categorized into Swarm Intelligence (SI), Evolutionary algorithms (EAs), and bacterial foraging. Swarm Intelligence (SI) algorithm is one of the bio-inspired methods that is inspired by the collective behavior of distributed and self-organized models e.g. bees swarm. Routing protocol that adopts bee-inspired principle utilizes collective foraging behavior to find source of food. as discussed in detail in Section III.

Fig. 1. Bio-Inspired Engineering.

State-of-the-art research efforts are introduced to present a routing protocol inspired by several biological models in the IoT ecosystem. In [11-14], energy-aware ant routing algorithms are proposed based on ant colonies foraging behavior in network. In WSN, many artificial bee colony-based algorithms are proposed to minimize energy consumption [16-18]. However, these protocols mainly optimized energy efficiency and designed in the context of WSNs, and are not directly applicable to D2D communication in the IoT environment due to its unique features. Several studies exist that use Ant Colony method to meet some D2D challenges the IoT framework [19-21]. However, the aforementioned studies are not suitable for D2D communication in IoT environment as they did not address the heterogeneity of devices along with the coexistence of different wireless standards. In this study, an energy-aware Bee-Inspired Routing Algorithm (BIRA) for D2D communication in IoT is introduced.

BIRA is a reactive routing protocol that is inspired by bee's behavior for searching the food source. BIRA is an on-demand, unicast, self-configured and distributed routing approach. Considering the existence of multiples interfaces in today's devices, BIRA attempts to find optimal path between a source and a destination utilizing multiple interfaces having different wireless standards. In routing discovery, BIRA chooses selective neighbors having acceptable residual energy to decrease the number of control packets. It generates a two types of control packet for route discovery namely, forward and backward scout packets and delivers data packet to destination using foragers. As a routing metric, expected transmission count (ETX) is utilized to measure link cost of different type of Wi-Fi communication. BIRA is simulated using ns3 [23] to assess the efficiency and compare the performance of BIRA with the traditional AODV protocol.

The paper is outlined as follows: Section II introduce the related works. The overview of Bee inspired algorithm in biological system is presented in Section III. The protocol description is discussed in detail in Section IV. Section V explains simulation results, and finally Section VI concludes the paper.

II. Related Work

As of now, a number of studies presented routing protocols in D2D communication [8-10]. Laha A. et al. proposed a quasi-group routing protocol (QGRP), a distributed cluster-based routing scheme focusing energy efficiency [8]. In [9], the authors presented a D2D routing protocol to avoid interference to cellular users. Their attempt aimed to minimize delay, hop count, power consumption, and computational complexity. Utilizing location information, Park J. proposed fast and energy-efficient multihop D2D routing algorithm focusing to improve the end-to-end packet delivery ratio while increasing network lifetime [10]. However, all aforementioned approaches are devised considering smartphone networks, and not appropriately applicable in IoT environment. Also, these protocols cannot provide efficient packet delivery.

Biology inspired algorithm exploits the behavior from nature to solve problems [7]. Ant colony optimization (ACO) attempts to devise solutions motivated by the self-organizing collective behavior of social ant colonies [7]. In [11], an Energy-aware ant routing algorithm (EARA) is proposed utilizing the foraging behavior of ants in multihop communication. EARA introduced energy information with the pheromone value in the route discovery phase and performed energy information maintenance process to update energy values in the routing table in case of change the residual energy of node. The results revealed that EARA performed better than ant routing algorithm. Energy-aware routing protocol and gradient-based routing (GBR) protocol also exploits ACO [12] to minimize energy consumption. Some solutions presented schemes to address mobility, energy consumption, self-optimization and reliability through modifying the traditional AC technique [13], [14]. Enhanced Ant Colony algorithm (EAC) is proposed focusing to improve reliability [13]. The proposed algorithm categorized the node states into active node, sleep node, and critical node. Here, nodes having less residual energy are considered as critical nodes. The path discovery is done before critical packet transmission. The scheme shows negligible effect on energy consumption than AODV. In [14], Bio-inspired optimization for sensor network (BiO4SeL) is presented aiming to reduce energy consumption of sensor nodes. It uses battery power information to create and update the routing table in the case of mobility. However, the approach ignored other performance metrics such as delay and packet delivery ratio. In [15], Ismail et al. presented a modified version of AODV for low-power wireless personal area networks (6LoWPAN). They intended to repair the link failure by proposing a local repair using bio-inspired artificial bee colony routing protocol. The protocol is inspired from the bee foraging process to find the food source. The study can optimize the route performance, decrease route maintenance delay and save energy. However, the authors did not present detailed study including implementation. Few other studies are proposed to minimize energy consumption in WSN exploiting artificial bee colony algorithm [16-18]. Kumar R. et al. proposed artificial bee colony based energy-efficient clustering mechanism to improve network lifetime and cluster head selection [16]. The proposed algorithm is designed to find an optimal shortest path with less energy consumption. It was analyzed and compared with LEACH and PSO and ABC-based routing protocol. It concluded that the proposed method could maximize the network lifetime. Okdem S. et al. provided performance testing and complexity analysis of Cluster-based Wireless Sensor Network routing protocol utilizing artificial bee colony algorithm (CWA) [17]. CWA method consists of four main steps including initialization, setup, broadcast setup configuration, and data gathering. The performance test result shows that the CWA algorithm enhances the network lifetime by saving more energy. In [18], Zheng W. et al. proposed intelligent routing protocol based on ABC algorithm to solve delay-energy trade-off challenge in WSNs. They concluded that the algorithm exhibits good performance in balancing

power consumption, maximizing network lifetime, and optimize delivery ratio. However, these algorithms mainly optimized energy efficiency and designed in the context of WSNs and cannot be applied directly to D2D communication in the IoT environment due to its unique features.

In the IoT environment, few studies exist exploiting AC method to address some D2D characteristics [19-21]. Y. Lu et al. proposed routing algorithm model to address the mobility of IoT network topology as well as the scalability [19]. They concluded that the time for establishing route decreased with the increasing number of nodes and it reduced broadcast storm. Another routing scheme is proposed using AC algorithm considering the node mobility, self-organizing network and node energy [20]. It reduced the time for establishing a route with the increasing number of nodes, and decreased broadcast storm. In [21], researchers presented an ant colony foraging routing scheme utilizing Markov decision model to find an optimal route. The study reduced network overhead and energy consumption while increasing network lifetime. However, the proposed algorithms are not suitable for D2D communication in IoT environment as they did not address device heterogeneity as well as the coexistence of different wireless standards.

III. Bee Colony Optimization Algorithm: An Overview

This section presents detailed overview of bee's foraging principle in biological system and how it is tuned to a realistic biological model in routing protocol.

A. Bee Algorithm in Biology

Artificial bee colony algorithm is a type of swarm intelligence models that simulates intelligent foraging behavior of bees. In such algorithm, food source searching is the main uses of bee communication [24]. To perform a food source searching process, the model has several types of bee groups: scouts and foragers. In particular, the bee foraging behavior begins with the "scouts". Scout bees fly and explore the beehive neighborhood to discover a food source. They carry out a random search with *n* dimensions in the area. Moreover, they are able to fly around to a radius of up to three kilometers from the hive to discover the food sources. Whenever food source is discovered, the scout bees return to the hive with their information to inform other bees about detected food source using dance language. This information related to the quality of food source is exchanged among bees to build collective knowledge through "waggle dance" in the dance area. In bee colony algorithm, a possible solution to the problem is represented by the position of a food source and the quality of the related solution is corresponded to the nectar amount of a food source. Some bee "foragers" obtain the collected information and associate to a particular food source to start forage. The number of foragers is proportional to the quantity of food information shared by the scouts with their nectars. Foragers assess the nectar information taken from all scouts and selects a food source with a probability related to its nectar amount according to (1).

$$P_a = \frac{fit_a}{\sum_{n=1}^{SN} fit_n} \quad (1)$$

where, fit_a is the fitness rate of individual in a population and SN is the number of individuals in the population.

This step is termed as exploration phase followed by the exploitation step [25]. Here, the forager bee collects food and estimates its quantity to make a new decision. The forager either memorizes the food source location or marks it as exhausted food source and returns to the hive as a scout to start finding a new food source [26][27].

B. Bee Inspired Routing Algorithm

In Bee inspired routing protocol, the mapping of bee behavior to routing function is achieved as per follows: Source node in the network is considered as beehive that consists of forward scouts that can work exactly as real forward scout do to discover food source which is represented as destination node. Relay node has both forward and backward scouts which are sent to assist beehive to find path to a food source. The forward/backward scouts provide the neighbors with information about the route they discovered. A routing table represents the dance floor where the forward/backward scouts deliver the information about the path quality. Application data packet is considered as forager which access the information of routing table to deliver the data in efficient way.

IV. Protocol Description

BIRA is inspired by bee colony algorithm in which source node is represented as beehive and the food source resembles destination node. In such model, relay nodes are simulated as worker bees which are neither in beehive nor on the food, called workers. Worker bees might be aware the path to the food at one point. BIRA is an on-demand hop-by-hop routing algorithm. Each node maintains two tables namely routing table and neighbors table. The proposed protocol stores only next-hop and previous-hop in node's routing table rather caching the whole path. The following subsections discuss the protocol operations in detail including packet types, neighbor discovery, route discovery and route maintenance.

A. Routing Packet Types

Forward scout: Forward scout is used to discover the path toward the destination on-demand. This packet consists of some fields initialized by source node as it is shown in Fig. 2. It includes scout identifier (Scout ID) which is a unique incremental value. Scout ID is used to ensure the uniqueness of the route to be discovered. It is generated by source node only and its value is unchangeable. Moreover, Forward scout packet includes Beehive identifier (Beehive ID) that represents the source node ID, and food ID referred to as destination ID. Beehive ID and food ID cannot be manipulated by other nodes. The combination of scout ID and beehive ID preserves the uniqueness of the route request.The hop-count field caches the hop count the scout traversed from the source node to the node it currently passes. Next-hop ID refers to the address of the next neighbor and it changes from hop-to-hop. Expected Transmission Count (ETX) is used to store the path cost information whenever the scout traverses toward destination, the value of ETX is accumulated from source node to the node it current passes. Estimation of ETX is explained in Section IV-b.

Forward scout packet

Scout ID	Beehive ID	Food ID	Hop count	Next-hop ID	ETX

Fig. 2. Packet Structure of Forward Scout.

a) Backward scout: Once the path is discovered, the destination node propagates the scout as backward scout packet toward the source node along the reverse path.

Similar to forward scout, the backward scout uses the same field value for scout ID, the beehive ID, and the food identifier as shown in Fig. 3. Backward scout uses the hop count field to refer the number of hops from the source node to the destination. When the desired route is found, this field is initialized using the hop-count field of the forward scout. Through the ETX field, the source node knows about the path cost to the destination. Upon receiving the forward scout, the Destination node initiates a unique path identification, and embed the value in the path ID field.

b) Forager: Application data is delivered to the destination using forager packet. Forager packet includes several fields which help to direct the waiting data packets to the intended node. Forager ID is a unique incremental value generated by beehive. The Beehive, Food ID, Hop-count and Path ID represents the same as scout packets. The Path ID helps intermediate node to forward the forager packet to the next hop associated with that particular path ID stored in the routing table. Data field carries the application data sent by source node to intended node. Forager packet structure is shown in Fig. 4.

B. Neighbor Discovery

In BIRA, every node broadcasts a periodic refresh packet to update its immediate neighbors about the active connections. The refresh packet also includes the residual energy of the node. To periodically measure the link quality, each node unicast a "probe" packet to its "eligible" neighbors through multiple interfaces (in our scheme we consider two interfaces) to maintain the link quality of the neighbors. We define the "eligible" neighbors are the nodes for which residual energy exceeds some threshold. Thus, BIRA preserves the information of energy efficient neighbors in its neighbor table in a distributed fashion.

Backward scout packet

Scout ID	Beehive ID	Food ID	Hop count	Path ID	ETX

Fig. 3. Packet Structure of Backward Scout.

Forager packet

Forager ID	Beehive ID	Food ID	Hop count	Path ID	Data

Fig. 4. Packet Structure of Forager.

C. Link Cost Estimation

BIRA exploits ETX as a link cost metric for the eligible neighbors, and stores the value in the neighbor table. ETX is defined as the average number of transmissions including retransmissions a node expects to make to successfully forward a packet to the destination [28]. The successful delivery is guaranteed through the reception of link-layer acknowledgment. ETX over the link (a_i,a_j) for node a_i can be estimated as:

$$ETX_{a_i} = \frac{N_{Tx}^{a_i,a_j}}{N_{ACK_{rx}}^{a_i,a_j}} \quad (2)$$

Where, $N_{Tx}^{a_i,a_j}$ denotes the total number of transmissions (including retransmission) from $a_i \; to \; a_j$, and $N_{ACK_{rx}}^{a_i,a_j}$ designates total number of ACK receptions by node a_i from a_j which also signifies the total number of successful transmission from $a_i \; to \; a_j$.

BIRA utilizes Exponential weighted moving averages (EWMA) for the smooth estimation of ETX_{a_i}, to address the abrupt changes in link condition. The running average of this ETX is estimated as:

$$ETX_{tr}^{a_i} = \alpha\left(ETX_{tr}^{a_i}(curr)\right) + (1-\alpha)ETX_{tr}^{a_i} \quad (3)$$

Where, $ETX_{tr}^{a_i}(curr)$ refers to the current observation of ETX and α is the tuning parameter that satisfy $0 < \alpha < 1$.

D. Route Discovery

When a node requires to transmit data, it first checks the availability of a valid path information of the destination in its routing table. If a valid route is found with sufficient foragers, the source node forwards the data. Otherwise, it launches the route discovery procedure. Here, the source node generates a forward scout packet with unique scout ID, and replicates a number of forward scouts with the same scout ID and transmits to its eligible neighbors. The number of forward scouts depends on its eligible neighbors. The forward scout is transmitted through the interface having minimum ETX value. Thus the path traversed by forward scout toward the destination node comprises different interfaces with optimal link cost as illustrated in Fig. 5. Forward scout marks temporarily its path in the routing tables of the visited nodes which will be utilized later for the return packet (i.e., backward scout).

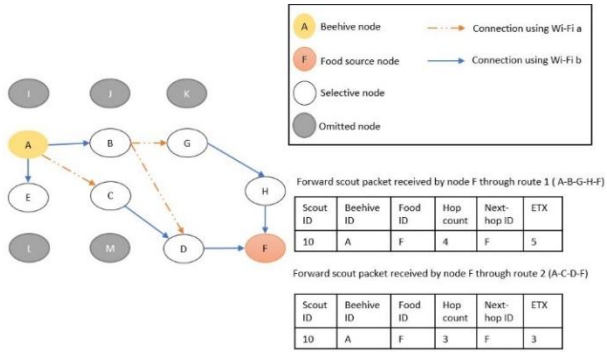


Fig. 5. Forward Scout Delivery.

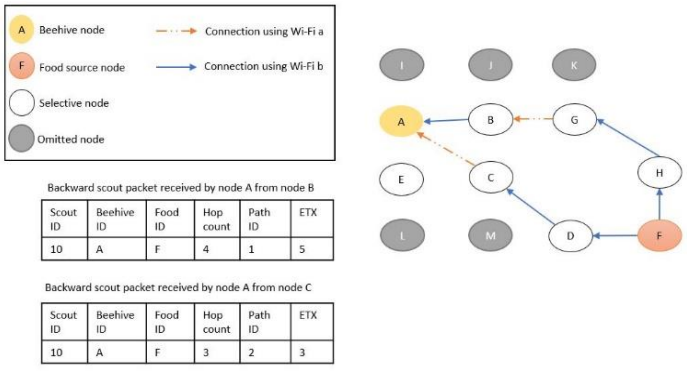
When the forward scout passes an eligible neighbor, the ETX value of the traversed link is accumulated in the ETX field of the forward scout. The intermediate node verifies whether it has already received a forward scout with the same scout ID and beehive ID. In this case, the intermediate node compares the ETX value of the currently received forward scout packet with the previous one and drops the current forward scout if its ETX value exceeds than that of the previous one. Afterwards, the node forwards the forward scout packet to its own eligible neighbor following the same manner toward the Beehive. Every intermediate node updates its routing table fields whenever it received scout packets. Fig. 5 illustrates how node A finds a route to node F if it doesn't have a valid route. Node A creates forward scouts and transmits to only selective nodes B, C and E based on acceptable percentage of remaining energy. Then node S communicates with selective nodes through optimal interface. Node A communicates with node B and E through wi-fi b link, but it communicates with node C through wi-fi a interface since the selected links have lower link cost. After receiving forward scout, node B, C and E check if the received forward scout is duplicate or not and then update the routing table and change hop count and ETX fields of the received packet, and forward it in a similar way as done by node A. In this example, node D receives duplicated forward scout packet; one from node C and the other from node B. But node D drops the forward scout packet from node B as it has higher ETX value than that of node C. Node C communicates with its neighbor node D through Wi-Fi b link while node B communicates with its neighbors, node G and D through Wi-Fi a link. Here, node F which is the food source (i.e., destination) receives forward scout packets from both node H and D through path A-B-G-H and A-C-D respectively as shown in Fig. 5.

Upon receiving a forward scout, the intermediate node may generate a backward scout and transmits it along the reverse path toward the source node, if the path information is available in its routing table. This is done by looking at the previous-hop field at the routing table. Otherwise, the node continues forwarding the forward scout. When the forward scout packet reaches to the destination, a backward scout packet is created and transmitted along the reverse path to the source node. When a node receives backward scout, it permanently stores the routing information, and utilizes the next hop field to determine a path to the destination. After the arrival of the backward scout at the beehive node, the path cost is retrieved from the ETX field of backward scout packet, and is stored as a routing table entry. As mentioned earlier, the destination node initializes the Path ID in the backward scout packet to indicate the traversed route. Fig. 6 shows the delivery of backward scout at the beehive node which is node A. Backward scout packets generated by node F are sent along reverse path through similar interfaces the forward scout has been received. Node A received two backward scout packets from different path (F-H-G-B and F-D-C) as depicted in Fig. 6. Node A updates ETX field of received backward scout and records the information carried in routing table. Eventually, the beehive node recruits forager to deliver data to the destination as depicted in Fig. 7. Node A select optimal path based on ETX value stored in the routing table. It selects path 2 because the ETX value is 3 which is less than path 1. Using path 2, the forager packet is forwarded from node A to node C through wi-fi an interface then node C communicates with node D through wi-fi b link and forward the received forager packet. Also, node D sends the forager to the intended node F through Wi-Fi b interface.

The flowcharts of BIRA in routing discovery process for Beehive, intermediate node and food source are illustrated in Fig. 8, 9 and 10, respectively.

E. Route Maintenance

In order to maintain route, BIRA utilizes two types of timer: route validity timer and route request waiting timer. After transmitting the forward scout the beehive node waits for the timeout period of route request waiting timer and retransmits the forward scout after timeout occurs. On the other hand, a node purges the expired entry from the routing table upon the expiry of route validity timer.



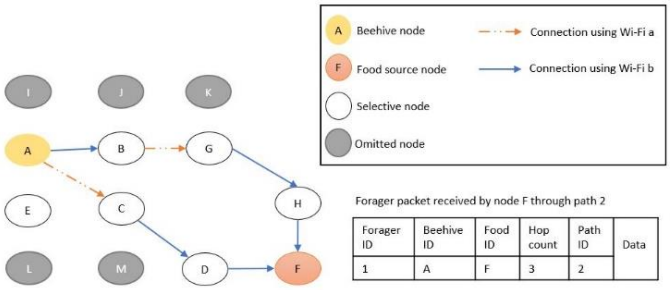
Backward scout packet received by node A from node B

Scout ID	Beehive ID	Food ID	Hop count	Path ID	ETX
10	A	F	4	1	5

Backward scout packet received by node A from node C

Scout ID	Beehive ID	Food ID	Hop count	Path ID	ETX
10	A	F	3	2	3

Fig. 6. Backward Scout Delivery.



Forager packet received by node F through path 2

Forager ID	Beehive ID	Food ID	Hop count	Path ID	Data
1	A	F	3	2	

Fig. 7. Data Delivery using Forager.

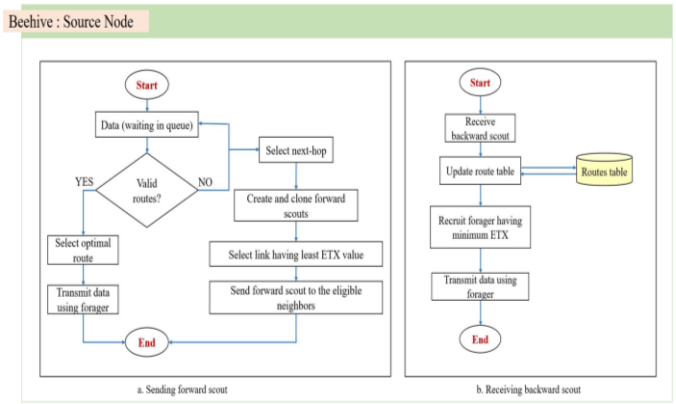


Fig. 8. Beehive's Route Discovery Flowchart.

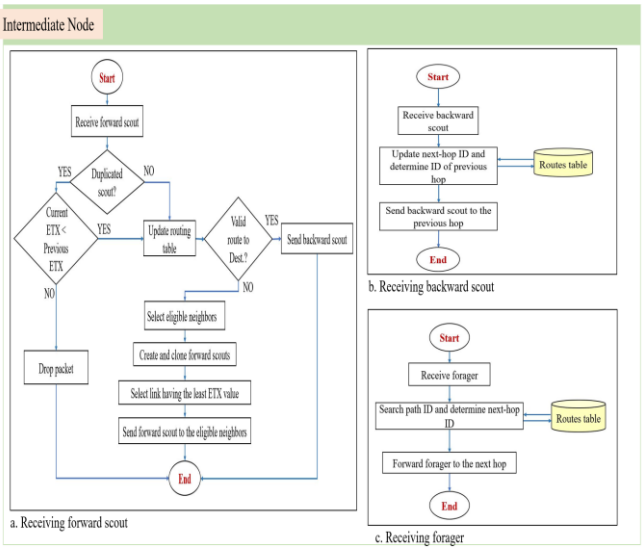


Fig. 9. Intermediate Node's Route Discovery Flowchart.

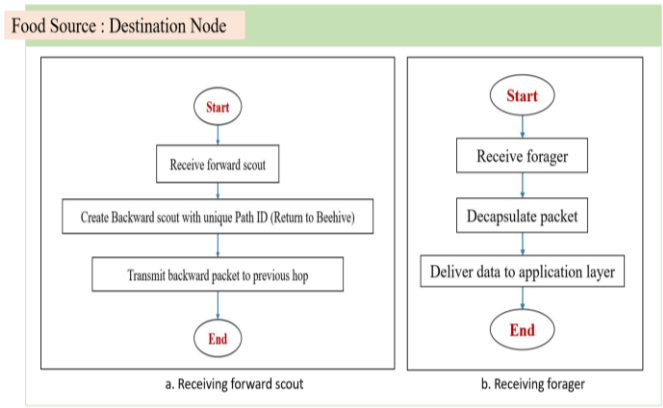


Fig. 10. Food Source's Route Discovery Flowchart.

In an IoT network, a node broadcast periodic refresh packet to know about the active neighbors. A node marks the routing table entry of its neighbor as invalid if it does not receive the refresh packet from its neighbor for a specified period of time. Then, an error scout packet is transmitted to notify the other nodes regarding the broken link. Upon receiving the error scout packet, all the nodes eventually delete the corresponding routing table entry from their routing table. The beehive node might initiate the route discovery procedure again if required.

V. Performance Evaluation

In this section, the performance of BIRA is assessed and compared its outcomes with AODV using simulation.

A. Simulation Parameters

BIRA is simulated using NS3 version 3.28.1 on the Ubuntu 16.04 LTS Linux Operating System. Varying number of nodes are deployed in a 1000x1000 m^2 area. The simulation is carried out for 800 seconds. Two MAC layer protocols are implemented in each node with two different interfaces: IEEE 802.11a and IEEE 802.11b. In the simulation scenarios, a log distance propagation model is utilized and Constant Bit Rate generators (CBR) is used over (User Datagram Protocol) UDP in the application layer with different traffic rate (packet/sec). Varying number of source and destination pairs is used in different experiments of the simulation. The payload size that considered is 512 bytes and initial energy of node is set to 100 joules. In each trial, the outcomes are acquired by taking the average over 10 random runs. Table I depicts the simulation parameters. The performance of the protocol is evaluated using three metrics: average end-to-end delay, packet delivery ratio (PDR) and average energy consumption. The performance of BIRA is compared with AODV, one of the most widely used routing protocols for IoT. To comprehensively assess the performance of BIRA, we conduct three types of experiment. In the first experiment, the results of all performance metrics are obtained using different numbers of nodes range from 10 to 50 nodes with one source/destination pair and 1packet/sec application traffic rate. Second experiment results are extracted from 50 nodes and source/destination pairs vary from 1 to 5 pairs with 1 packet/sec application traffic rate. In the last experiment, different traffic rate is implemented ranging from 2 to 10 packets/sec with fixed 50 nodes and 4 source/destination pairs.

B. Performance Metrics

Three performance parameters have been used in the experiment to analyze the performance of BIRA protocol [29]. These metrics are explained as follows:

- Average end-to-end delay comprises all possible delays including route discovery latency, queuing at the interface queue, retransmission delays at the MAC, propagation and transmission delay. It is measured from the time a packet is queued to the time the acknowledgment of the packet is received.
- Packet delivery ratio (PDR) is measured by dividing the number of received packets at the destination node to the number of packets created by the source nodes.
- Average energy consumption parameter of nodes which caused by event transmitting, receiving, checking the channel for incoming transmission and idle. It can be measured as follows [30].

$$E_{rx} = P_{rx} * d_{size} / S \quad (4)$$

$$E_{tx} = P_{tx} * d_{size} / S \quad (5)$$

$$E_{min} = E_{check} + E_{idle} \quad (6)$$

$$E_{cons} = E_{min} + d_{rx} * E_{rx} + d_{tx} * E_{tx} \quad (7)$$

$$E_{re} = E_{init} - E_{cons} \quad (8)$$

Where E_{tx} is energy consumed to transmit a data packet (J), P_{tx} is Tx power (W),E_{rx} is energy consumed to receive a data packet (J), P_{rx} is Rx power (W), d_{size} is data packet size (bit), d_{tx} is number of data packet transmitted (packet), d_{rx} is number of data packet received (packet), S is communication speed (bits per second), E_{min} is minimum consumed energy (J), E_{check} is energy for checking channel (J), E_{idle} is energy in idle time (J), E_{cons} is total energy consumed (J), E_{init} is initial energy (J) and E_{re} is remaining energy (J).

TABLE. I. SIMULATION CONFIGURATIONS

Parameter	Value
Routing protocols	BIRA and AODV
Payload size	512 bytes
Simulation time	800 sec
Application data traffic	CBR
Application traffic rate	2,4,6,8 and10 packet/sec
Wireless network	Wi-Fi IEEE802.11
MAC layer protocols	IEEE 802.11a and IEEE802.11b
Propagation loss model	Log distance propagation model
Number of nodes	10,20,30,40 and50 nodes
Number of source and destination pairs	1,2,3,4 and 5
Initial energy	100 Joule

C. Experimental Results

Packet delivery Ratio: As shown in Fig. 11, 12 and 13, BIRA achieves better PDR than AODV in all the three experiments. Fig. 11 shows that BIRA deliver data successfully ranging from 89% to 100% packets but the value for AODV ranges from 75% to approximately 91%. Although the experiment is conducted having 1 source/destination pair, but with the increasing number of nodes, the PDR slightly decreases for BIRA as shown in Fig. 11. This is due to the fact that, in larger network size, the hop-count increases, and packet retransmissions also increases due to varying link quality for each hop. In the first experiment, it is observed that the PDR difference in both the protocol is about 14%. Due to pure flooding of route request packet in AODV, it shows poorer performance as compared to BIRA. BIRA achieved better PDR because it restricts the number of forward scouts which speeds up the convergence.

Despite increasing sending packets from different pairs of source and destination as illustrated in Fig. 12, BIRA outperforms significantly AODV because it unicasts forward scout to selective neighbors and communicates with them through interface having the least ETX value. Fig. 13 shows the delivery ratio of both protocols while increasing the number of packets per second; in the best case, AODV attains 94% PDR at 6 packets/sec traffic rate while BIRA, at that traffic rate, shows 99% PDR, and the best performance of BIRA achieves at 2 packets/sec traffic rate which is 100% PDR.

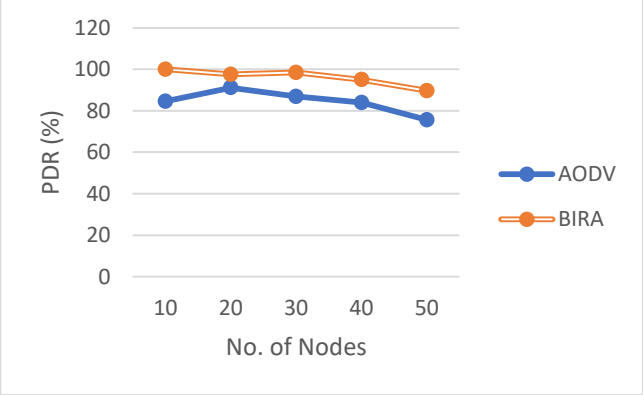


Fig. 11. Packet Delivery Ratio of Varied no. of Nodes.

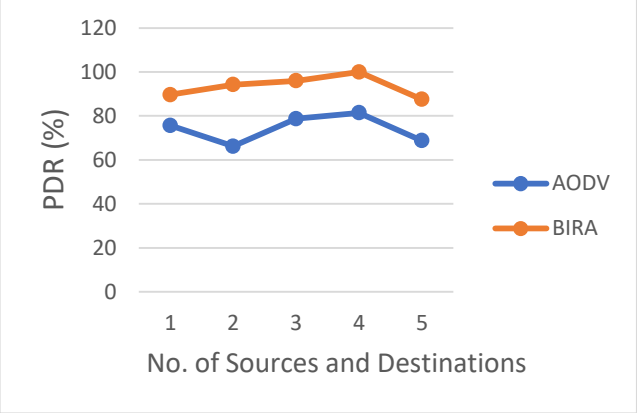


Fig. 12. Packet Delivery Ratio of different Source/Destination Pairs.

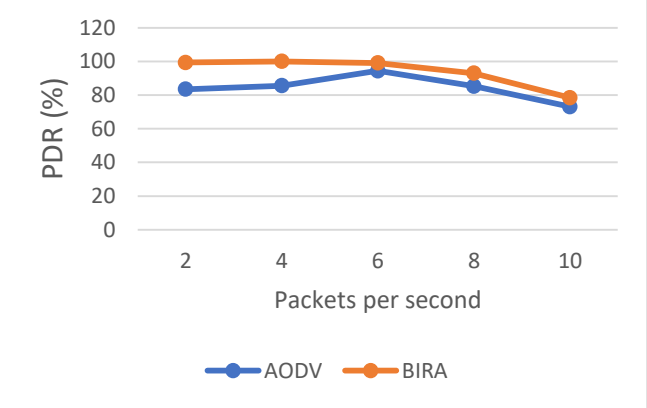


Fig. 13. Packet Delivery Ratio of different Traffic Rate.

a) End-to-End Delay: Fig. 14, 15 and 16 illustrate that the BIRA can deliver the data to the destination faster than AODV. As shown in Fig. 14, the average end-to-end delay increases as the number of nodes increases due to the increase in hop-counts, and BIRA outperforms AODV in all cases. This is self-evident since, BIRA chooses ETX as routing metric as well as select the interface having optimal link cost. Hence, the total path cost of BIRA is always lower than AODV in terms of end-to-end delay. Moreover, BIRA does not drop all duplicated forward scout packet rather it drops only if duplicated packet has highest ETX value than the earlier one and thus chooses a route having lower path cost which in turn reduces the end-to-end latency. The better delay performance of BIRA is also observed in second experiment with varying source/destination pairs. When the number of source/destination pairs are 5 (i.e., more sources are generating traffic), the end-to-end latency of AODV is around 0.06 seconds while BIRA achieves a delay of 0.03 seconds as illustrated in Fig. 15. BIRA also outperforms AODV in varying traffic load as shown in Fig. 16. In all the traffic loads, AODV shows poorer delay performance than BIRA since AODV routes the packet depending on hop count metric and does not exploit the heterogeneity of different wireless standards for choosing optimal route.

b) Energy Consumption: The energy consumption performance of BIRA is evaluated in all the three experiments

as illustrated in Fig. 17, 18 and 19. Fig. 17 indicates that BIRA is energy efficient and can increase the network lifetime as compared to AODV in different network size. This is because, BIRA always chooses eligible neighbors having higher residual energy. Moreover, it selects the links with least ETX value that signifies reduced retransmissions which in turn minimizes energy consumption. The lower energy consumption of BIRA has also been observed while varying source/destination pairs as depicted in Fig. 18, and with different traffic rate as shown in Fig. 19. AODV has poor energy performance than BIRA because it consumes a lot of power when broadcasting route request packets in the route discovery phase, also AODV cannot choose link with lower retransmission value.

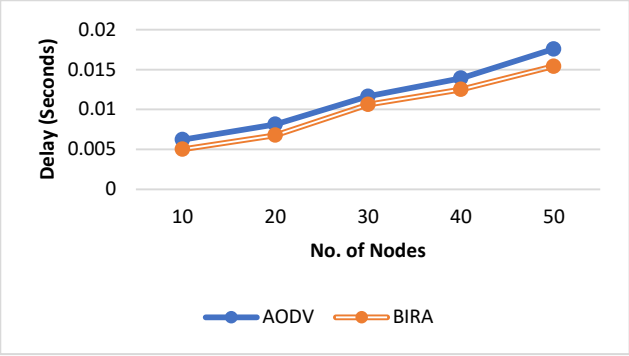


Fig. 14. Average End-to-End Delay of different no. of Nodes.

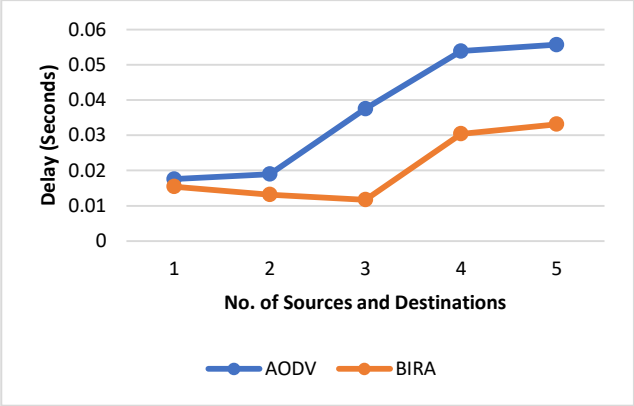


Fig. 15. Average End-to-End Delay of different no. Source/Destination Pairs.

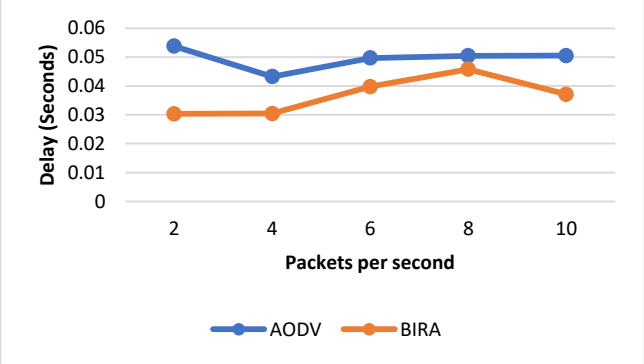


Fig. 16. Average End-to-End Delay of different Traffic Rate.

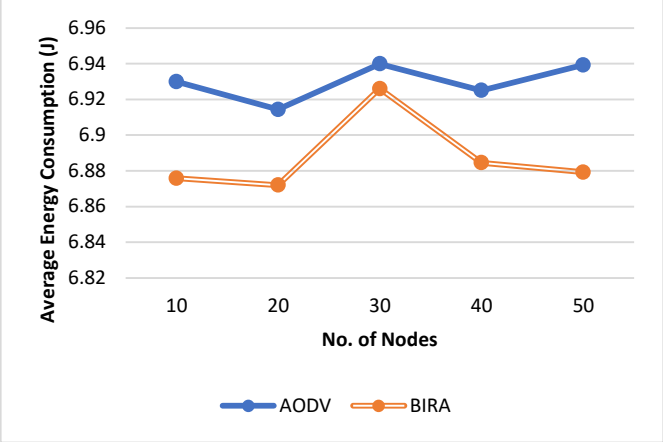


Fig. 17. Average Energy Consumption of different no. of Nodes.

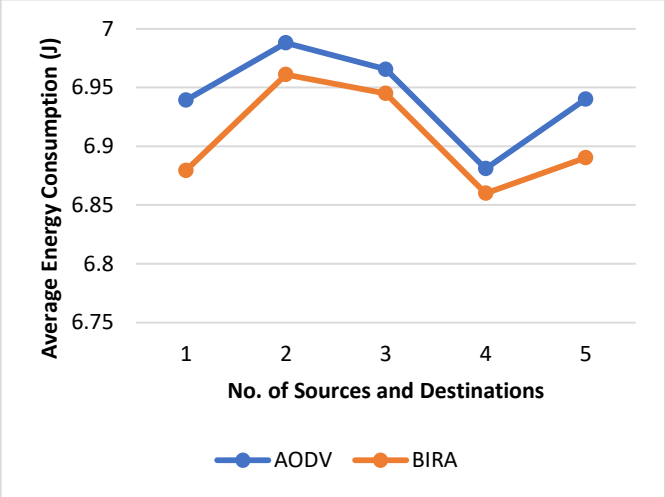


Fig. 18. Average Energy Consumption of different no. of Source/Destination Pairs.

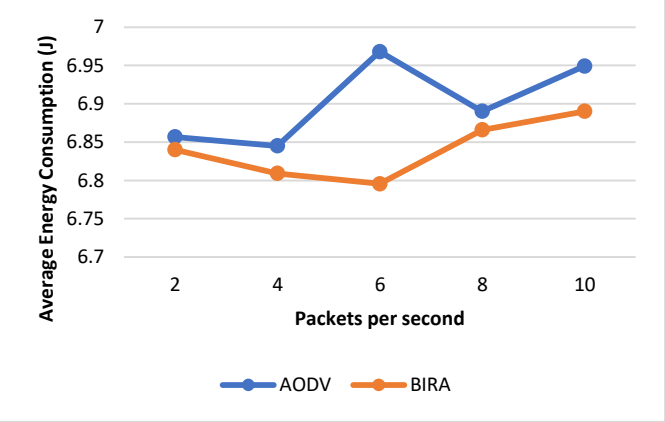


Fig. 19. Average Energy Consumption of different Traffic Rate.

VI. Conclusion

In this paper, we introduced BIRA routing protocol inspired from the bee communication principle to meet the challenges of Device-to-device communication in IoT environment. BIRA is an on-demand hop-by-hop routing

algorithm that exploits the multiple interfaces of a node having different wireless standards to find an optimal path between a source and destination. The performance of BIRA is realized through extensive simulations using ns-3 simulator. The outcomes deduced that as compared to AODV, the BIRA obtains least energy consumption and longer network lifetime. In addition, BIRA performs better packet delivery ratio than AODV as well as it achieves lower average end-to-end delay in different traffic loads. In future work, authors aim to consider dynamic network topology and interference.

REFERENCES

[1] Majeed, "Internet of Things (IoT): A verification framework," in 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1–3, 2017,.

[2] P. Barnaghi and A. Sheth, "On searching the internet of things: Requirements and challenges," IEEE Intell. Syst., vol. 31, no. 6, pp. 71–75, 2016.

[3] R. M. Mare, C. Luiz Marte, and C. E. Cugnasca, "Visible Light Communication Applied to Intelligent Transport Systems: An Overview," IEEE Lat. Am. Trans., vol. 14, no. 7, pp. 2101–2106, 2016.

[4] F. Deicke, W. Fisher, and M. Faulwasser, "Optical wireless communication to eco-system," Futur. Netw. Mob. Submit, pp. 1–8, 2012.

[5] P. E. Hockberger, "A history of ultraviolet photobiology for humans, animals and microorganisms.," Photochem.Photobiol., vol. 76, no. 6, pp. 561–579, 2002.

[6] M. Frey, F. Grose, and M. Gunes, "Energy-aware Ant Routing in Wireless Multi-hop Networks", IEEE International Conference on Communications (ICC), pp. 190-196, 2014.

[7] O. Bello and S. Zeadally, "Intelligent Device-to-Device Communication in the Internet of Things," IEEE Systems Journal, vol. 10, no. 3, pp. 1172–1182, Sep. 2016.

[8] A. Laha, X. Cao, W. Shen, X. Tian, and Y. Cheng, "An energy efficient routing protocol for device-to-device based multihop smartphone networks," in Communications (ICC), 2015 IEEE International Conference on, 2015, pp. 5448–5453, 2015.

[9] P. Ren, Q. Du, and L. Sun, "Interference-aware routing for hop-count minimization in wireless D2D networks," in Communications in China-Workshops (CIC/ICCC), 2013 IEEE/CIC International Conference, pp. 65–70, 2013.

[10] J. Park, "Fast and Energy Efficient Multihop D2D Routing Scheme," International Journal of Distributed Sensor Networks, vol. 12, no. 5, p. 2148734, May 2016.

[11] M. Frey, F. Grose, and M. Gunes, "Energy-aware Ant Routing in Wireless Multi-hop Networks", IEEE International Conference on Communications (ICC), pp. 190-196, 2014.

[12] R. Arya and S. C. Sharma, "Analysis and Optimization of Energy of Sensor Node Using ACO in Wireless Sensor Network," Procedia Computer Science, vol. 45, pp. 681–686, 2015.

[13] M. Umadevi and M. Devapriya, "An Enhanced Ant Colony Based Approach to Optimize the Usage of Critical Node in Wireless Sensor Networks," Procedia Computer Science, vol. 47, pp. 452–459, 2015.

[14] M. F. de Castro, L. B. Ribeiro, and C. H. S. Oliveira, "An autonomic bio-inspired algorithm for wireless sensor network self-organization and efficient routing," Journal of Network and Computer Applications, vol. 35, no. 6, pp. 2003–2015, Nov. 2012.

[15] N. H. A. Ismail and R. Hassan, "6LoWPAN Local Repair Using Bio Inspired Artificial Bee Colony Routing Protocol," Procedia Technology, vol. 11, pp. 281–287, 2013.

[16] R. Kumar and D. Kumar, "Multi-objective fractional artificial bee colony algorithm to energy aware routing protocol in wireless sensor network," Wireless Networks, vol. 22, no. 5, pp. 1461–1474, Jul. 2016.

[17] S. Okdem, D. Karaboga, and C. Ozturk, "An application of wireless sensor network routing based on artificial bee colony algorithm," in Evolutionary Computation (CEC), 2011 IEEE Congress on, 2011, pp. 326–330, 2011.

[18] W. Zheng and D. Luo, "Routing in Wireless Sensor Network Using Artificial Bee Colony Algorithm," pp. 280–284, 2014.

[19] Y. Lu and W. Hu, "A Research on the dynamic routing of Internet of Things Based on Ant Colony Algorithm," The Proceeding of 2nd International Conference on Computer and Applications (CCA), 2013.

[20] Y. Lu and W. Hu, "Study on the application of ant colony algorithm in the route of internet of things," International Journal of Smart Home, vol. 7, no. 3, pp. 365–374, 2013.

[21] C. Cheng, Z. Qian, and G. Ji, "An IoT Ant Colony Foraging Routing Algorithm Based on Markov Decision Model," The Proceeding of International Conference on Soft Computing in Information Communication Technology (SCICT), 2014.

[22] F. Dressler and O. B. Akan, "A survey on bio-inspired networking," Computer Networks, vol. 54, no. 6, pp. 881–900, Apr. 2010.

[23] "ns-3", ns-3, 2019. [Online]. Available: https://www.nsnam.org/. [Accessed: 21- Aug- 2019].

[24] D. Karaboga, B. Gorkemli, C. Ozturk, N. Karaboga, "A comprehensive survey: Artificial bee colony (ABC) algorithm and applications", Artif. Intell. Rev., pp. 1-37, 2012.

[25] H. F. Wedde et al., "BeeAdHoc: an energy efficient routing algorithm for mobile ad hoc networks inspired by bee behavior," in Proceedings of the 2005 conference on Genetic and evolutionary computation - GECCO '05, Washington DC, USA, p. 153, 2005.

[26] H. F. Wedde, M. Farooq, and Y. Zhang, "BeeHive: An Efficient Fault-Tolerant Routing Algorithm Inspired by Honey Bee Behavior," in Ant Colony Optimization and Swarm Intelligence, vol. 3172, M. Dorigo, M. Birattari, C. Blum, L. M. Gambardella, F. Mondada, and T. Stützle, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 83–94, 2004.

[27] S. Bitam, M. Batouche, and A. Mellouk, "QoSBeeManet: A new QoS multipath routing protocol for mobile ad-hoc networks," in 2010 IEEE Globecom Workshops, Miami, FL, USA, pp. 1648–1652, 2010.

[28] D. Couto, D. Aguayo, J. Bicket and R. Morris, "a high-throughput path metric for multi-hop wireless routing", Wireless Networks, vol. 11, no. 4, pp. 419-434, 2005.

[29] C.E. Perkins, E.M. Royer, S.R. Das, M.K. Marina, "Performance comparison of two on-demand routing protocols for ad hoc networks", IEEE Personal Communications, vol. 8, no. 1, pp. 16-28, 2001.

[30] D. Sharma and K. Sharma, "Grid based multipath energy aware routing protocol for WSNs", International Journal of Computer Trends and Technology (IJCTT), Vol. 4, Issue 4, pp. 729-733, 2013.

Computerized Drug Verification System: A Panacea for Effective Drug Verification

Oketa Christian Kelechi[1], Alo Uzoma Rita[2], Okemiri Henry Anayo[3], Richard-Nnabu Nneka Ernestecia[4]
Achi Ifeanyi Isaiah[5], Chinazo I. Chima[6], Afolabi Idris Yinka[7], Mgbanya Praise Chinenye[8]
Department of Mathematics / Computer Science / Statistics / Informatics
Alex Ekwueme Federal University Ndufu Alike
Ebonyi State, Nigeria

***Abstract*—Computerized Drug Verification System (CDVS) is a research work geared towards establishing the means of identifying authentic drugs in Nigeria with emphasis on identifying the manufacturing date and expiration date respectively using an interactive mobile app. At the point of drug purchase, only few drugs have methods of verifying their authenticity using both Personal Identification Number (PIN) and National Agency for Food and Drug Administration Control (NAFDAC) number. However, the production and expiry dates of such drugs are not always known, hence drugs that have long expired may still report authentic thereby endangering the lives of the consumers. This research addresses the challenges providing an interactive platform for drug verification, with especially strength in its ability to incorporate the manufacturing date, the manufacturer, expiry date and the authentication of drugs. The product of this research, "NAFDAC VERIFY" an interactive mobile and web application can be used in verifying the authenticity of drugs in Nigeria in partnership with Mobile Authentication Service (MAS). This system which runs both on android mobile phones and web-based have been tested using some raw data and the system proves to be very robust and achieves the set out objectives.**

Keywords*—NAFDAC verify; security; drug verification; information system; drug authentication

I. INTRODUCTION

In, today's world, computer has permeated every part of life with its benefits. With this information processing which is the life wire of any organization is not left out in the scheme of computerization of Drug verification. Drugs are chemical substances that are administered to patients for curative purposes; it can also be called medicine because it is the essential part of people care. In recent years, rapid growth in fake drugs incident has claimed a lot of lives in Nigeria which led to the establishment of National Agency for Food and Drug Administration and Control (NAFDAC).

NAFDAC is a Nigerian government agency and they are responsible for controlling and regulating the manufacture, importation, exportation, advertisement, distribution, sale and use of food, drugs, cosmetics, medical devices, chemicals and prepackaged water. Its creation was inspired by a 1988 World Health Assembly resolution requesting countries to "help in combating the global health threat posed by counterfeit pharmaceuticals, and amidst growing concerns about the growing problem of fake and poorly regulated drugs in Nigeria" [1]. In 1989, over 150 children died due to error in a drug formulation. Consequent upon these problems, the Federal Government of Nigeria established NAFDAC to help create a fake-drug free environment in the country. NAFDAC's effort in this direction and in pursuance of the nation's vision for its establishment has been commendable, but researchers still wonder why there remain many unanswered questions: "why does Nigeria still have in existence open drug markets? Why do Nigerians in Drug business breech the stipulated drug laws and get away with it, and continue with their business, committing mass murder and smiling to their banks? How long are we to fight the battle of fake drug even with the threats on the lives of those who fight to preserve the health of the nation?" [1].

NAFDAC has a Mobile Authentication Service (MAS) that enables people to check whether a drug is original or fake with their mobile phones. NAFDAC Mobile Authentication Service (MAS) scratch and text authentication codes were launched in 2010 by NAFDAC's Director General, Dr. Paul Orhii to save consumers from the menace of fake drugs. However, the verification only shows that the drug in question is either authentic or fake without displaying the manufacturing and expiration dates respectively. NAFDAC MAS is not also enabled for all drugs as it is limited to certain drugs; mainly anti-malaria and antibiotic drugs. This research not provide a holistic solution to all the problems of drug faking in Nigeria. However, it carves a niche in providing means by which fake drugs and even genuine drugs which have expired can be tracked, knowing fully well that expired drugs are as harmful as their fake counterparts.

In Nigeria setting, a dubious drug manufacturer can preferably use the NAFDAC number of one drug for another, and the mode verification of drug authenticity does not include the status of the drug in respect to its manufacturing date and expiration date. Individuals who use these drugs are very reluctant to use verify the status of their drugs due to the non-interactive nature of the current drug verification system. At the point of drug purchase, there are no known methods of verifying; how genuine an assumed NAFDAC number is, the manufacturing date and expiry date of a drug. No working method to access NAFDAC database to know if a given NAFDAC number corresponds with the name of the drug in question. This research work is undertaken to uncover some of the problems with conventional drug reporting and verification systems.

II. Statement of the Problem

In Nigeria setting, a dubious drug manufacturer can preferably use the NAFDAC number of one drug for another. At the point of drug purchase, there are no known methods of verifying; how genuine an assumed NAFDAC number is, including the manufacturing date and expiry date of drugs. No working method to access NAFDAC database to know if a given NAFDAC number corresponds with the name of the drug in question. This research work is undertaken to uncover some of the problems with conventional drug reporting and verification systems with emphasis on production dates and expiration dates of such drugs and beverages.

III. Objectives of the Study

In view of the problems mentioned above, the aim of this project is to design and implement a computerized drug verification system. With the following objectives in mind:

1) To develop an android application and a Web-based application to query the NAFDAC drug database for verification using verification pin.

2) To develop a web-based NAFDAC application for drug manufacturers registration, drug registration and verification pin generation.

3) To use a single pin (verification pin) to get complete information about a given drug.

4) To review literatures on the effects of drug faking on the Nigeria populace.

5) To create awareness on the negative implications of expired drugs and its unchecked circulations.

6) To carry out a SWOT analysis on the existing system in order to build a strong system that will stand the test of time.

IV. Review of Related Literature

Drug Faking is a worldwide health problem as the effects can be felt by both the manufacturing country of such drugs and the recipient countries. Olike, [1] noted that "National measures for combating of fake drugs in country might be insufficient because of the advanced sophistications of those who manufacture and sell them" [1]. Furthermore, Beverley Glass asserted that "product counterfeiting is increasing worldwide, both in terms of the volume, level of sophistication, and in the number the countries affected, both in the developed and developing world" [2]. Till date, Nigeria is not excluded in the challenges of fake drugs. Some people still prefer to undertake self-medication when they are ill, and often times the drugs are purchased from unlicensed drug vendors, whose drug quality is not certain. Moreover, fake drugs proved the main factor in contributing to high rate of deaths. "Over 150 children died in 1989 as a result of a formulation error in a drug. Such problems led to the establishment of NAFDAC, which would help create a fake-drug-free environment. The reason was to ensure effective registration of good quality drugs that are inexpensive in Nigeria. In April 2001, Professor Dora Akunyili the Director General of NAFDAC worked hard in combating the problems of sale of fake drugs" [1]. Dora Akunyili's comment on the menace of drugs faking and its negative impact on the Nigeria was captured properly by Olusegun, [3], "the problems of Fake drugs have embarrassed our healthcare providers and denied the confidence of the public on the nation's healthcare delivery system". The result of fake drug proliferation has led to treatment failures, organ dysfunction or damage, worsening of chronic disease conditions and the death of many Nigerians. The situation became so bad that even when patients were treated with genuine drugs, there were no response due to resistance caused by previous intake of fake drug*"* [4].

Fake drugs comes in variant forms and styles as observed by Buowari [4], "A counterfeit drug may contain inappropriate quantities of active ingredients, or none, may be improperly processed within the body (e.g., absorption by the body), may be supplied with inaccurate or fake packaging and labeling, or may contain ingredients that are not on the label (which may or may not be harmful). Medicines which are deliberately mislabeled to deceive consumers including genuine generic drugs but mislabeled are otherwise counterfeit. Counterfeit drugs are related to pharmacy fraud. Drug manufacturers and distributors are increasingly investing in countermeasures, such as traceability and authentication technologies, to try to minimize the impact of counterfeit drugs" [5].

The high occurrence of fake drugs in Nigeria is as a result of the arbitrary ways by which importation licenses on drugs were issued to people by then politicians and military leaders in the 80's, without considering the eventual implications their actions could bring on the health of the public. As the drug market boomed and competition for market share increased, some of the beneficiaries of drug importation licenses looked at the option of importing fake products in order to have a favorable margin over their competitors. "In Nigeria today, it is common knowledge that drugs are treated as general merchandise, which can be obtained easily from open markets, moving vehicles, faceless medicine stores, ferries, and even in the provision stores. This is because the drug distribution business has been left in the hands of non-professionals who just want to make profit at the expense of the consuming public. Poor people are faced with a confusing myriad of health providers and drug sellers" [1].

In February 2006, World Health Organization (WHO) established the first global partnership known as the International Medicinal Products Anti-Counterfeit Taskforce (IMPACT). This enabled them to create awareness and action in the fight against fake drugs. IMPACT focuses on key areas such as

- "Regulatory implementation approach to ensure that standards for quality, safety and efficacy are implemented and distribution chains effectively controlled.
- IMPACT will also help countries in monitoring borders through co-ordination of action between customs, police and the judiciary by working with the World Customs Agency, INTERPOL, and informal networks of enforcement officers.

- Creating stronger legislation that will help empower those who deal with counterfeits and counterfeiters in the course of their work." [6].

However, improving International interaction especially to the developing world in order to identify the actual source of fake medicines tormenting their markets. IMPACT will also create International information networks to strengthen and monitor good trafficking, issue alerts from country to country.

A. Technology Previously Used

Most countries have developed strategic means of combating the prevalence of food and drug counterfeiting. The Mobile Authentication Service (MAS) works with mobile phones that are SMS enabled. It allows anyone with mobile phones to check whether drugs are fake or original without direct contact with the manufacturer. This service was introduced into Nigeria by National Agency for Food and Drug Administration and Control (NAFDAC) in response to the increasing rate of fake drugs distributed in the country. It is an attempt to turn mobile phones into tools that can help check this problem as many lives have been claimed and many are still endangered by the circulation of such fake products. The Mobile Authentication Service (MAS) uses Truscan technology which enables Consumers to input a twelve-digit number scripted on the back of the packaging of a drug and send the same via SMS to a dedicated number which returns a Negative or Positive response about the authenticity and genuineness of the product. This service was made possible through a partnership between NAFDAC, GSM operators and as well the representatives of national and international pharmaceutical companies in the Nigeria. Recently, the SMS drug verification was confirmed to be an efficient method in some Nigeria, but it is yet to be implemented on all drugs [7].

V. The M.A.S Experience in Nigeria

The National Agency for Food and Drug Administration and Control (NAFDAC) on February 2, 2010, launched the NAFDAC Mobile Authentication Service (MAS) with the aim of putting the power of checking for originality of product in the hands of consumer. The National Agency for food and Drug Administration and Control (NAFDAC) has made the implementation of MAS compulsory by all Pharmaceutical companies for all their major products especially the Antibiotics and the Anti-malarial drugs. The official deadline for implementation was set for 2nd of March, 2013. It was hoped that this will be implemented and definitely allow Nigerians to check the authenticity of their medicines and eliminate fake drugs from circulation. To date it is clear that out of every 200 Nigerians only 1 knows about MAS representing 0.5% irrespective of the fact that the service is free. Well over a million text messages have been sent by Nigerians and millions of products now carry the unique sproxil label to show that such product is MAS ready (sproxil.com/mas.html). In February 2011, GlaxoSmithKline in partnership with National Agency for Food and Drug Administration and Control (NAFDAC), ran a six-month pilot anti-counterfeiting programme in Nigeria and in all, a total of 145,000 texts from 115,000 unique users were received. This figure represents approximately 10% of use from the total products sent out for the pilot programme. Ninety percent of text returned a genuine confirmation, 2.5% received a counterfeit alert and other received a message indicating a duplicate PIN.

A. How to use NAFDAC MAS

The step by step method that can be applied by the consumer even at the point of purchase of some selected drugs in order to verify their authenticity and genuineness via mobile phone are as follows:

1) "Scratch the sproxil label on the product to be purchased to reveal a unique 12-digit PIN.

2) Text the Unique 12-digit PIN as SMS to the short code 38353.

3) Immediately receive a response in form of SMS on your phone affirming the validity of the PIN either positively or negatively. With the support of all mobile network provider" [8] as seen in Fig. 1.

B. Functions of NAFDAC in Nigeria

1) Ensures compliance to standard: Ensuring compliance to standard is the major functions of the National Agency for Food and Drugs Administration Control (NAFDAC). Therefore what the National Agency for Food and Drugs Administration Control does is to carry out test in locally manufactured drugs and food and sometimes, those foods and likes imported into the country with the view to ensure that those products meet specified standard. There is designated standard specification to every product manufactured within Nigeria or imported into the country, there must be total consent by the council and until these product are approved by the council, such products is not consumable and it ensures effective quality control of these goods, such as food, and drugs, and cosmetics and medical devices, and packaged water, and chemicals, etc.

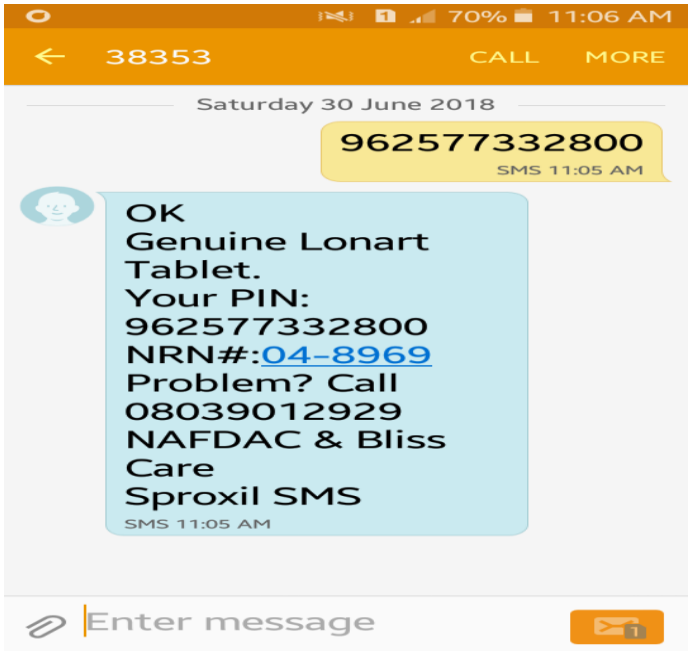


Fig. 1. Interface of NAFDAC MAS.

2) Inspects imported food: Some of the goods imported into the country include, imported foods, imported drugs, imported cosmetics, imported bottle water, and imported chemical among others. NAFDAC ensures that what comes into the country meets the standard, the specification and make sure that there is an establishment of the relevant quality assurance system. They also ensure that, there is certification of the products and regulated product.

3) Investigates into production houses: Ensuring that premises, where these good are produced are in order and hygienic has been the primary function of the National Agency for Food and Drugs Administration Control. From time to time, the Agency will embarked on investigation tours to these places, where these goods are manufactured, and its investigation into these production houses is always appropriate and proper without leaving any stone unturned, the Agency will inspect and investigate raw material of these goods, be it drugs, food, cosmetics, etc. NAFDAC would check all accurately.

4) Registration of food and drug products: No products within this sector can thrive without National Agency for Food and Drugs Administration Control not knowing about such product, that's why NAFDAC itself has been registering these products.

5) Controls exportation: Exportation control is one of the function of the National Agency for Food and Drugs Administration Control, so far such product for exportation is product that is controlled and regulated by the National Agency for Food and Drugs Administration Control (NAFDAC). And they make sure that quality certification of all these products for exports are issued.

VI. Analysis of NAFDAC MAS

The National Agency for Food and Drug Administration and Control (NAFDAC) established in January 1993, is responsible for regulating and controlling the manufacture, importation, exportation, advertisement, distribution, sale and use of food, drugs, cosmetics, medical devices, chemicals and packaged water. To achieve this aim, this agency has deployed several means such as demanding all manufactured drugs to be registered and assigned a NAFDAC registration number.

The next challenge facing drug end-users is how to verify the authenticity of a registration number imprinted on drugs. To achieve drug verification, the agency as well as other third-party organizations have to work in collaboration to achieve proper drug verification. On February 2, 2010, NAFDAC and Sproxil launched the NAFDAC Mobile Authentication Service (MAS), putting the power of product verification right in the hands of the consumer as shown in Fig. 2. MAS is powered by Sproxil's award-winning cloud-based Sproxil Defender™ technology, and remains the world's largest nation-wide implementation of consumer-facing SMS anti-counterfeiting technology in the world (Sproxil®, 2018). Others including mPedigree, PharmaSecure etc, also provide drug verification alternatives to drug end-users in Nigeria (Nigerian Law Intellectual Property Watch Inc., 2018).

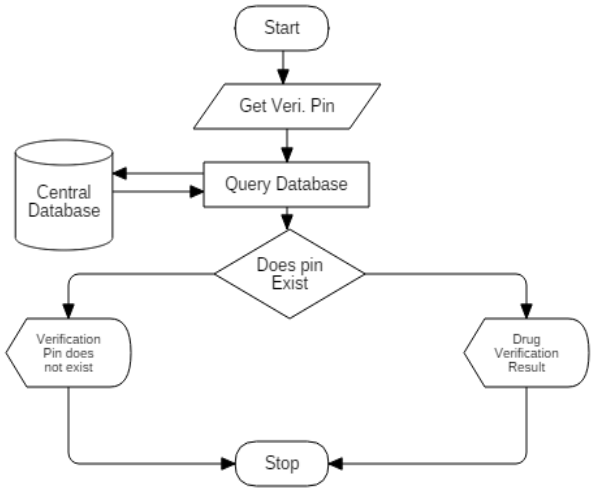


Fig. 2. Flowchart for NAFDAC MAS.

This system is a drug verification system built on the android operating system. This works in parallel with a database of drug manufacturers, drugs and drug verification pins representing the NAFDAC database hosted on a server online, from where the mobile application queries to supply user with the authenticity information of a given drug.

The NAFDAC database is populated with data from a NAFDAC admin web application. This web application is deployed to do the following:

Drug Manufacturers Registration

Drug Registration

Drug Verification Pin Generation

Drug Registration Slip Printing

Overview of Manufacturers

Overview of Registered Drugs

High level model of the developed system Fig. 3 shows a high level model of the developed system.

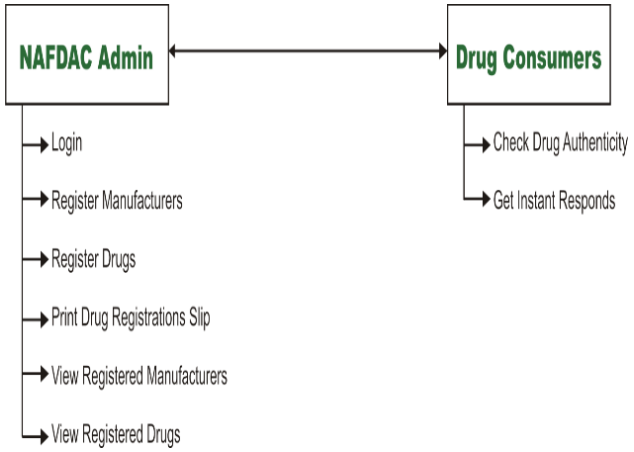


Fig. 3. High Level Model of the Developed System.

VII. System Architecture

A set of artifacts (that is: principles, guidelines, policies, models, standards, and processes) and the relationships between these artifacts, that guide the selection, creation, and implementation of solutions aligned with business goals [9]. Fig. 4 is a system architecture diagram of the developed NAFDAC VERIFY which is meant to partner with other platforms for effective drug verification.

A. Use Case Diagram

Use Case Diagram is used to represent how actions are performed by the different characters and to whom/what the actions are performed. It is a mimic of the use of the system in real-time. Fig. 5 shows the interaction between actors/users of the system and the system itself.

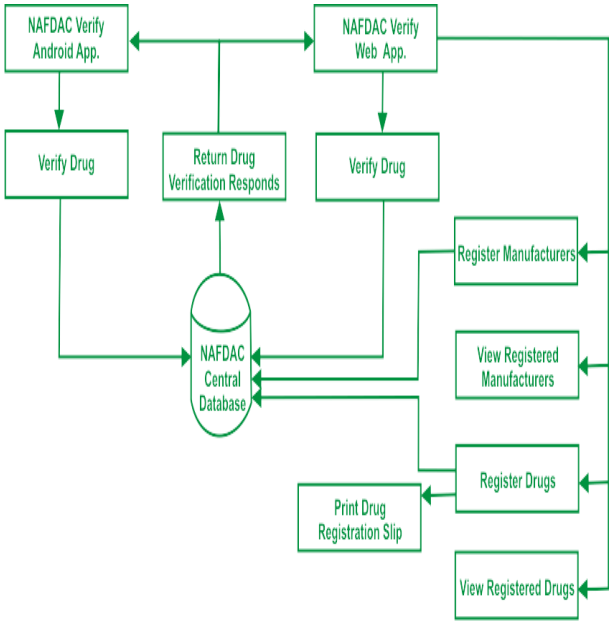


Fig. 4. System Architecture Diagram of the NAFDAC Drug Verification System.

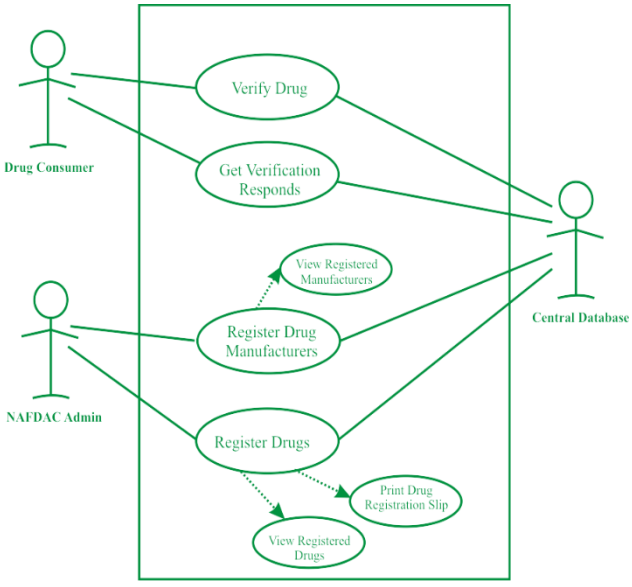


Fig. 5. The use Case Diagram of the System.

1) Entity Relationship Diagram (ERD): Entity Relationship Diagram as shown in Fig. 6 provides a clear view of the relationships among the different entities associated with the system. This gives one idea of the flow of action performed by the system entities.

B. Menu Design

The Home Page of the system is depicted in Fig. 7 while Fig. 8 shows the process decomposition of the drug verification system and provides a navigation to several aspects of the system that work in synchronization with each other to achieve the main aim.

This menu serves as an authentication to NAFDAC admin panel. The authentication was implemented using PHP and MYSQLi. This menu is most concerned with security, therefore certain precautions were considered such as preventing Structured Query Language (SQL) injection. SQL injection is a malicious use of the SQL statement in a form. The attacker consciously login using a database query, this may grant them access to a system is the query successfully implement.

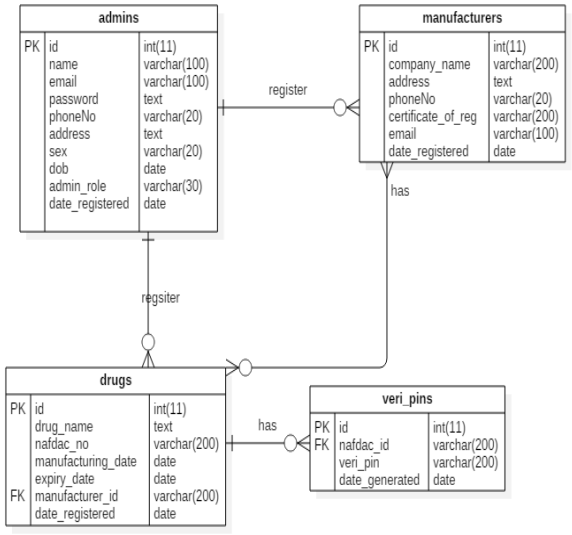


Fig. 6. Entity Relationship Diagram (ERD) of the New System.

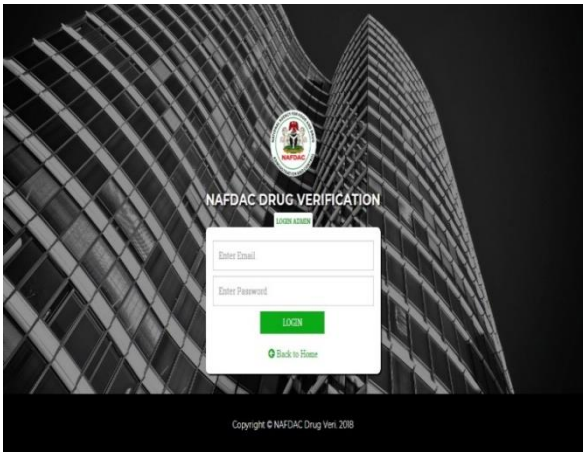


Fig. 7. The Admin Login Page for NAFDAC Staff in Order to Register New Drugs.

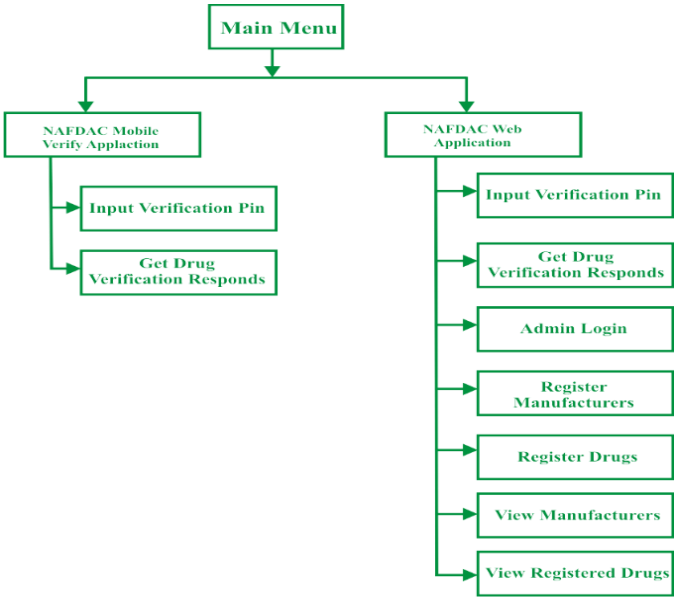


Fig. 8. Main Menu Design Showing the Functionalities of the Systems.

2) Subsystem design: The entire system is made up of subsystems whose functionalities help to achieve the goal of the system. The subsystems include:

a) The Android Platform Sub System as shown in Fig. 9 is built to get input data into the system in the form of verification PIN. It runs on the Android Operating System.

b) The Web Platform Subsystem as shown in Fig. 10 is designed for both drug verification and the NAFDAC Database Management System.

3) Program module design: The proposed system is made up of several modules which make up the entire system. Few of these include:

- Admin Login Module

The login module is based on the web application. It serves as a security restriction to the NAFDAC database operation. It is an access to the following admin privileges:

a) Drug manufacturers registration

b) Drug registration

c) Overview of registered drug manufacturers

d) Overview of registered drugs

e) Printing of drug registration slip containing drug verification pins.

f) Reprinting of drug verification pins for a given drug.

- Android Drug Verification Module

This module is based on the android operating system. It supports drug consumers who choose to use the android application for drug verification. It also queries the central database if the inputted drug verification pin is valid, if valid it gives the entire details about the drug including the manufacturer name, NAFDAC number, manufacturing date, expiry date, originality status (expired, original); else if it is not valid it gives a pin not found responds.

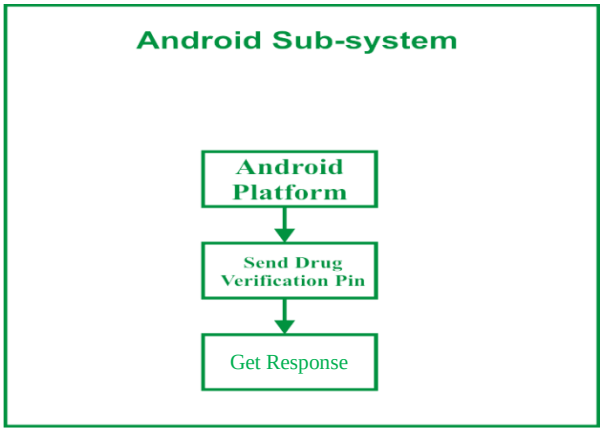


Fig. 9. The Android Platform.

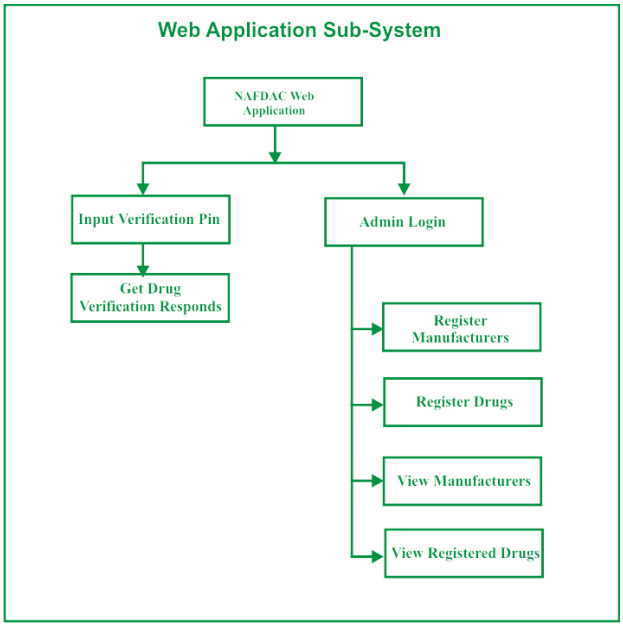


Fig. 10. The Web Platform.

4) Mobile drug verification menu implementation: The mobile drug verification menu shown in Fig. 11 was designed to enable drug verification PIN collection and initiate a query to the central database. This was implemented using java as a programming language on the android operating system. The HttpURLConnection was used to pass the verification pin data through the web hosted Application Programming Interface (API) http://mirarticles.com/nafdac /api/verify.php. The API processes the verification pin and queries the central database, then returns responds to the android application for display.

C. Register Drugs Menu Implementation

This is a menu that enables the NAFDAC admin do a drug registration. It was designed to accept inputs such as drug name, number of drugs, NAFDAC number, manufacturing date, expiry date and manufacturer. When a registration is initiated through the register button, a drug verification pin is generated following a predefined algorithm.

Fig. 11. Mobile Drug Verification Menu Implementation.

```
109  function newVeriPin(){
110      GLOBAL $con;
111      //GENERATE NEW VERI. PIN CONTAINING 12 DIGITS
112      $n1 = rand(10,99); //GENERATE 2 DIGITS
113      $n2 = rand(0,9);  //GENERATE 1 DIGITS
114      $n3 = rand(0,9); //GENERATE 1 DIGITS
115      $n4 = rand(100,999); //GENERATE 3 DIGITS
116      $n5 = rand(10,99); //GENERATE 2 DIGITS
117      $n6 = rand(0,9); //GENERATE 1 DIGITS
118      $n7 = rand(0,9); //GENERATE 1 DIGITS
119      $n8 = rand(0,9); //GENERATE 1 DIGITS
120
121      $veri_pin = $n1.$n2.$n3.$n4.$n5.$n6.$n7.$n8; // 12 digits veri pin
122
123      //CHECK DRUG PINS IN DATABASE IF PIN IS ALREADY GENERATED
124      $result = mysqli_query($con, "SELECT * FROM veri_pins WHERE veri_pin = '$veri_pin'");
125      if ($result->num_rows>0) {
126          //VERI PIN EXISTS
127          //RECURSE
128          $this->newVeriPin();
129
130      }else{
131          //VERI PIN DOES NOT EXIST
132          return $veri_pin;
133      }
134  }
```

Fig. 12. The Drug Verification Pin Generation Program Code.

The code, part of which is shown in Fig. 12 comprises a concatenation of 8 variables, each generating random numbers. n1, n2, n3, n4, n5, n6, n7, n8 generates 2, 1, 1, 3, 2, 1, 1, 1 digit(s) respectively, this forms a total of 12 randomly generated digits called the verification pin. From line 124 of the code, the generated PIN is checked if it already exists to avoid duplication of the same PIN. If it already exists, the function will call itself again in line 128. The flowchart for the drug verification PIN generation is shown in Fig. 13.

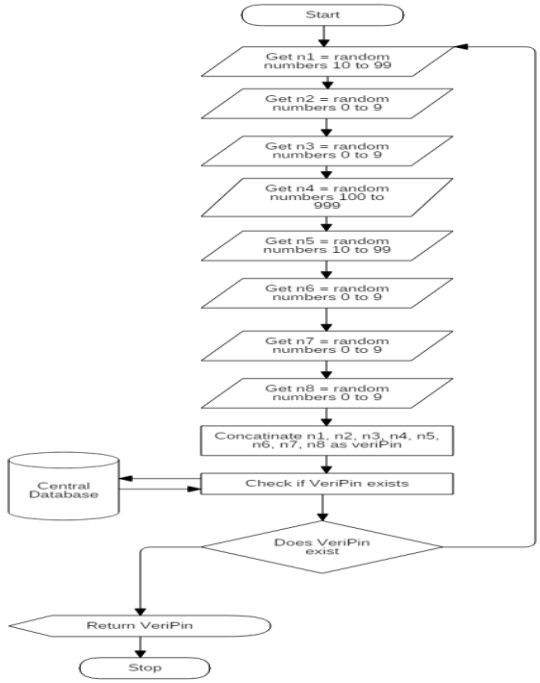


Fig. 13. Flowchart Diagram for Drug Verification PIN Generation.

D. Subsystem Implementation

1) The mobile drug verification subsystem: The mobile drug verification subsystem is a part of the entire drug verification application which was designed to enable drug verification pin collection and initiate a query to the central database. This was implemented using java as a programming language on the android operating system. The HttpURLConnection was used to pass the verification pin data through the web hosted Application Programming Interface (API) http://mirarticles.com/nafdac/api/verify.php. The API processes the verification pin and queries the central database, then returns response to the android application for display. Fig. 14 shows the drug verification API code.

2) The web subsystem implementation: The web subsystem was implemented and designed using HTML, CSS, JavaScript, PHP and MYSQL.

3) Program module implementation: In this section we paid more attention on the underground implementation of the program module: Android Drug Verification Module.

4) Android drug verification module implementation: The drug verification pin was collected using the android method called findViewById, and in precision we used veripin = (EditText) findViewById(R.id.veri). veri is the id given to the drug verification pin input field in the android application. The data was immediately authenticated if it is empty using the java code Veripin.trim().equals(""). If empty, an android toast is made to display an error message, else a class called Background which extends the AsyncTask predefined class was called. The AsyncTask helps some processes to run at the background of the application. These processes include using the java httpURLConnection to send the verification pin

through the apihttp://mirarticles.com/nafdac/api/verify.php. Once a responds text is returned, it is sent to a new android activity class called Result.java. Here the result is interpreted and displayed to the user.

The response text is a string of results separated with "|". It looks like thisstatus|drug_name|nafdac_no|manufacturer_name |manufacturing_date|expiry_date|veri_pin. This result is interpreted and formatted for user's view.

```
<!--
 * Copyright (c) 2018 Tutapis
 * @license MIT
 -->
<?php
  include('../class.php');
  $class = new main();
  $class->db_connect();

  //Receive Verification Pin From Android
  $veri_pin = $class->purify($_POST['veri_pin']);

  //Query the Drug Database
  $query = $con->query("SELECT * FROM veri_pins WHERE veri_pin = '$veri_pin'");
  if ($query->num_rows == 0) {
    //No Result Found for this verification Pin
    echo "Verification Pin Does not Exist.";
  }else{
    //Verification Pin Exists
    //Return Results Associated with this Veri_pin
    $row1 = $query->fetch_assoc();
    $nafdac_id = $row1['nafdac_id'];
    //USe the nafdac ID to get the information about the drug

    $drug_name = $class->getDrugDetailsByNadfacId($nafdac_id,"drug_name");
    $nafdac_no = $class->getDrugDetailsByNadfacId($nafdac_id,"nafdac_no");
    $manufacturing_date = $class->getDrugDetailsByNadfacId($nafdac_id,"manufacturing_date");
    $expiry_date = $class->getDrugDetailsByNadfacId($nafdac_id,"expiry_date");
    $mname = $class->getManufacturerDetailById($class->getDrugDetailsByNadfacId($nafdac_id,"manufacturer_id"));
    $manufacturer_name = $mname['company_name'];

    //Check if this drug has expired
    $today = date("Y-m-d");
    if($class->dateDifference('d', $today, $expiry_date, $relative=true) < 1 ){
      $status = "Expired";
    }else{
      $status = "Genuine";
    }

    $return_data = $status."|".$drug_name."|".$nafdac_no."|".$manufacturer_name."|".$manufacturing_date."|".$expiry_date."|".$veri_pin;
    echo $return_data;
  }
?>
```

Fig. 14. The Drug Verification API Source Code.

VIII. Test Result

These are outputs of the workability of the tested system. They give evidence of a fully implemented system when tested with real-time data. Fig. 15 and 16 shows a sample output for a genuine and expired drug verification on the web application respectively. Furthermore, Fig. 17 shows a sample output of registered drugs database with the registered PINs.

Fig. 15. An Output Sample for a Genuine Drug Verified on the Web Application.

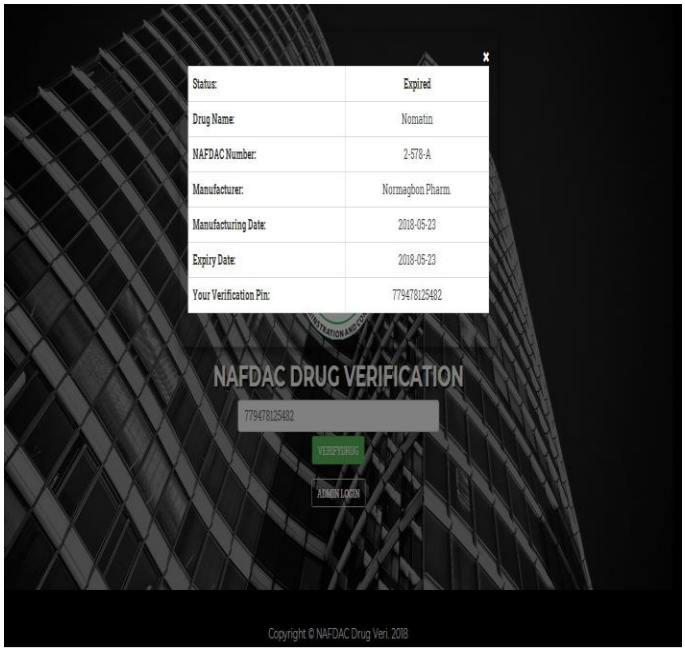


Fig. 16. An Output Sample for an Expired Drug Verified on the Web Application.

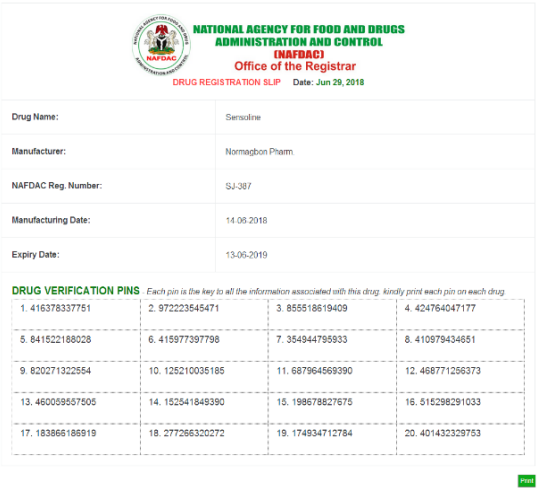


Fig. 17. Test Result for a Registered Drug with the Verification PINs.

A. System Integration

System integration as shown in Fig. 18 depicts how the various subsystems were bought together to form the main system. The following were integrated:

a) The Android Verification Application
b) The NAFDAC Web Application
c) The Central Database

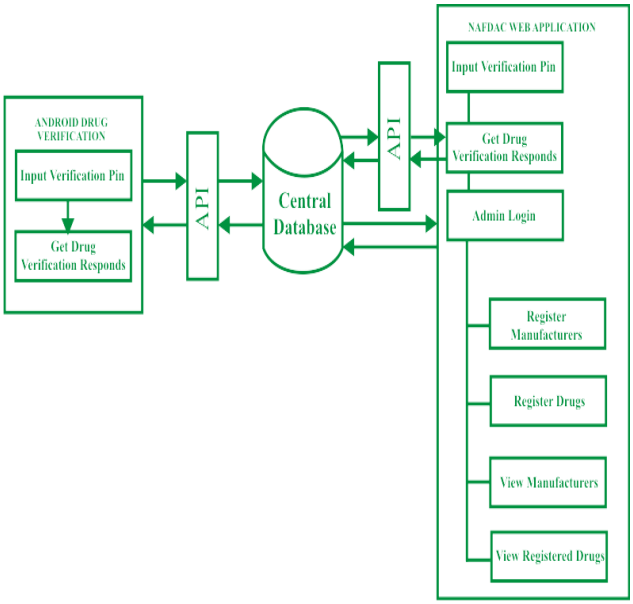


Fig. 18. An Integration of the System Subsystems.

IX. Conclusion

The challenges of verifying fake drugs and the rampant distribution of expired rugs and beverages which cause hazard to the health of humans, led to this research. Despite the existence successive strategies to eliminate consumption of fake and expired drugs. The anchor points of the crime of fake and expired drug production distribution can be tackled by the introduction of a new drug verification system that is centered on absolute and effective verification of all drugs using an interactive mobile and web application. Therefore, the following were put in consideration during this research:

1) Verification beyond NAFDAC number, knowing that the correctness of a NAFDAC number does not declare the drug originality. It was discovered that many manufacturers print the NAFDAC number of an original drug and replicate same on fake ones, capitalizing on the reluctance of the consumers to scratch and verify.

2) Drug Information centered verification which pulls out the real information of the drug in the database for comparison with the one in question provides in-depth information such as if a drug has expired at the point of verification or if the verification pin does not exist.

3) To provide a real-time drug verification platform, this research has provided alternative means for consumers to verify the authenticity of their drugs including that of partnering with a third party technology company Sproxil® that heralds the Mobile Authentication Service (MAS). This research was made to advance these existing strategies of the agency to ensure a consumer oriented drug verification in real-time and covers the loopholes of the preceding systems.

References

[1] Olike, "The Fight Against Fake Drugs by NAFDAC in Nigeria," in 44th International Course in Health Development, Vrije Universiteit Amsterdam, 2007-2008.

[2] G. D. Berverly, "Counterfeit Drugs and Medical Devices in Developing Countries," Dove Press, vol. V, pp. 11-22, 2014.

[3] A. Olusegun, "Counterfeit drugs in Nigeria: A threat to public health," African journal of Pharmacy and Pharmacology, vol. VII, no. 36, pp. 2571-2576, 2013.

[4] O. Y. Buowari, "Fake and Counterfeit Drug: A review," AFRIMEDIC Journal, vol. 3, no. 2, pp. 1-4, 2012.

[5] A. C. Emmanuel, Deborah, O.O. and E. N. Gloria, "Food And Drug Counterfieting In The Developing Nations; The Implications and Way-Out," Academic Research International, vol. III, no. 2, pp. 24-31, 2012.

[6] Impact, Impact: International Medical Products Anti-Counterfeit Taskforce, Roma: Agenzia, Roma: Intaliana del Farmaco, 2006.

[7] C. Obinna and G. Oluwale, "NAFDAC introduces new anti-counterfeiting technologies Nigeria," Nigerian Vanguard, Lagos, 2010.

[8] J. A. Osuolale, "Institut Francais de Recherch Afrique/French Institue for Research in Africa," 2015. [Online]. Available: https://www.ifra-nigeria.org/files/44/Habitele/22/Joseph-Ayodokun:-Identification-drugs-MAS.pdf. [Accessed 31 May 2019].

[9] F. Jeane-Claude, "New York University," 2011. [Online]. Available: www.nyu.edu>jcf>session2. [Accessed 20 May 2019].

Empirical Analysis of Context based Content Delivery for M-Learning Scenarios using ANFIS

Sudhindra B.Deshpande[*, 1]
Department of Information Science & Engineering
Gogte Institute of Technology
Belagavi, Karnataka
India

Shrinivas R.Mangalwede[2]
Department of Computer Science & Engineering
Gogte Institute of Technology
Belagavi, Karnataka
India

Abstract*—Today's world is unusually popular with the internet and mobile devices in everyday life. It offers unprecedented possibilities learning with mobility. This kind of learning can be called "M-learning" (Mobile Learning) at any point in the world. Meeting learners ' needs in the current scenario in which collaborative electronic and mobile education systems have become more evolving. Every learners ' needs differ in terms of learning; context, content, learning styles, speed of learning, even including preferences, places. Mobile learning enables the learner to learn while moving, enabling the learner to learn in any time and any place. Learning styles have evolved along with advances in technology; specifically advances in mobile technology and mobile networks. Portable devices such as mobile phones, tabs, iPods, etc. are commonly used today by all. The way we learn has been changed with the use of these devices in education. In M-learning environment the learner has access to the contents everywhere and every time through mobile devices. Customization and learner context awareness are the important factors in providing the learner with relevant content. Appropriate content delivery based on a learner's context is a complex process. So a content delivery system is needed that takes into account learners ' needs such as context, style and devices features. To model such a system neural network with fuzzy reasoning can be used, to accommodate the dynamically changing learning styles, contexts and characteristics of smart device. "If-then" conditions can be formed to build the suggestion rules required for such a content delivery system. ANFIS i.e. Artificial Nero Fuzzy Inference System is an integral asset to create fuzzy systems with IF-THEN guidelines. To model and analyze this type of context aware and adaptive content delivery system for an M-learning environment, ANFIS can be used. In this article, use of ANFIS tool is demonstrated for various m-learning scenarios with different contexts. Four different contexts are constructed based on the inputs given by the student learners. Using ANFIS these four scenarios have been analyzed empirically for their performance based on the RMSE of various membership functions.

Keywords*—Personalization; adaptive learning e-learning; distance learning; mobile learning; M-Learning; ANFIS; fuzzy system; neural networks; personalization; context awareness; content adaption; content delivery; expert system

I. INTRODUCTION

Learning is acquiring the knowledge through study, experience, or being taught. The traditional learning is through physical existence i.e. a teacher used to deliver lectures in classroom and through group discussions, white board or with slide projections, teaching learning process happens in classroom. To improve the learning ability, students need to ask questions, take notes and write assignments. Here learning is restricted only to books and teachers.

In recent years Internet and information technology has brought great changes into learning phenomenon. A verity of technological advancements like wireless communication and mobile devices have emerged and revolutionized the education. Education, via Internet and intelligent, portable gadgets, is one of the most interesting domains in today's world. The increasing use of computers, mobile devices and networks makes people more attractive in mobile learning. From distance learning to e-learning, and from e-learning to m-learning, has revolutionized the education with technology, reducing the constraints of space and time, reduced costs of learning, particularly increasing the learning efficiency. Applying environmental intelligence in educational scenarios allows students to immerse themselves in a digital environment that is aware of their presence and context. Mobile learning can provide new ways to deliver training, develop talent, and build a workforce with the skills and knowledge needed to meet current and future demands to enhance organizations' economic growth through mobile learning.

Learners have different cultures, education backgrounds, attentions, age, gender and social roles, and interests, making a significant difference in learning behavior. Providing the right ` content and tactics for teaching is very demanding in line with the learners ' needs. With new mobile technology, the education process must change and behave according to needs and interests of individual learner. Context-aware is a computing style in which situational and environmental information about places, people and things is used to anticipate immediate needs and to provide enriched, situation-aware and usable content, functions and experiences proactively. Context awareness originated as a term derived from ubiquitous computing or as so-called pervasive computing that links environmental changes with mobile systems.

Context awareness is a mobile device property that is defined in addition to location awareness. M-Learning tries to combine the possibilities provided by new mobile technologies, wireless infrastructure and E-Learning strategies. Various definitions of M-Learning are as following.

*Corresponding Author

1) M-Learning refers to "learning that may take place in multiple locations, across multiple times and addressing multiple content areas using static or portable equipments, such as wireless laptops, smart phones or personal media players" [1].

2) From pedagogical perspective M-Learning is defined as "any sort of learning that happens when learner is not at a fixed , predetermined location, or learning that happens when learner takes advantage of learning opportunities offered by mobile technologies" [2].

3) M-Learning is one new learning mode that users can use the mobile communication terminals to assist them to learn [3].

4) M- Learning can be defined as "the ability to learn independently without any constraint of place and time, facilitated by a range of mobile devices such as mobile phones, PDAs, iPod, pocket PCs and blackberries" [4].

For such m-learning system one more important aspect is content delivery. The formats of content should vary according to the context the learner is in. Delivery of the content depends on the context of a learner. The context may be location, time, network speed, battery and noise etc. According to the context of the learner different formats of the content need to be delivered. The formats of content may be in audio, video or text/pdf format.

To develop adaptive content delivery system, rules need to be framed. These rules can be treated as suggestions or circumstances to deliver the different format of the contents. To frame the rules, there is a need to find a technique which assist in building them; also to find the behaviour and performance. ANFIS i.e. Artificial Neural Fuzzy Inference System can be used to simulate the system; through which performance of various models and scenarios can be compared and measured.

II. LITERATURE

In today's society, which is considered as a Mobile Communication Society, mobile technology is also adapted in education. Qin Shuai and Zhou Ming-quan [5] in their paper propose M-Learning in cloud will have lot of opportunities. According to them adapting existing E-Learning services and didactic content to M-Learning is very much challenging in nature.

Yuan Jiugen,Xing Ruonan and Wang Jianmin [6] in their paper highlighted the importance of M-Learning in teaching. They discuss an analysis of the application of M-Learning in education in teaching with respect to learners' age and styles.

Jane Y-K and Mike Joy [7] have discussed about personalization in context aware M-Learning depending on learning preferences of learners. They have discussed varieties of scenarios of mobile learners to illustrate different M-Learning preferences. The scenarios distinguish between different learners' preferred locations, their preferences for the levels of noise/distractions in a location and time of the day to conduct their studies. They propose that building these preferences is challenging work.

Yi Jin [8] in his paper proposes and exemplifies a framework for M-Learning platform through a layered module and he explains about eight feasible functions including course learning, teaching, collecting of content, training of mobile education, outdoor task supporting , Q&A , blogs and games. Author concludes that there is requirement of new applied functions to be designed for existing system according to the new need of user.

Hsuan-Pu Chang [9] has proposed four modules to carry out the adaptive course caching and adaptive course presentation strategies. Author discusses about creating a package of learning material that will play on number of learning management systems.

Gwo-Jen Hwang and his co-authors [10] in their paper have investigated the effects of mobile learning model that integrates real world and digital world resources on the cognitive load and learning achievement of learners.

Fadi R.Shahroury [11] discusses in his paper that concept of data mining in academics is still at early stages and aims to highlight the potential of data mining in M-Learning and suggests various data mining tools that can be beneficial for M-Learning.

R.Madhubala and A.Ahila [12] have proposed a Context Aware and Adaptive Mobile Learning. It includes the detailed survey about context aware and adaptive m-learning, which serves as the base for new researches in this area. The system uses the clustering technique for generating adapted and personalized content for the user. Mobile learning helps the user to learn the course/subject at anytime and anywhere and also provides the unique experience to the learners in terms of its flexibility. In order to provide effective learning contents to the users, both device and learner context should be considered.

Richard A. W., *et al.* [13] have proposed a considering learning styles and context-awareness for mobile adaptive learning. An approach for providing mobile, personalized course content tailored to each individual's learning style while incorporating adaptive context awareness. The respective approach has been implemented as an iOS application and the evaluation with 45 students show that students were able to improve their comprehension of a subject matter by 23% after using the application.

Nikolaos, Polatidis, *et al.* [14] have proposed a Privacy-preserving recommendation in context-aware mobile environments. It was focused on user privacy by providing a method for context privacy-preservation and privacy protection at user interface level. Thus, a set of algorithms that are part of the method has been designed with privacy protection in mind, which was done by using realistic dummy parameter creation. To demonstrate the applicability of the method, a relevant context-aware data set has been used to run performance and usability tests. The method has been experimentally evaluated using performance and usability evaluation tests and was shown that with a small decrease in terms of performance, user privacy can be protected.

Mario, Casillo, *et al.* [15] have proposed a Context-Aware Mobile Solution for Assisting Tourists in a Smart

Environment It was introduced an adaptive Context Aware app able to collect not-structured data, belonging to heterogeneous sources and develop tailored recommendations for the user, in order to support a tourist inside a town. The solution found takes advantage of information technologies, like Internet of Thing and Internet of Services and the objective was reached through the use of a system of description of the context through a graphical formalism named Context Dimension Tree.

Singh, Gnana, *et al*. [16] have proposed a Mobile Application for m-Learning. The problem was addressed by m-learning using mobile application. In m-learning approach-based education, the course contents were uploaded to the cloud server by the teacher. Then, the students those who were not able to attend school or college can download the course content through internet using an m-learning mobile application and learn them. Once the course content was downloaded to the mobile devices through mobile application, the student can learn even without internet connectivity. This presents a mobile application for m-learning for the students those who are not able to attend the classes at engineering colleges.

Matlab ANFIS tool [17] [18] is used to demonstrate the Takagi– Sugeno fuzzy inference framework for the content delivery system.

III. Artificial Nero Fuzzy Inference System (ANFIS)

Takagi– Sugeno fuzzy inference framework is used in ANFIS architecture [14] which is an adaptive system that utilizes directed learning algorithm. Fig. 1 demonstrates the architecture of fuzzy reasoning component for Takagi–Sugeno model and ANFIS that consists of two sources of information "x" and "x", and one yield or output "f".

Takagi–Sugeno model follows the two variants of IF-THEN are:

RULE 1 = If 'x' is A1 and 'y' is B1 Then f1 = p1x+q1x+r1

RULE 2 = If 'x' is A2 and 'y' is B2 Then f2 = p2y+q2y+r2,

Where A1, A2 and B1, B2 are the membership elements of each info 'x' and 'y', while p1, q1, r1 and p2, q2, r2 are linear parameters to a limited extent *Then* (subsequent part) of Takagi– Sugeno fuzzy inference model.

Referring to Fig. 1, ANFIS architecture shows five layers. The first layer and fourth layer consist of adaptive nodes, while the remaining layers consist of fixed nodes.ANFIS tool in MATLAB provides Grid partitioning algorithm. The algorithm generates a single-output Sugeno-type FIS by using grid partitioning on the data.

genfis1 generates a Sugeno-type FIS structure used as initial conditions (initialization of the membership function parameters) for ANFIS training. It generates input membership functions by uniformly partitioning the input variable ranges, and creates a single-output Sugeno fuzzy system. The fuzzy rule base contains one rule for each input membership function combination.

ANFIS tool simulates and generates Root Mean Square Error (RMSE) values for data sets which include training, cheking and testing. It also facilitates to use various membership functions.

Membership function is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. They characterize fuzziness whether the elements in fuzzy sets are discrete or continuous. The membership functions trimf, trapmf, gaussmf, gauss2mf and gbellmf are explained below:

1) Triangular function: defined by a lower limit a, an upper limit b, and a value m, where a < m < b, as given by equation (1).

$$\mu_A(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{m-a}, & a < x \le m \\ \frac{b-x}{b-m}, & m < x < b \\ 0 & x \ge b \end{cases} \quad (1)$$

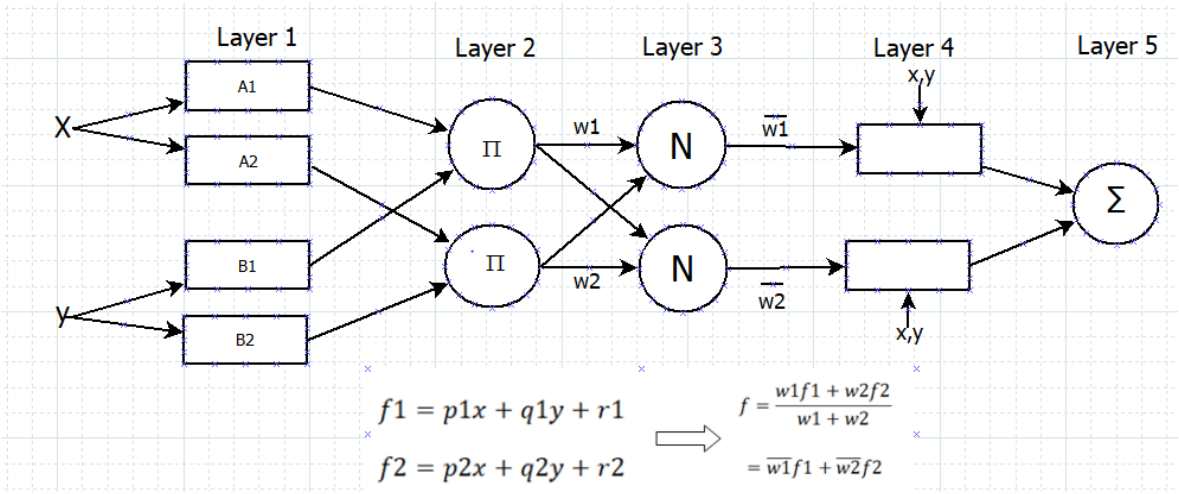


Fig. 1. Architecture of ANFIS with 5 Layers.

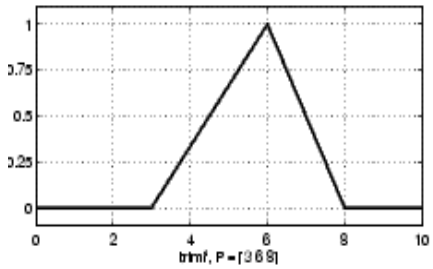


Fig. 2. Triangular Membership Function Curve.

The simplest is the triangular membership function is shown in Fig. 2, and it has the function name trimf. It's nothing more than a collection of three points forming a triangle.

2) Trapezoidal function: defined by a lower limit a, an upper limitd, as given by equation (2).

$$\mu_A(x) = \begin{cases} 0\,, & (x < a)\ or\ (x > d) \\ \frac{x-a}{b-a}\,, & a \le x \le b \\ 1\,, & b \le x \le c \\ \frac{d-x}{d-c}\,, & c \le x \le d \end{cases} \quad (2)$$

Fig. 3 shows the trapezoidal membership function, where a lower support limit b, and an upper support limit c, where a < b < c < d. The trapezoidal membership function, trapmf, has a flat top and really is just a truncated triangle curve.

3) Gaussian function: defined by a central value m and a standard deviation k > 0, as given by equation (3)

$$\mu_A(x) = e^{-\frac{(x-m)^2}{2k^2}} \quad (3)$$

The smaller k is, the narrower the "bell" is. This gaussmf, is built on the *Gaussian* distribution curve: a simple Gaussian curve as depicted in Fig.4.

4) The second gaussian function: depends on two parameters *sig* and *c* as given by equation (4).

$$f(x;\sigma,c) = e^{-\frac{(x-c)^2}{2\sigma^2}} \quad (4)$$

The function `gauss2mf` is a combination of two of these two parameters as shown in Fig. 5; two-sided composite of two different Gaussian curves. The first function, specified by *sig1* and *c1*, determines the shape of the left-most curve. The second function specified by *sig2* and *c2* determines the shape of the right-most curve. Whenever *c1* < *c2*, the `gauss2mf` function reaches a maximum value of 1. Otherwise, the maximum value is less than one. The parameters are listed in the order.

5) The generalized bell function: gbellmf, depends on three parameters *a*, *b*, and *c* as given by equation (5), where the parameter *b* is usually positive and the parameter *c* locates the center of the curve as shown in Fig. 6.

$$\mathrm{f}(\mathrm{x};\mathrm{a},\mathrm{b},\mathrm{c}) = \frac{1}{1+\left|\frac{\mathrm{x}-\mathrm{c}}{\mathrm{a}}\right|^{2\mathrm{b}}} \quad (5)$$

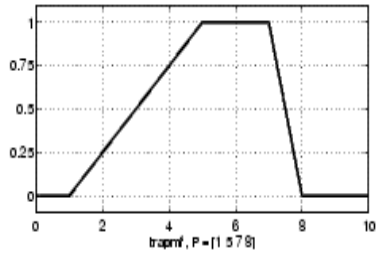


Fig. 3. Trapezoidal Membership Function Curve.

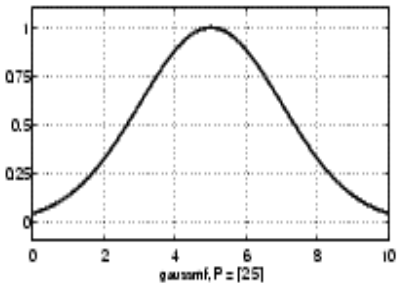


Fig. 4. Gaussian Distribution Curve.

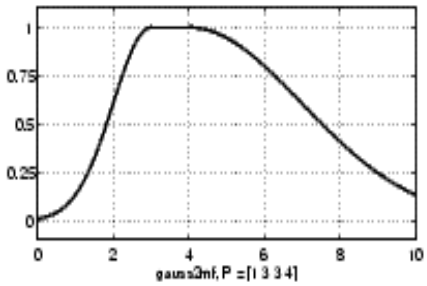


Fig. 5. Gauss2mf Function Curve.

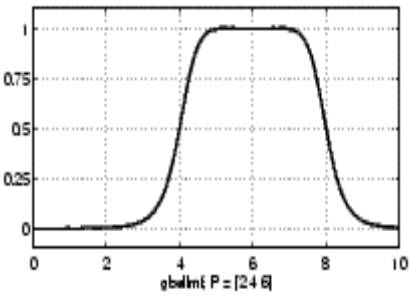


Fig. 6. Generalized Bell Function Curve.

IV. Proposed Model

In this article, four different context aware m-learning scenarios under different circumstances have been modeled. The context relative information has been collected in real time where 700 engineering students participated. Engineering students of branches Computer Science, Information Science, Electrical, Electronic and Mechanical have participated and data is collected through Google forms.

Students have answered various questionnaires related to their contexts, device characteristics and format of the contents that they are interested. The questionnaires are related to the following questions:

- Preferences of and within locations – home, college, etc.
- Personal factors like - friends, likes to be alone, in a group.
- Whether a student likes audio, video, text/pdf in learning.
- What time student prefers to study for the time of the day- daytime, morning, afternoon, evening, night, midnight?

The content delivery is adaptable with preferred contexts, entered by the students. Fig. 7 depicts the general architecture of the adaptive content delivery system.

The four scenarios deal with different contexts. All the scenarios generally have the domains: Preferred Location, Preferred Timing, Preferred Content Type, Noise, Network Speed and Battery Level. Each domain in turn has attributes that takes different values.

A. *Scenario I*

In this model, data pertaining to home context is collected, as shown in Fig. 8.

B. *Scenario II*

Here, data pertaining to home context is collected with addition to noise, as shown in Fig. 9.

C. *Scenario III*

Here, data with respect to the college context is collected, as shown in Fig. 10.

D. *Scenario IV*

Here, data is collected considering both contexts home and college, as shown in Fig. 11.

ANFIS is used to analyze all the four scenarios separately. Firstly, the data collected from the students, is partitioned into training dataset and checking dataset and further divided in four groups (training dataset %: checking dataset %) as 50:50, 60:40, 70:30 and 80:20, respectively. Secondly the behavior of these groups is measured with the five membership function: trimf, trapmf, gaussmf, gauss2mf and gbellmf. The grid partitioning algorithm is used as the number of input variables are in the range of 4 to 5. Further with respect to each group and with respect each membership function, the RMSEs are generated. Lastly the average RMSE is calculated for each membership function generated with respect to all the four groups of datasets. And the average RMSEs are compared and the membership function with the least RMSE is considered as the better performing result.

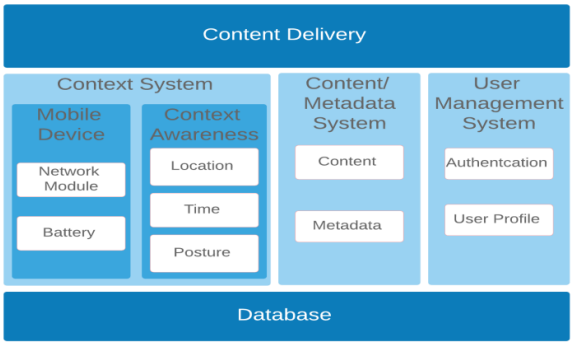


Fig. 7. Working Model of Adaptive Content Delivery Context System.

Scenario I : HOME	
Location:	Room,Hall,Gallery
Timing:	Morning,Afternoon,Evening,Night
Network:	2g,3g,4g
Battery:	Low,Average,High
Content Format:	Audio,Text/pdf,Video

Fig. 8. Home Scenario I.

Scenario II : HOME(NOISE)	
Location:	Room,Hall,Gallery
Timing:	Morning,Afternoon,Evening,Night
Network:	2g,3g,4g
Battery:	Low,Average,High
Noise:	Alone,Friends
Content Format:	Audio,Text/pdf,Video

Fig. 9. Home Scenario II.

Scenario III : COLLEGE	
Location:	Classroom,Library,Campus
Timing:	Morning,Afternoon,Evening,Night
Network:	wifi,4g
Battery:	Average,High
Noise:	Alone,Friends
Content Format:	Audio,Text/pdf,Video

Fig. 10. College Scenario III.

Scenario IV : HOME & COLLEGE	
Location:HOME	Room,Hall,Gallery
Location:COLLEGE	Classroom,Library,Campus
Timing:	Morning,Afternoon,Evening,Night
Network:	2g,3g,4g,wifi
Battery:	Low,Average,High
Noise:	Alone,Friends
Content Format:	Audio,Text/pdf,Video

Fig. 11. College and Home Scenario IV.

V. Results and Discussion

In this section, four different scenarios have been simulated using ANFIS.

A. Scenario I

In this approach the context of home is considered. The student learners have given the various preferences to study. The study material i.e. content will be delivered in text/pdf, or audio or video format in association with the context preferred and device characteristics at that moment. The attributes and their values are given as:

Location (Home) ={ room,hall,gallery}

Timing = { morning,afternoon,evening,night}

Network={2g,3g,4g}

Battery={low,average,high}

Content={ text/pdf, audio, video}

The suggestion rules are constructed based on the inputs given by the student learners; are in the format of IF-AND-THEN rules, for example:

IF Location(Home)=’room’ AND Timing=’morning’ AND Network=’3g’ AND Battery=’low’ THEN Content=’audio’

IF Location(Home)=’room’ AND Timing=’night’ AND Network=’4g’ AND Battery=’average’ THEN Content=’video’

IF Location(Home)=’hall’ AND Timing=’afternoon’ AND Network=’3g’ AND Battery=’high’ THEN Content=’video’

IF Location(Home)=’gallery’ AND Timing=’evening’ AND Network=’4g’ AND Battery=’average’ THEN Content=’text/pdf’

Fig. 12 depicts the ANFIS structure with 3 * 4* 3* 3 input membership functions.

Table I shows the performance behavior of the scenario I where RMSEs of the trimf,trapmf,gbellmf,gaussmf and gauss2mf are generated and compared. Out of five mf types, trimf gives 0.87 and gbellmf gives 0.90 with better performance. And remaining trapmf, gaussmf and gauss2mf give >1, which are with poor performance. But overall trimf with RMSE=0.87 performs well in comparison with others as depicted in Fig. 13.

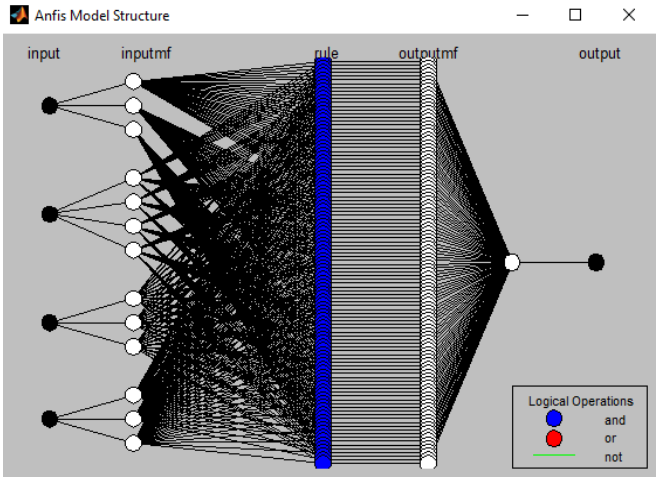


Fig. 12. ANFIS Structure with 3 * 4* 3* 3 Input mfs Scenario 1.

TABLE. I. Performance Behavior of Various MF Types for Scenario I

Number of fuzzy rules				**108 (3*4*3*3)**									
Total Number of Data Set				**408**									
Membership Functions				**trimf**		**Trapmf**		**gbellmf**		**Gaussmf**		**gauss2mf**	
Training set (%): checking set(%)	Number of training data pairs	Number of checking data pairs	Training RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	Epoch	Checking RMSE	epoch	Checking RMSE
50:50	204	204	0.788	5	0.80	10	1.261	15	1.226	20	1.39	20	1.39
60:40	245	163	0.794	5	0.920	10	1.543	15	0.788	20	1.039	20	1.085
70:30	286	122	0.791	5	0.995	10	0.952	15	0.824	20	1.091	20	1.063
80:20	326	82	0.802	5	0.787	10	1.175	15	0.781	20	1.0383	20	1.155
Average			**0.793**		**0.875**		**1.232**		**0.904**		**1.13**		**1.17**

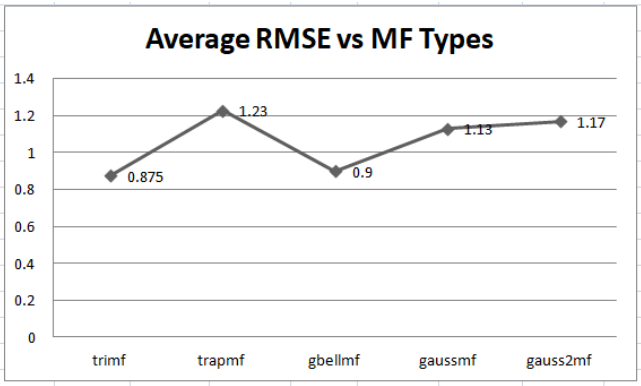


Fig. 13. Average RMSEs of Scenario I against Various MF Types.

B. Scenario II

In this approach the context of home with noise is considered. The student learners have given the preferences to study with additional attribute noise; whether learner wants to study alone or with friends. When with friends, generally it can be considered as discussion i.e. noisy environment. For the network attribute, additional 'wifi' value has been also added in this scenario. The study material i.e. content will be delivered in text/pdf, or audio or video format in association with the context and noise preferred and device characteristics at that moment. The attributes and their values are given as below:

Location (Home)={ room,hall,gallery}

Timing = { morning,afternoon,evening,night}

Noise={alone,friends}

Network={2g,3g,4g,wifi}

Battery={low,average,high}

Content={ text/pdf, audio, video}

The suggestion rules are constructed based on the inputs given by the student learners; are in the format of IF-AND-THEN rules, for example:

IF Location(Home)='room' AND Timing='morning' AND Noise='alone' AND Network='3g' AND Battery='average' THEN Content='text/pdf'

IF Location(Home)='hall' AND Timing='afternoon' AND Noise='friends' AND Network='4g' AND Battery='high' THEN Content='video'

IF Location(Home)='room' AND Timing='evening' AND Noise='alone' AND Network='4g' AND Battery='low' THEN Content='audio'

IF Location(Home)='gallery' AND Timing='evening' AND Noise='friends' AND Network='wifi' AND Battery='average' THEN Content='audio'

IF Location(Home)='room' AND Timing='night' AND Noise='friends' AND Network='2g' AND Battery='average' THEN Content='audio'

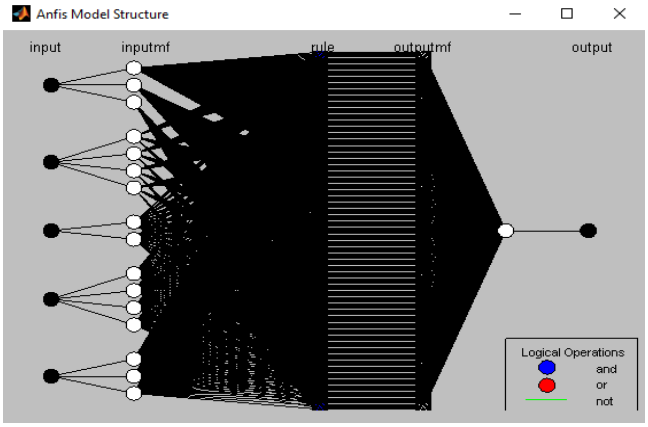


Fig. 14. ANFIS Structure with 3 * 4* 2* 4*3 Input mfs - scenario II.

Fig. 14 depicts the ANFIS structure with 3 * 4* 2*4* 3 input membership functions.

Table II summarizes the performances of the scenario 2 where RMSEs of the trimf,trapmf,gbellmf,gaussmf and gauss2mf are generated and compared. Out of five mf types, trimf gives 0.77, gbellmf gives 0.94 and gauss2mf gives 0. 85. And remaining trapmf and gaussmf give >1, performing poorly. But overall trimf with RMSE=0.77 performs well in comparison with others as depicted in Fig. 15.

C. Scenario III

In this approach the context of college is considered. The student learners have given the preferences to study in college. The study material i.e. content will be delivered in text/pdf, or audio or video format in association with the context and device characteristics at that moment. The attributes and their values are given as below:

Location (College)={ classroom,library,campus}

Timing = { morning,afternoon,evening,night}

Network={wifi,4g}

Battery={average,high}

Noise={alone, friends}

Content={ text/pdf, audio, video}

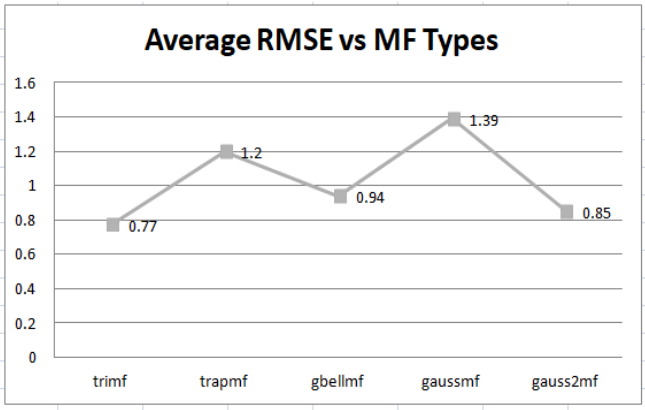


Fig. 15. Average RMSEs of Scenario II against Various MF Types.

TABLE. II. PERFORMANCE BEHAVIOR OF VARIOUS MF TYPES FOR SCENARIO II

Number of fuzzy rules				**288 (3*4*2*4*3)**									
Total Number of Data Set				**510**									
MF Types				**trimf**		**trapmf**		**gbellmf**		**gaussmf**		**gauss2mf**	
traning set(%): checking set(%)	Number of training data pairs	Number of checking data pairs	Training RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE
50:50	255	255	0.68	5	0.867	10	1.588	15	1.022	20	1.408	20	1.129
60:40	306	204	0.686	5	0.844	10	1.716	15	0.694	20	0.855	20	0.695
70:30	357	153	0.679	5	0.67	10	0.814	15	0.977	20	1.549	20	0.689
80:20	408	102	0.669	5	0.703	10	0.693	15	1.105	20	1.755	20	0.922
Average			**0.678**		**0.771**		**1.202**		**0.949**		**1.391**		**0.858**

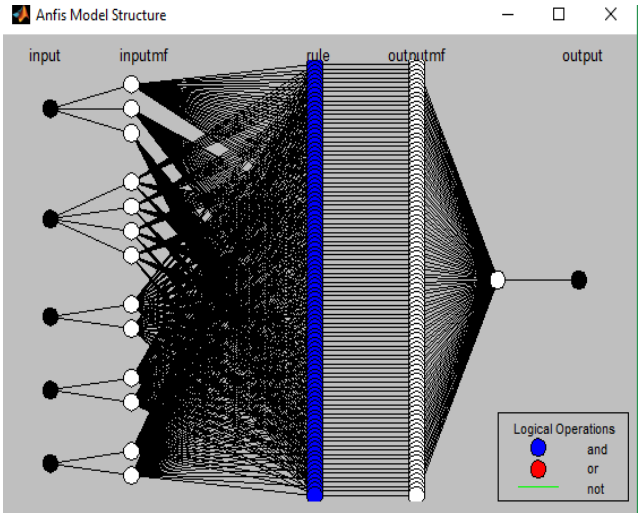


Fig. 16. ANFIS Structure with 3 * 4* 2* 2*2 Input MFS - Scenario III.

Fig. 16 depicts the ANFIS structure with 3 * 4* 2*2* 2 input membership functions.

Table III provides the performances of the scenario III where RMSEs of the trimf,trapmf,gbellmf,gaussmf and gauss2mf are generated and compared. Out of five mf types trapmf with RMSE=0.96 performs better. And remaining mf types give >1, performing poorly. Fig. 17 shows the comparison of all mfs.

D. Scenario IV

In this scenario the both the contexts- home and college are considered. The student learners have given the preferences to study in home and as well at college. The attributes and all corresponding values of Scenarios I, II, and III are taken into consideration. The study material i.e. content will be delivered in text/pdf, or audio or video format in association with the contexts and device characteristics at that moment. The attributes and their values are given as below:

Location (College)={ classroom,library,campus} (Home)={ room,hall,gallery}
Network={2g,3g,4g,wifi}

Timing = { morning,afternoon,evening,night}

Battery={low,average,high}

Noise={alone, friends}

Content={ text/pdf, audio, video}

Table IV provides the performances of the scenario 4 where RMSEs of the trimf,trapmf,gbellmf,gaussmf and gauss2mf are generated and compared. Out of five mf types, trapmf =0.78,gbellmf=0.80 and gaussmf=0.81 perform better with least RMSE values. And remaining mf types give >1, performing poorly. Trapmf with RMSE=0.78 performs much well, as Fig. 18 shows the comparison of all mfs.

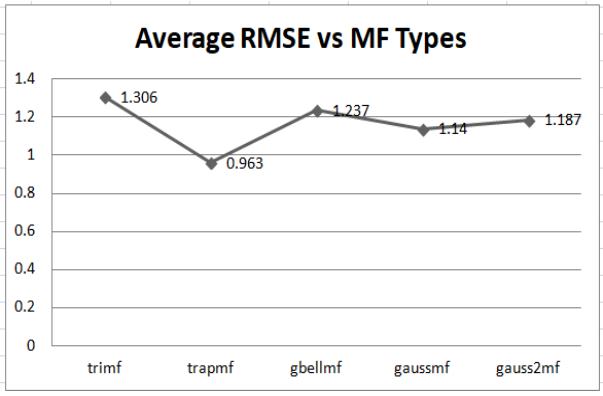


Fig. 17. Average RMSEs of Scenario III against Various MF Types.

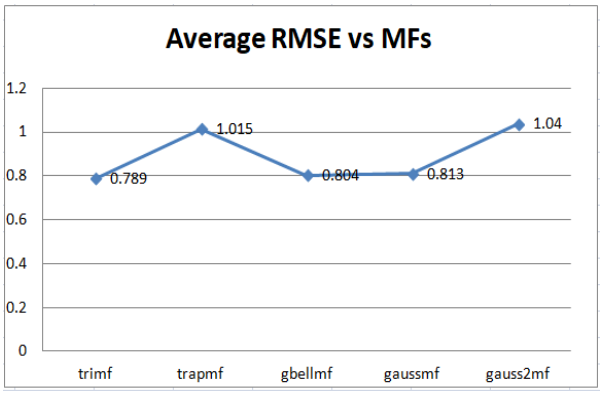


Fig. 18. Average RMSEs of Scenario IV against Various MF Types.

TABLE. III. PERFORMANCE BEHAVIOR OF VARIOUS MF TYPES FOR SCENARIO III

Number of fuzzy rules				**288 (3*4*2*2*2)**									
Total Number of Data Set				**240**									
MF Types				**trimf**		**trapmf**		**gbellmf**		**gaussmf**		**gauss2mf**	
Training set (%): checking set(%)	Number of training data pairs	Number of checking data pairs	Training RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epochs	Checking RMSE
50:50	120	120	0.801	5	1.211	10	1.319	15	1.233	20	1.256	20	1.318
60:40	144	96	0.80	5	1.63	10	0.831	15	0.887	20	0.902	20	0.9435
70:30	168	72	0.816	5	1.509	10	0.824	15	1.271	20	0.824	20	0.824
80:20	192	48	0.812	5	0.877	10	0.881	15	1.557	20	1.579	20	1.665
Average			**0.807**		**1.306**		**0.963**		**1.237**		**1.140**		**1.187**

TABLE. IV. PERFORMANCE BEHAVIOR OF VARIOUS MF TYPES FOR SCENARIO IV

Number of fuzzy rules				**576 (6*4*4*3*2)**									
Total Number of Data Set				**660**									
MF Types				**trimf**		**trapmf**		**gbellmf**		**gaussmf**		**gauss2mf**	
Training set (%): checking set(%)	Number of training data pairs	Number of checking data pairs	Training RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE	epoch	Checking RMSE
50:50	330	330	0.687	5	0.927	10	1.765	15	0.924	20	0.865	20	1.711
60:40	396	264	0.708	5	0.709	10	0.707	15	0.69	20	0.7	20	0.707
70:30	462	198	0.758	5	0.743	10	0.715	15	0.766	20	0.766	20	0.765
80:20	528	132	0.766	5	0.78	10	0.875	15	0.838	20	0.924	20	0.977
Average			**0.729**		**0.789**		**1.015**		**0.804**		**0.813**		**1.04**

TABLE. V. ALL SCENARIOS WITH AVERAGE TRAINING AND CHECKING RMSES

Scenarios	**Average Training RMSE**	**Average Checking RMSE**				
		trimf	**trapmf**	**gbellmf**	**gaussmf**	**gauss2mf**
Scenario 1	0.793	0.875	1.232	0.904	1.13	1.17
Scenario 2	0.678	0.771	1.202	0.949	1.391	0.858
Scenario 3	0.807	1.306	0.963	1.237	1.140	1.187
Scenario 4	0.729	0.789	1.015	0.804	0.813	1.04

Table V refers to all above scenarios and presents the average training RMSEs and Average Checking RMSEs of all mf types.

Observations:

1) For Scenario I: trimf membership function performs well.

2) For Scenario II: trimf membership function performs well.

3) For Scenario III: trapmf membership function performs well.

4) For Scenario IV: trimf membership function performs well.

5) Overall impression 1: Scenario III gives very poor performance compared to all others; suggestion rules built for this scenario poorly perform.

6) Overall impression 2: Scenario II performance is better in comparison with Scenario I.

7) Overall impression 3: Scenario IV performs best.

The rules constructed for Scenario IV, that considers the data of both contexts home and college (refer Fig. 12) are more precise giving good performance when compared to all other scenarios in providing the adaptive content delivery based on the context. The simulation of Scenario IV using ANFIS gives comparatively better RMSE when compared to all others.

VI. Conclusion

M-learning, learn while on move, is the revolutionized education concept with increased mobility and technology advancements. To deal with mobile learning, there is a need to understand the context of various learners. According to the context preferences- like place, timing, network speed, battery level etc., there is requirement to deliver the content. The content delivery should be adaptable based on the preferences given by the learners. In this paper, four such different m-learning scenarios have been identified. Finding out of the performance of these scenarios, and which one performs best is the work behind this paper. ANFIS i.e. artificial neuro fuzzy inference model is used to simulate all the four scenarios. Individual scenarios are compared within themselves using five membership types - *trimf, trapmf, gbellmf, gaussmf and gauss2mf.* And based on RMSEs which membership type performs well has been identified for each scenario. At the end, comparison of all the four scenarios has been made to find out which scenario performs best. And it has found that *trimf* (triangular membership function) performs mostly well in all four scenarios with least RMSEs. And Scenario III gives very poor performance compared to all others; Scenario II performance is better in comparison with Scenario I and Scenario IV performs the best.

References

[1] Yuan Jiugen , Xing Ruonan , Wang Jianmin : “Applying Research of Mobile Learning Mode in Teaching”, International Forum on Information Technology and Applications (IFITA), Volume: 3, Pages: 417 – 420, 2010.

[2] Yau, J.Y.-K. , Joy, M. “A Context-Aware Personalized M-learning Application Based on M-Learning Preferences”, The 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), Page(s): 11 - 18,2010.

[3] Yi Jin : “Research of One Mobile Learning System”, International Conference on Wireless Networks and Information Systems,(WNIS '09) , Pages: 162 - 165 , 2009.

[4] Fadi R.S.: “Data Mining in M-Learning Domain”, Trends in Innovative Computing-Information Retrieval and Data Mining, http://www. mirlabs.net/ict12/download/Paper11.pdf, 2012.

[5] Qin Shuai , Zhou Ming-Quan: "Cloud Computing Promotes the Progress of M-Learning”, International Conference on Uncertainty Reasoning and Knowledge Engineering (URKE), Volume: 2, Pages: 162 - 164, 2011.

[6] Yuan Jiugen , Xing Ruonan , Wang Jianmin : “Applying Research of Mobile Learning Mode in Teaching”, International Forum on Information Technology and Applications (IFITA), Volume: 3, Pages: 417 – 420, 2010.

[7] Yau, J.Y.-K., Joy, M.: “A Context-Aware Personalized M-learning Application Based on M-Learning Preferences”, The 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), Page(s): 11 - 18 , 2010.

[8] Yi Jin : “Research of One Mobile Learning System”, International Conference on Wireless Networks and Information Systems,(WNIS '09) , Pages: 162 - 165 , 2009.

[9] Hsuan-Pu Chang “Applying Adaptive Course Caching and Presentation Strategies in M-Learning Environment” , IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Pages: 1314 – 1318, 2010.

[10] Gwo-Jen Hwang , Wan-Ling Kuo , Po-Han Wu ,Yueh-Min Huang , Ya-Yan Zhuang: “An Investigation on Students' Cognitive Load and Learning Achievements for Participating in a Local Culture Mobile Learning Activity”, 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), Pages: 27 – 33, 2010.

[11] Fadi R.S.: “Data Mining in M-Learning Domain”, Trends in Innovative Computing-Information Retrieval and Data Mining, http://www. mirlabs.net/ict12/download/Paper11.pdf, 2012.

[12] Madhubala, R., and A. Akila. "Context Aware and Adaptive Mobile Learning: A Survey." Advances in Computational Sciences and Technology 10.5 (2017): 1355-1370.

[13] Tortorella, Richard AW, and Sabine Graf. "Considering learning styles and context-awareness for mobile adaptive learning." Education and Information Technologies 22.1 (2017): 297-315.

[14] Polatidis, Nikolaos, et al. "Privacy-preserving recommendations in context-aware mobile environments." Information & Computer Security 25.1 (2017): 62-79.

[15] Casillo, Mario, et al. "A Context-Aware Mobile Solution for Assisting Tourists in a Smart Environment", (2017).

[16] Singh, Gnana, et al. "Mobile Application for m-Learning, "International Journal of Advanced Research in Computer Science 8.3”, (2017).

[17] https://in.mathworks.com.

[18] https://edoras.sdsu.edu/doc/matlab/toolbox/fuzzy/.

Performance of LoRa Network for Environmental Monitoring System in Bidong Island Terengganu, Malaysia

Nur Aziemah Azmi Ali[1], Nurul Adilah Abdul Latiff[2], Idrus Salimi Ismail[3]
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Terengganu, Malaysia

***Abstract*—Recent wireless communication network which is Low Power Wide Area Network (LPWAN) bring a huge potential in monitoring system as the integration between sensors application and LPWAN technology itself contribute to greater efficiency, higher productivity, and better quality of life in multiform sectors such as in healthcare, smart city application, and monitoring system. The project involves the implementation of Low Power Wide Area Network (LPWAN) while delving into performances of the LoRa Technologies in the development of the environmental monitoring system in Bidong Island located in state of Terengganu, Malaysia. An Arduino Uno microcontroller stacked with LoRa module shield via SPI connection and in conjunction with few sensors work as end device to capture the environment data and send it to the gateway over the long range at a very low-data-rate with low power consumption for remote monitoring system. In this paper, we identify the best spreading factors that works well in the island region by manipulating all the spreading factor from SF6 to SF12. The result shows SF11 provides the best packet delivery ratio and the most stable Receive Signal Strength Indicator (RSSI) compared to other SFs. Moreover, we evaluate the external factors which caused the packet loss data and provide ways to improve the signal quality and to achieve optimal results for signal transmission and communication performance.**

***Keywords*—LPWAN; LoRa; environmental monitoring; WSN**

I. INTRODUCTION

Radio communication with wireless network becomes one of the most life-changing technology trends for these past few decades and it keeps growing with ubiquitous connectivity in every field as everything is interconnected. This interconnection of things refers to the idea of "Internet of Things" (IoT) in smart city applications, intelligent agriculture, smart manufacturing, health-care monitoring system and etc. [1]. Conventional cellular networks and Wi-Fi are the most common networks that offer high data rates and provide the needs of bandwidth-heavy applications such as video streaming and cloud computing to consumers. However, these networks cannot be operated without the main power and are difficult to be deployed in remote and underground areas due to high cost in installation and maintenance. Besides, several issues such as low coverage range, low robustness against network interference and jamming are parts of challenges identified that do not meet some application requirement [2][3][4]. Notwithstanding these limitations, researchers come with a new invention of Low Power Wide Area Network (LPWAN) triggered from Machine-to-Machine (M2M) communication idea to create a communication of information transmission without human intervention.

In this paper, a prototype of an environment monitoring system using LPWAN and LoRa technology was proposed. The environmental sensor had been used to measure the environmental parameters such as temperature and humidity. Then, LoRa performances were analysed in term of signal reliability such as Packet Delivery Ratio (PDR) and Receive Signal Strength Indicator (RSSI). The analysis was done using different spreading factors and different antenna's situation setups at both transmitter and receiver. This paper focuses on the Low Power Wide Area Network which is suitable for a monitoring system in a remote area that does not require high data rates and has minimal power consumptions. Table I presents the comparison between the existing wireless network and LPWAN itself according to [5], [6]. From the tabulated data, it explains the reason why LPWAN was selected to be implemented in this environmental monitoring system.

In the next Section II, we describe the main characteristics of LPWAN and its technologies. Followed by section III; we discuss the related work and in Section IV we explain the methodology used in the development of environmental monitoring system. Then, the results conducted in the experiments are presented and discussed in Section V. Lastly, Section VI concludes this paper.

TABLE. I. THE CHARACTERISTICS OF EXISTING WIRELESS NETWORK

Wireless System	**Long Range, High Data Rate (LTE)**	**Long Range, Low Data Rate (LPWAN)**	**Short Range, High Data Rate (WI-FI)**	**Short Range, Low Data Rate (ZigBee)**
Coverage	Large	Large	Small	Small
Range	Long	Long	Short	Short
Latency	Low	High	Low	Low
Bandwidth	200kHz-900MHz	500kHz-900MHz	2.4, 3.6, 5, and 60GHz	<2.4GHz
Power Consumption	Low	Low	High	Low
Operating Cost	Expensive	Cheap	Expensive	Cheap
Topology	Star	Star and Mesh	Star	Point-to-point
Data Rate	High	Low	High	Low

II. LOW POWER WIDE AREA NETWORK (LPWAN)

LPWAN arises and complement the conventional networks with the capability to transmit a signal with minimal transmission power about 10mW [1]. The transmission of data to the gateway by end device is only few packets per day and it follows below 1% of duty cycle restrictions set by most country depending on the frequency band used. For instance, Malaysian Communications and Multimedia Commission (MCMC) has set a standard for technical regulatory aspect in a frequency band, maximum power transmit and duty cycle for radio communication devices [7] to be used in Malaysia. Duty cycle in radio communication allows end device to work only when data need to be transmitted or received and it will be turned off if not required. Moreover, LPWAN implies star topology to avoid multiple hops communication and network congestion towards the gateway. The end device transmits the data directly to the gateway resulting in minimal power usage as the end device does not require to listen to the other devices to pass the information. These mechanisms permit the battery to live longer more than 10 years of operation and make it possible to cut down the cost while using an AA and a coin cell batteries for some certain applications [8].

LPWAN offers wide range transmission up to 50km and able to cover large areas with a single base station while supporting multiple numbers of connected devices [9]. This is because LPWAN signal propagates in unlicensed sub-1 GHz radio frequency band and adopted unique modulation technique. In Malaysia, radio frequency band of 919 - 923MHz is used for LPWAN applications [1]. Sub-1 GHz network band is less congested and less occupied that make it safe from signal interference compared to 2.4 GHz ISM band that is used commonly for Wi-Fi, Cell-Phone peripherals, Microwave ovens, ZigBee, and Bluetooth. Therefore, this Sub-1 GHz band serve a better network propagations for signal performance and more resilient when passing through obstacles by creating more wave diffraction around it and experiences less signal attenuation that allows it to transmit for a longer range [10].

There are two types of modulation technique used by LPWAN technology at the physical layer such as Ultra Narrow Band (UNB) and Spread Spectrum (SS) technique. Sigfox Company [11] uses UNB modulation technique to transmit a very narrow bandwidth signal typically less than 1kHz. This is because it is modulated by passing two significance stages which are known as an abrupt phase shift modulation stage and UNB filtering stage [12]. These stages protocols produce a suitable bandwidth for small uplink traffic and narrowing bandwidth to increase the link budget as the signal power become concentrated [13]. On the other hand, the SS modulation technique adopted by Semtech in LoRa technology [14] spreads the signal energy by reduces the spectral power density into a wide frequency band while preserves the same power intensity. Besides, the signal is secure and clear from traffic and interference as the peak is indistinguishable from the noise signal [15][16].

There are several examples of LPWAN technologies such as LoRa, Sigfox, RPMA [17] and Weightless [18]. However, this project focused on LoRa from Semtech technology. The term LoRa is referred to as "Long-Range" wireless communication system proposed by LoRa Alliance and developed by Semtech for commercial usage. LoRa utilizes Chirp Spread Spectrum (CSS) that has been used in military and space communication to achieve long range and low power energy characteristic [19]. In the CSS scheme, the signal data frequency chipped decreases or increases over a specific time in the form of the chirp signal. The condition of upper-chirp occurs when the frequency changes from lowest to highest, while down-chirp occurs vice versa. This means that the chirp signals have constant amplitude with variable frequency. The technique reduces the design complexity of the receiver as the timing signal and frequency offset between the LoRa transmitter and receiver are equivalent to each other [14]. This situation provides immunity against Doppler Effect and contributes to high receiver sensitivity achieved down to -148 dBm [20]. Furthermore, the product between high bandwidth and time gives the ability to resist against radio signals band-in or band-out interference while the best selection of broadband chirps able to avoid multipath fading and raise the robustness level. Hence, increasing the maximum link budget up to 157 dB and lowering the energy consumption by reducing the transmit power [21].

III. RELATED WORKS

A. Environmental Monitoring System using Prior Wireless Networks

Wireless Sensor Network (WSN) has revolutionized the world in many ways. Most of them focus on implementing common wireless networks on their environmental monitoring system. In [22], the authors describe snails monitoring by deploying a ZigBee protocol in a base mote for agricultural plantations and detects the presence of snails by using capacitive sensor to collect the data and pass it to the end device through the router. The research used WSN simulator to design the system and deploy a field test to support the real-life condition. In [23], the authors discussed the issues in Southern Africa on rainfall monitoring system by utilizing the General Pocket Radio Service (GPRS) via cellular network. The data

are sent directly to a website server for data storage. Authors in[24] propose a new novel of low cost and low power system for smart irrigation specialised in leaf sensing using backscatter radio communication such as in Radio Frequency Identification (RFID) application with 2m of communication distance between the tag and the emitter/reader antennas. However, it is preferable if the project implements the IoT system to get easy access of the data. Meanwhile in [25], the authors analysed three different sensor networks for environmental monitoring system using IoT capabilities. One of them employed Wi-Fi communication technology based on User Datagram Protocol (UDP) and Hypertext Transfer Protocol (HTTP) system platforms and the other used power-harvesting on Bluetooth Smart beacon (BLE). This research compared the developed systems in terms of power consumption, network design flexibility, and complexity. The results showed that using WIFI networks to transfer the collected data consumes more energy compared to transfer using the BLE. In addition, data transmitted using WIFI are totally dependable on Internet connection quality where the reduction of signal quality leads to packets loss. On the other hand, to implement BLE in monitoring system, the forwarder device is required to send the collected data from sensors to the cloud.

B. Environmental Monitoring System using LPWAN

Moving on to the implementation of LPWAN in environmental monitoring system, a few recent studies used LoRa in their project such as in [26]. The authors proposed the idea for a water monitoring system based on IoT and LoRa. The system was set up in a farm area from 1km to 3km between cattleman's house to the barn. The star topology communication has been implemented in this project and it consists of several nodes that detect and sense the water level in trough and sensor hub is used to collect all the data from the nodes and then the data pass to the gateway. HopeRF RFM95 was used for communication module on both side. At node, the controller used was Atmel ATMega328 with a float switch GE-1307 in order to reach low power system for remote area. While at the sensor hub part, Raspberry Pi v.1 type B was used as a controller because it has SPI protocol to communicate with HopeRF RFM95. The float switch in this project represented the condition of water quantity in the trough. The result showed that the longer the distance, the higher the signal power loss during transmission. Meanwhile, the elevation and polarization between hubs and nodes devices gave impacts on transmission quality. Horizontal polarization is the best position for this water monitoring application. Besides, researchers in [27] presented a new approach on smart water grid management using LPWAN technology by Indian communities in Mori village, India. The project aimed to provide real-time information to the villagers and authorities on quality of water supply and implemented alert mechanism to control the water flow either to step up or step down the water locks and at the same time create awareness on flood management. Moreover, residents around the area were able to use optimal quality water for their home and irrigational purpose according to the needs. The project composed of several sensors such as Oxidation Reduction Potential (ORP), pH, salinity level, turbidity, temperature and flow water sensor. Then, the sensors were mounted to LoRa RF Module and Arduino Nano as major components for the entire system. The collected data is then broadcasted to the LoRa Gateway and stored in the cloud. In this pilot project, the LoRa modules transmitted the data to the gateway via single-hop communication with MAC protocol has its own node address while the *App_Key* specifically identified which sender that the data comes from. The data that have been transmitted from a gateway and forwarded to the Ericsson cloud were then stored in the cloud according to device ID and next it went through an analytical process that can trigger alert that can be sent through SMS or emails. Lastly, residents were able to reach real-time updates from the dedicated website. To conclude, the project explores the potential of LoRa and at the same time solve the problem such as clean water supplies. Additionally, the authors in [28] deployed an environmental monitoring system with wireless self-powered based on LoRa technology. The system was consist of peripheral nodes or known as sensor nodes such as smoke detection (MQ2), alcohol (MQ3), carbon monoxide (MQ7), luminance (TEMT6000), methane(MQ4), humidity and temperature (HDC1080) sensors. The ATMega328 was used as a controller and LoRa RN2483 as a communication module. Besides detecting and acquiring information, peripheral nodes also transmitted the input to the LoRa gateway via LoRaWAN protocol. The gateway then processed the collected information and upload it to the cloud system via GPRS. This information was accessible to everyone through website or mobile application via ThingSpeak.com. The built prototype was also self-powered by solar energy and it was shown that the system was reliable as there was no packet loss detected.

IV. METHODOLOGY

The installation of monitoring system involves hardware and software components. We conducted several experiment to identify the best practical setup that can enhance the quality of LoRa signal and increase the packet delivery ratio. The measurements were carried out in Bidong Island (5.6186°, 103.0584°) with rough terrain topography and only one square kilometer in an area across the South China Sea. The transmitter node was placed 50m above the sea water aligned with the receiver node 75m from the sea level on a tophill of Bidong Island. Both LoRa nodes are in a slant path direction to minimize the interruption from obstacles that come rough topography and from other devices that might be use the same channel with our monitoring system. The distance between transmitter and receiver was around 1km and in near line-of-sight radio link as presented in Fig. 1.

Fig. 1. Location of Experiment in Bidong Island.

A. System Architecture of Monitoring System

The system architecture of the developed environmental monitoring system can be simplified into two parts which are LoRa sensor node and LoRa gateway as shown in Fig. 2. LoRa sensor node consists of sensors that functions to collect specific environment data and to define how many time the data must be sent to the gateway. While the gateway will always programmed to receive the data and it became a platform to store the data and allow to remotely monitor and configure the system in the very moment the data has been sent. These platform also enable us to track data loss rate.

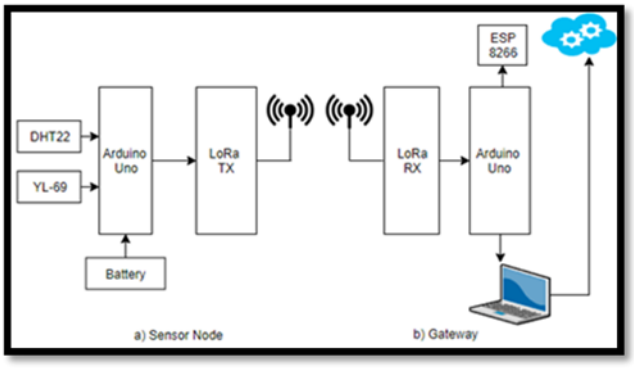


Fig. 2. System Architecture of the Monitoring System.

B. Hardware Component

The components used in this project composed of sensors, microcontrollers, radio modules, and two types of the antenna as shown in Table II.

C. Wiring Connection for the Transmitter and Receiver Node

The wiring connection for the sensor node and the gateway can be seen from the Fig. 3 and Fig. 4,– respectively.

The measurement of environmental parameters equipped with sensors such as temperature and humidity sensor (DHT22) and soil moisture sensor (YL-69) embedded with microcontroller for data acquisition and stack up with LoRa radio module connected all together to form a sensor node before transmit the data to the gateway. Both sensor node and gateway must be connected with correct choice of antenna types with optimal gain value to achieve a certain communication distance either using omnidirectional antenna or Yagi-Uda antenna.

D. Software Setup

The research was carried out using the Arduino Integrated Development Environment (IDE) software to write and program the system. The sensor node operated at 5V sampled the data from environmental sensors and transmitted to the gateway. The gateway received the data and only the successful packet data transmission displayed at the personal computer and store to a data logger. If the packet was lost, the sensor node will not retransmit the data lost to the gateway instead it skipped the data packet for the next data sensing and transmission. Both sensor node and gateway were set up using the same value of LoRa physical setting to ensure data received via the same channel.

1) LoRa physical setting: LoRa modulation scheme depends on five main configurable setting which are Transmission Power (TP), Spreading Factor (SF), Coding Rate (CR), Bandwidth (BW), and Carrier Frequency (CF). These parameters can be tuned into certain value to meet an optimal condition hence give the best data signal transmission and communication performance.

a) Transmission Power (TP): The power of the propagation signal to transmit a packet in LoRa transceiver and this parameter affects the power expenditure. It can be fine-tuned from 2dBm to 20dBm in 1 dB steps depending on hardware limitation while power usage higher than +17dBm limits for 1% radio duty cycle [29]. Increase the TP will increase the transmission range and data rate.

b) Spreading Factor (SF): The ratio of symbol rate and chip rate. The chip rate is a multiplication of bit rate and 2^{SF} [30] while varying the SF value from SF7 to SF12 can avoid concurrent transmission and does not cause packet collisions because every spreading factors are orthogonal to each other. Higher SFs provide long range but in trade-off low data rate.

c) Bandwidth (BW): The width of a selected frequency used in a transmission band. The transmission of a data sent as a chip rate per second is equal to the bandwidth. LoRa network can be operated in bandwidth either BW125, BW250, and BW500 correspond to 125kHz, 250kHz, and 500kHz respectively. This project used a lower bandwidth as it offer long range transmisson and higher sensitivity.

d) Coding Rate: The factor that increases the protection against corrupted bits in a radio link and known as Forward Error Correction (FEC) rate that allows bits information's recovery. A larger CR is more reliable and offers more protection but consumes more energy and increase the time-on-air. CR can be select either 4/5, 4/6, 4/7 or 4/8. LoRa radio always used CR 4/8 for the payload stored in the header packet and if being used as an explicit header the radios with different CR are still able to communicate with each other.

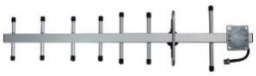
e) Carrier Frequency: The central frequency band that transmit the signal and depends on the allowed ranges of local regional frequency. For instance in Malaysia, Malaysian Communication Multimedia Commission (MCMC) has set a specific frequencies band for radio communication devices to be operated only between frequencies of 919MHz to 923MHz. While, in Europe frequency band available were 433MHz and 868MHz.

The whole experiment was conducted based on LoRa Physical setting as tabulated in Table III.

E. Experimental Setup

An antenna is a conductor that converts high-frequency alternating current (AC) signals to radio frequency signals and radiates into the air to form radio waves. In this experiment, two types of antennas were used to be connected with the LoRa module which 5dBi gain omnidirectional antenna was shown as in Fig. 5 and 11dBi gain the Yagi-Uda antenna as shown in Fig. 6 while varying the situation of the transmitter and the receiver at both locations as in Table IV.

TABLE. II. LIST OF COMPONENT USED IN THIS EXPERIMENT

Component	Explanation
DHT22 Temperature Sensor Module 	This 3-pins small size sensor measures the relative humidity and temperature values of its surrounding. It consumes only 0.3mA when operating and 60µA when on standby. Able to withstand severe application types with great reliability and stability. It has an accuracy of ±1%RH for humidity with range 0% to 100% and for the temperature range of -40°C to 80°C with ±0.5°C of accuracy.
YL-69 with LM393 	This simple sensor was used to measure the soil moisture and it comes with two parts which are the probes and the electronic module. Operates at voltage 3.3 to 5V and able to tuning the sensitivity via digital potentiometer. When the moisture in the soil is high, the output voltage decreases, while if the moisture is low, the output voltage increases.
Arduino Uno 	ATmega328p was selected as a microcontroller module on the Arduino Uno to control the sensors and it was embedded with communication module since it has the capability to convert 10-bit analog-to-digital (ADC) and has high-performance feature such as 16 clock speed and consists of 14 digital I/O pins that allows various type of application with low cost and low power consumption.
RFM95W 	LoRa module featured with SX1278 transceivers as a communication module and able to operate well with the microcontroller via Serial Peripheral Interface bus (SPI). LoRa can reach long distance communication and the data transmission is robust through their chirp spread spectrum and this module operates in the 915MHz frequency band and has -146 dBm receiver sensitivity.
Omni-Directional Antenna 	This high performance of 5dBi omnidirectional antennas operates in ranges between 868MHz to 915MHz of ISM band is ideal for multipoint and non-line-of-sight radio link. Connected with SMA Male copper connector and works in vertical polarization.
Yagi-Uda Antenna 	A powerful 11dBi gain Yagi-Uda directional antenna able to create strong focusing signal in one direction for maximum signal strength and performance over the long distances.

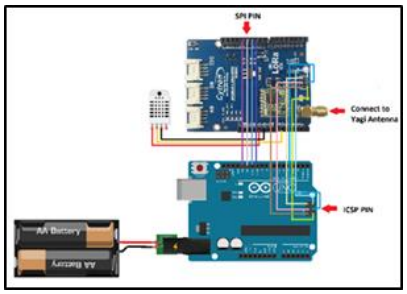


Fig. 3. Sensor Node act as LoRa Transmitter.

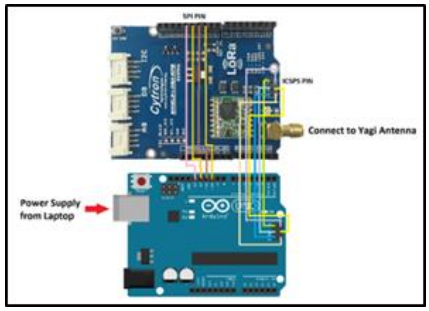


Fig. 4. Gateway Act as LoRa Receiver.

TABLE. III. LORA PHYSICAL SETTING USED IN THIS EXPERIMENT

TP	SF	BW	CR	CF	Payload
23dBm	7-12	125kHz	4/8	919MHz	5 Bytes

Fig. 5. The Sensor Node Connected with Omnidirectional Antenna.

Fig. 6. The Sensor Node Connected with Yagi-Uda Antenna.

TABLE. IV. ANTENNA TYPE AND SET UP FOR LORA NODES

Types of Antenna	Transmitter	Receiver
Situation A	Omni-directional	Omni-directional
Situation B	Yagi-Uda	Yagi-Uda
Situation C	Yagi-Uda	Omni-directional
Situation D	Omni-directional	Yagi-Uda

V. RESULT AND DISCUSSION

A. Signal Quality Performance

The device prototype has been developed and the results define the quality of signal transmission over the LoRa physical layer by using different spreading factors in different situations of antennas. The spreading factor was varied from SF7 to S12 and has been set up according to the situations of antenna mentioned in Table IV of this experiment.

From the Fig. 7, SF11 is noticeable to give better reliability among other SFs and works well in every situation of the antenna including in situation A where the omnidirectional antenna was used for both transceiver nodes and also in situation D, where the omnidirectional used as a transmitter and received by Yagi-Uda antenna. The value of RSSI is slightly affected by varying the spreading factor but mostly affected by the changes in the antenna type and antenna gain as well as transmission power. The whole experiment was decided to use 23dBm of transmission power (TP) because lower TP cannot transmit the data packet in this type of non-line of sight (NLOS) radio link. This result proved that to increase the signal quality, it is crucial to set up the transmitter and receiver in the most suitable situation with suitable LoRa PHY layer and also by considering the surrounding obstacles.

The types of the antenna are also an important factor to transmit or receive data. In this experiment, we observed that Yagi–Uda antennas provide more reliable data compare to omnidirectional antenna because it transmits and receives signals directly from a specific direction with a concentrated and strong signal beam. Meanwhile, the omnidirectional antenna radiates or receives the signal uniformly from all directions in 360 degrees horizontally with equal signal power. For that reason, the result obtained in this experiment might be slightly affected by the unmatched polarization between the transmitter and receiver position. The orientation of the antenna's nodes that we fixed depends on our smartphone's compass application and the calibration can be slightly inaccurate due to metal and other electronic interfered.

Next, Fig. 8 presents the graph of Signal-to-Noise Ratio (SNR) while Fig. 9 presents the graph of the Packet Delivery Ratio (PDR) and both are over the Spreading Factor (SF). The results show it is possible to increase the rate of PDR and the range of communication by using the higher SF (e.g.; SF12). However, it seems that increased SF caused more packets collision due to longer transmission time and enhanced the Signal-to-Noise Ratio (SNR) and increased the data lost. While it is also possible to use lower SF (e.g.; SF7) but, we must consider the link budget and the type the antenna for optimum signal transmission in a particular range.

B. External Factors that affect Signal Performance

The results obtained also correlated with Free Space Path Loss (FSPL) factor between two antennas. FSPL is a power loss at a rate inversely proportional to the distance travelled and proportionally to the wavelength of the signal illustrated as in Fig. 10.

$$FSPL = 20\ log_{10}\left[\frac{4\pi d}{\lambda}\right]\{dB\} \quad (1)$$

The calculation of path loss theory for every situation is recorded as in Table V below by referring to the equation (1). The value of FSPL for both situations in C and D is the same, but the result indicated that the antenna setup in situation C is more reliable compared to the situation in D because as mentioned in the previous section Yagi-Uda antenna that acts as transmitter radiates a strong signal and gives ability to directs the signal in a long distances and it is easier to the receiver to catch that signal compared to situation D where omnidirectional acts as a transmitter. The signal power from the omnidirectional antenna is not strong enough to be caught by the receiver.

Another important factor that contributes to this phenomenon is the Fresnel zone between transmitter and receiver. The Fresnel zone is a concentric ellipsoid–shaped area around the LOS path as shown in Fig. 11 and the radius of the Fresnel zone in this experiment is 9.03 meters calculated as in equation (2). The equation refers to the radius of the zone at its widest point where the signals propagate after leaving the antenna. Almost 40% of the Fresnel zone at the receiver location is blocked by the obstacle such as growing trees, old water tank, branches, plant canopy, and shrubs nearby the hilltop of Bidong Island. This obstruction affects the signal propagation by inducing the multi-path issues as the signal will be reflected, refracted, and diffracted away from the receiver and causing signal degradation and loss.

$$Radius\ (m) = 17.31 \times \sqrt{\frac{Distance\ (km)}{4f\ (GHz)}} \quad (2)$$

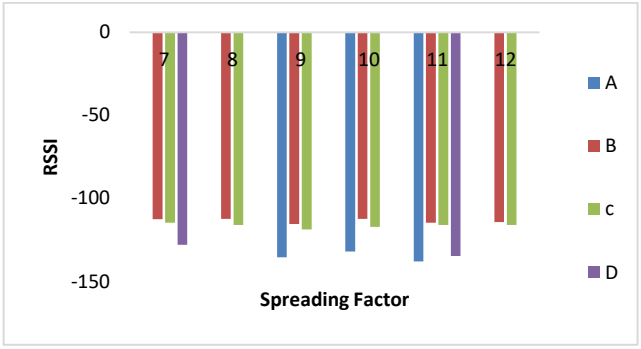


Fig. 7. Graph of RSSI Varies with different SF and different Antenna Situation.

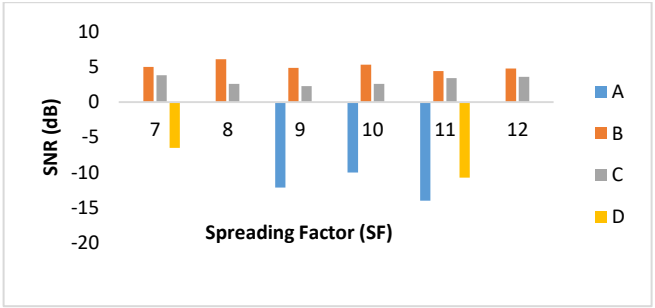


Fig. 8. Graph of SNR Varies with different SF and different Antenna Situation.

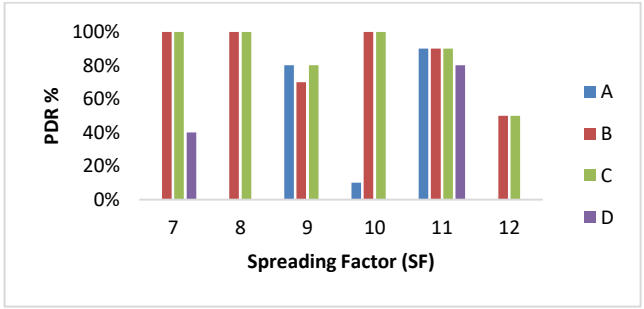


Fig. 9. Graph of PDR Varies with different SF and different Antenna Situation.

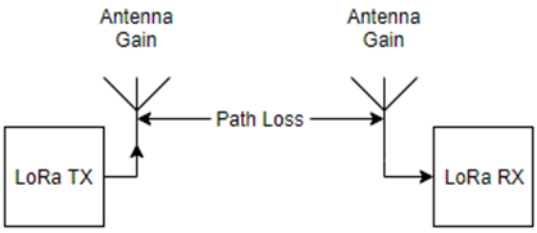


Fig. 10. Block Diagram of Free Space Path Loss.

TABLE. V. CALCULATION OF FSPL

Situation	FSPL (dB)
A	81.71
B	69.71
C	75.71
D	75.71

Fig. 11. The Fresnel Zone of Experimental Location.

VI. CONCLUSION

The paper has presented the hardware and software structures of the monitoring system and the details of each component used. The project also highlighted the implication of antenna settings as its effect on the signal transmission and explained the significance of selection in LoRa physical layer settings. The results of this research support the idea that Yagi-Uda directional antenna is the best choice for long-distance signal propagation that radiates strong beams transmission compared to an omnidirectional antenna. Meanwhile, dealing with the right antennas gains and polarization can lower the data rate loss and increase the signal quality. Besides, increase the antenna height for both transmitter and receiver also improves the signal propagation because it can reduce the ground refraction. Another important practical implication is to clear the Fresnel zone from any signal interferer and blockage that leads to signal loss such as growing trees, shrubs, and branches. Besides, one of the findings of the project was the value of the spreading factor that can be used in most of the situations is SF11 because of its variability. This is proven that with a suitable selection of SF can greatly impact the result of LoRa performances and improves the maximum range of coverage. Lastly, future research might explore the cloud server for the IoT system to be applied using ThingSpeak or any other open clouds platform for online processing data storage and analytics system.

ACKNOWLEDGMENT

The authors would like to be obliged to Luimewah (M) Sdn Bhd because of the project assistance throughout this project progress and this research was supported by University Malaysia Terengganu (UMT) and Ministry of Education Malaysia with financial assistance under project no. 59466.

REFERENCES

[1] MCMC, Iot-Technical-Regulatory-Aspects-Key-Challenges-180406014 354.Pdf..

[2] A. Lavric and V. Popa, “A LoRaWAN : Long Range Wide Area Networks Study,” pp. 417–420, 2017.

[3] X. Vilajosana, F. Adelantado, X. Vilajosana, P. Tuset-peiro, B. Martinez, and J. Melia, “Understanding the limits of LoRaWAN Understanding the limits of LoRaWAN,” no. July, 2016.

[4] I. S. Ismail, “Performance analysis of data recovery via application layer for LPWAN,” 2019 IEEE 89th Veh. Technol. Conf., pp. 1–5, 2019.

[5] M. M. Alsulami and N. Akkari, “The role of 5G wireless networks in the internet-of- things (IoT),” 2018 1st Int. Conf. Comput. Appl. Inf. Secur., pp. 1–8, 2018.

[6] J. E. Shuda, A. J. Rix, and M. J. Booysen, “Towards Module-Level Performance and Health Monitoring of Solar PV Plants using LoRa Wireless Sensor Networks,” 2018 IEEE PES/IAS PowerAfrica, pp. 172–177, 2018.

[7] M. Communications and M. Commission, “CLASS ASSIGNMENT NO . 1 OF 2017,” no. 1, 2017.

[8] U. Raza, P. Kulkarni, and M. Sooriyabandara, “Low Power Wide Area Networks: An Overview,” IEEE Commun. Surv. Tutorials, vol. 19, no. 2, pp. 855–873, 2017.

[9] E. De Poorter et al., “Sub-GHz LPWAN Network Coexistence, Management,” 2017.

[10] S. Aust, “Analysis of the Performance Boundaries of Sub-1 GHz WLANs in the 920 MHz ISM-Band,” pp. 1–5.

[11] Sigfox, “Sigfox Technical Overview,” vol. 1, no. May, p. 26, 2017.

[12] X. Xiong, K. Zheng, R. Xu, W. Xiang, and P. Chatzimisios, “Low power wide area machine-to-machine networks: Key techniques and prototype,” IEEE Commun. Mag., vol. 53, no. 9, pp. 64–71, 2015.

[13] M. Anteur, V. Deslandes, N. Thomas, and A. L. Beylot, “Ultra narrow band technique for low power wide area communications,” 2015 IEEE Glob. Commun. Conf. GLOBECOM 2015, pp. 1–6, 2015.

[14] Semtech, “LoRaTMModulation Basics Semtech,” 2015.

[15] “Understanding Spread Spectrum for Communications - National Instruments.” [Online]. Available: http://www.ni.com/en-my/innovations/white-papers/06/understanding-spread-spectrum-for-communications.html. [Accessed: 28-Apr-2019].

[16] S. Kartakis, B. D. Choudhary, A. D. Gluhak, L. Lambrinos, and J. A. McCann, “Demystifying low-power wide-area communications for city IoT applications,” pp. 2–8, 2016.

[17] “Technology - Ingenu.” [Online]. Available: https://www.ingenu.com/technology/. [Accessed: 13-May-2019].

[18] “What is Weightless - Weightless.” [Online]. Available: http://www.weightless.org/about/what-is-weightless. [Accessed: 13-May-2019].

[19] LoRa Alliance, “A technical overview of LoRa ® and LoRaWAN TM LoRaWAN TM What is it?,” no. November, 2015.

[20] “Sx1276/77/78,” no. July, p. 2012, 2012.

[21] K. Mikhaylov, M. Pettissalo, and J. Peta, “Performance of a low-power wide-area network based on LoRa technology : Doppler robustness , scalability , and coverage,” vol. 13, no. 3, 2017.

[22] D. García-lesta, D. Cabello, E. Ferro, P. López, and V. M. Brea, “Wireless Sensor Network With Perpetual Motes for Terrestrial Snail Activity Monitoring,” vol. 17, no. 15, pp. 5008–5015, 2017.

[23] E. M. Mangundu, J. N. Mateus, G. L. Zodi, and J. Johson, “A Wireless Sensor Network for Rainfall Monitoring , using Cellular Network : A Case for Namibia,” pp. 240–244, 2017.

[24] S. N. Daskalakis, G. Goussetis, S. D. Assimonis, M. M. Tentzeris, and A. Georgiadis, "A uW Backscatter-Morse-Leaf Sensor for Low-Power Agricultural Wireless Sensor Networks," IEEE Sens. J., vol. 18, no. 19, pp. 7889–7898, 2018.

[25] G. Mois, S. Folea, and T. Sanislav, "Analysis of Three IoT-Based Wireless Sensors for Environmental Monitoring," IEEE Trans. Instrum. Meas., vol. 66, no. 8, pp. 2056–2064, 2017.

[26] Lukas, W. A. Tanumihardja, and E. Gunawan, "On the application of IoT: Monitoring of troughs water level using WSN," 2015 IEEE Conf. Wirel. Sensors, ICWiSE 2015, no. August 2015, pp. 58–62, 2016.

[27] M. Saravanan, A. Das, and V. Iyer, "Smart water grid management using LPWAN IoT technology," GIoTS 2017 - Glob. Internet Things Summit, Proc., 2017.

[28] K. Tzortzakis, K. Papafotis, and P. P. Sotiriadis, "Wireless Self Powered Environmental Monitoring System for Smart Cities based on LoRa," 2017 Panhellenic Conf. Electron. Telecommun., 2018.

[29] M. Bor and U. Roedig, "LoRa transmission parameter selection," Proc. - 2017 13th Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2017, vol. 2018-Janua, no. June, pp. 27–34, 2018.

[30] M. H. Habaebi, I. J. Chowdhury, M. R. Islam, and N. A. B. Zainal, "Effects of shadowing on LoRa LPWAN radio links," Int. J. Electr. Comput. Eng., vol. 7, no. 6, pp. 2970–2976, 2017.

Vulnerability Analysis of E-voting Application using Open Web Application Security Project (OWASP) Framework

Sunardi[1]
Department of Electrical Engineering
Universitas Ahmad Dahlan
Yogyakarta
Indonesia

Imam Riadi[2]
Department of Information System
Universitas Ahmad Dahlan
Yogyakarta
Indonesia

Pradana Ananda Raharja[3]
Master Program of Informatics
Universitas Ahmad Dahlan
Yogyakarta
Indonesia

***Abstract*—This paper reports on security concerns in the E-voting used for the election of village heads. Analysis of the system and server uses two different tools to determine the accuracy of scanning vulnerabilities based on the OWASP Framework. We reported that the results of the scanning using the ZAP tool got vulnerability information with the following risk level, one high level, three medium levels, and eleven low levels. The Arachni tool got vulnerability information with the following risk level, one high level, three medium levels, and two low levels. ZAP has a more complex vulnerability view than Arachni. Fatal findings on E-voting in this E-voting system is XSS, which impacts clients, which can be exploited by attackers to bypass security. Directory Traversal allows attackers to access directories and can execute commands outside of the web server's base directory. Cyber Hiscox Readiness report in 2018 in several European countries such as The United States, Britain, Germany, Spain, and the Netherlands, that the Attackers target through the most vulnerable security holes such as injection, Broken Authentication, Sensitive Data Exposure, XXE, Merged, Security Misconfiguration, XSS, Insecure Deserialization, Using Components with Known Vulnerabilities, Insufficient Logging, and Monitoring. The purpose of cyberattacks alone can threaten the stability of the country and disturb other factors. E-voting, as part of an electronic government system, needs to be audited in terms of security, which can cause the system to disrupt.**

***Keywords*—Vulnerability; e-voting; open web application security project framework; attacker**

I. INTRODUCTION

The development of information technology is used to replace previous habits that have not been computerizing. E-voting is a web-based technology that can be utilized by the government in carrying out election activities [1] [2] [3]. The implementation of E-voting[1] in government can have a severe impact if it is not ready to respond to cyber-attacks [4] [5] [6]. Fig. 1 shows cyberspace statistics that, on average, there are millions of data breaches, and that dangerous attacks continue to increase. On average, 73% of businesses are not ready to respond to cyber-attacks is by the 2018 Cyber Hiscox Readiness Report[2].

The United States, Britain, Germany, Spain, and the Netherlands found that most organizations were not prepared and would be severely affected by cyber-attacks [7].

OWASP TOP 10 2017[3] explains ten application security risks, as in Fig. 2. Zed Attack Proxy[4] (ZAP) from OWASP is one of the most popular free security scanning tools in the world and is actively managed by hundreds of international volunteers. ZAP can automatically scan for security vulnerabilities in web applications when they are developed and tested. ZAP is a reliable tool for experienced penetration testers to be used as automatic safety testing tools [8][9].

In addition to explaining application security risks, OWASP Top 10 is a Guide for developers and security teams to control weaknesses in web applications that are vulnerable to attack and to anticipate. These various vulnerabilities make it easy for intruders to embed malware, search for data, or completely take over the site [10].

Even though the web server is physically protected, web applications that run in the environment are not protected from attacks through computer networks. The attacks referred to according to OWASP Top 10-2017 among other things, Injection Weaknesses such as SQL injection[5], NoSQL[6], OS[7], and LDAP[8] are caused when fake data is sent to the server as part of the order [11].

Broken Authentication is an application function related to authentication and sessions misapplied so that an attacker can ignore passwords, tokens, and exploit weaknesses as other implementations to use the identity of other users [12].

[1] https://github.com/pradavanraharja/ta_dup

[2] https://www.hiscox.com/sites/default/files/content/2018-Hiscox-Cyber-Readiness-Report.pdf

[3] https://www.owasp.org/index.php/Main_Page

[4] https://www.owasp.org/index.php/OWASP_Zed_Attack _Proxy_Project

[5] https://en.wikipedia.org/wiki/SQL_injection

[6] https://en.wikipedia.org/wiki/NoSQL

[7] https://en.wikipedia.org/wiki/Operating_system

[8] https://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol

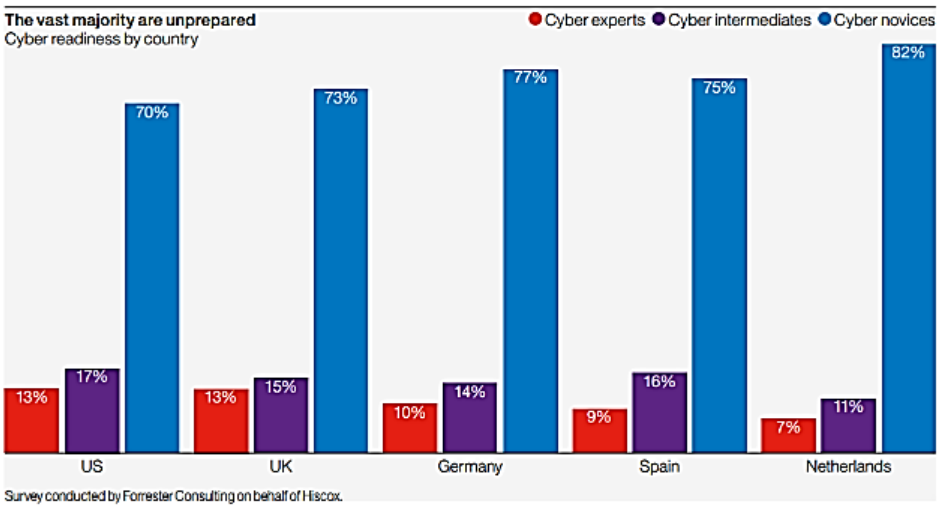


Fig. 1. The Organization is not Ready to Face Cyber-Attacks in 2018.

Sensitive Data Exposure is a matter related to web applications and the Application Programming Interface (API)[9]. It does not protect sensitive data accurately, such as financial, health, and personally identifiable information (PII). Rescue can be done, or change data is protected, but some are protected, so it is necessary to remove from credit cards, identity protection, and other crimes [13]. This sensitive data is not well protected because it is not encrypted when there is an exchange of data with the browser [14]. XML External Entities (XEE) are XML[10] documents that are outdated or not properly configured to be an attacker reference. Attackers can uncover internal files using URI Handler documents, internal file shares, internal port scanning, remote code execution and Denial of Service (DoS) attack.

OWASP Top 10 2017
A1:2017 – Injection
A2:2017 – Broken Authentication and Session Management
A3:2013 – Sensitive Data Exposure
A4:2017 – XML External Entity (XXE) [NEW]
A5:2017 – Broken Access Control [Merged]
A6:2017 – Security Misconfiguration
A7:2017 – Cross-Site Scripting (XSS)
A8:2017 – Insecure Deserialization [NEW, Community]
A9:2017 – Using Components with Known Vulnerabilities
A10:2017 – Insufficient Logging & Monitoring [NEW, Comm.]

Fig. 2. Top Ten Attacks by OWASP Top 10-2017.

Broken Access Control is a limitation of the actions that can be performed by authenticated users that are often misused. Attackers can exploit this weakness to access unauthorized functionality and data, such as accessing other users, viewing sensitive documents, modifying other users' data, and changing access rights [15]. Security Misconfiguration is the most commonly seen problem. Generally, it is the result of default configuration that is not safe, incomplete or ad hoc configuration, open cloud storage, configuration errors in the HTTP header, and error messages that contain sensitive information [16].

Cross-Site Scripting (XSS) can occur when an application is inserted fake data on a web page without validation or can update web pages with data entered through a browser, resulting in resulting HTML and Javascript. XSS is used by attackers to execute scripts in victim browsers that can hijack when accessed, delete the web, or redirect users to malicious sites [17]. Insecure Deserialization can cause remote code execution, even if insecure deserialization can be used to carry out attacks, including replay attacks, injection attacks, and privilege escalation attacks [18].

Using components with known vulnerabilities includes components, libraries, frameworks, and other software modules. If a component is vulnerable to exploitation, attacks like that have an impact on data loss or server takeover. The application and API use components with known vulnerabilities, thus damaging the application's defence and activating various attacks [19]. Insufficient logging and monitoring that is incomplete and ineffective with the incident response so that the attacker can carry out further attacks for maintaining position, expand attacks on the system, damage, extract, and destroy data. Research reports that approved for more than 200 days are usually known or detected from the private sector or monitoring [20].

ZAP is a tool that implements the OWASP Top 10 method with the main features of Intercepting Proxy, Automated Scanner, Passive Scanner, Brute Force Scanner, Spider, Fuzzer, Port Scanner, Dynamic SSL Certificate, API,

[9] https://en.wikipedia.org/wiki/Application_programming_interface
[10] https://en.wikipedia.org/wiki/XML

Beanshell Testing. As if an attacker from outside to break into the system to get data or make DoS Attack [21][22].

This research focuses on preventing attacks targeting web applications, servers, and artefacts related to vulnerability analysis in the E-voting tested concerning source code, libraries, folders, encryption, and web interfaces. The aim is to find out the possibility of an attack [23].

The graph in Fig. 3 shows that "exploitation of security misconfiguration" is 11%, and "Exploitation of software vulnerability" is 5%, including eight primary enemy targets [24]. The following can be a source of problems in security devices and is a challenge for practitioners to provide appropriate solutions and assistance to find out the source of problems, treatment, prevention, and repair, OWASP Top 10 is the best choice.

ZAP, as a testing tool designed explicitly for installing web applications, ZAP is known as a main-in-middle proxy. ZAP works between the browser tester and the web application so that it can intercept and read messages sent between the browser and the web application, manage the content needed, and can install packages to the destination we can see in Fig. 4 [25].

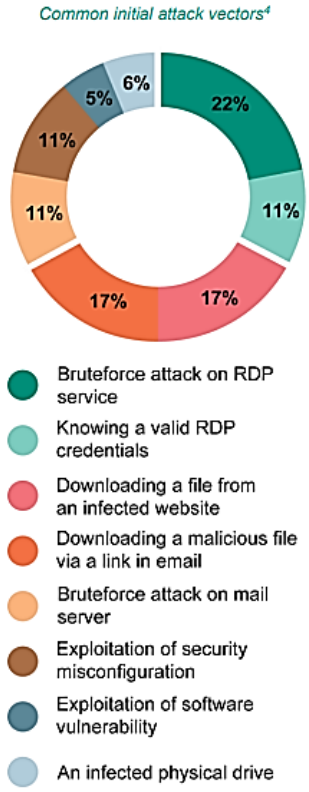


Fig. 3. Adversary Attack Vectors in 2018.

Fig. 4. The Concept of main-in-the-Middle Proxy on ZAP.

The functionality provided by ZAP for various levels ranging from Developers and Security Testing Specialists [26]. The results received by the tester are in the form of reports from ZAP in the form of HTML and XML files. In ZAP Scanning Report consists of Summary of Alerts, which are grouping in certain levels along with the amount.

II. Literature Review

This paper consists of two parts: Risk rating study based on OWASP on E-voting and comparison of reporting results from ZAP and Arachni.

A. Risk Rating

Currently, finding a vulnerability to be necessary, but being able to estimate risks associated with the business is equally important. One can identify security problems in architecture or design using threat modelling. Someone has the possibility of finding security problems using source code or penetration testing, and problems not yet found until the application is in production and entirely compromised.

This method might be used to estimate the severity of all risks to the web application and make decisions based on information about what to do with those risks. Reviewing the ranking of risks to the system will save time and eliminate polemics about priorities. The results of research on web applications are not affected by small risks, and more significant risks not known with certainty.

As shown in Table I, explained that in this step, the likelihood estimation and impact estimation are combined to calculate the overall severity of this risk. It is finding out if the likelihood is low, medium, or high and then doing the same for impact. Scale 0 to 9 into three parts. Ideally, there will be a universal risk assessment system that will accurately estimate all risks for all organizations. However, vulnerabilities that are very important for one organization may not be too significant for another. So the basic framework presented here is for specific organizations [27].

B. OWASP Top 10-2017

OWASP Top 10-2017 is an update of OWASP as an open-source community dedicated as an organization that develops, funding, maintains applications and APIs. The benefits of OWASP (1) Tools and application security standards, (2) Complete books on application security testing, (3) code of development, and safe code review (4) Presentations and videos, (5) Cheat sheets on many general topics , (6) Control and standard security libraries, (7) local chapters around the world, (8) Leading research, (9) Extensive conferences around the world, (10) Lists of correspondence [28].

C. Architecture of Networks

The network architecture applied in this research consists of a web server acting as an E-voting server and a client acting as a Vulnerability Scanner, and there is a ZAP Application. Both the server and client use the Windows 10 operating system. The network architecture in Fig. 5 is a network architecture that simulates vulnerability scanning activities on the e-voting server to obtain important information related to vulnerabilities on the server and web application.

TABLE. I. RISK LEVEL

Likelihood and Impact Levels	
0 to <3	Low
3 to <6	Medium
6 to 9	High

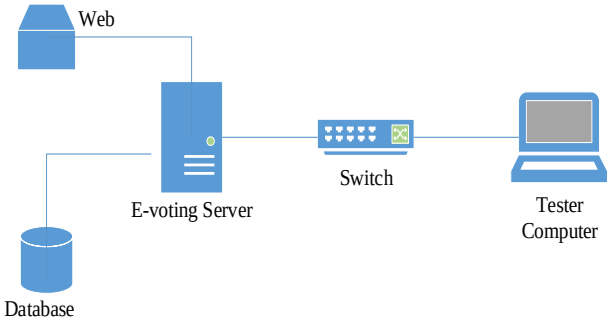


Fig. 5. NIST Methodology.

D. Methodology

The method used in this research is the forensic method based on the National Institute of Standards and Technology (NIST) with the forensic stages of acquisition, inspection, utilization, and review, as described in Fig. 6 [29][30][31].

NIST is the organization responsible for developing minimum standards, guidelines, and requirements to provide adequate information security for all assets and parties with digital forensic competencies [32].

1) Acquisition: The first step in the research process is to identify the source of the data, the phase of data acquisition related to a particular event that will be identified, collected, and protected. Table II shows the equipment and material requirements needed.

2) Examination: The data obtained in the next phase is to examine the data, identify, collect, and organize relevant information from the data obtained. This phase can also involve bypassing or mitigating operating system or application features that obscure data and code, such as data compression, encryption, and access control mechanisms. It is a testing phase of appropriate tools and techniques for the types of data collected during the first phase to identify and analyse relevant information from the data obtained.

3) Utilization: The utilization of data is the process of preparing and presenting information generated from the inspection stage. Many factors influence data utilization, including data reduction, alternative explanations, audience considerations, and actionable information. The final phase, which involves the reporting process and practice in the context of current events to identify policy deficiencies, then procedural errors, and other problems that need to be corrected.

To identify each computer in the network, in this research we provide 192.168.130.0 networks on two computers as shown in Table III.

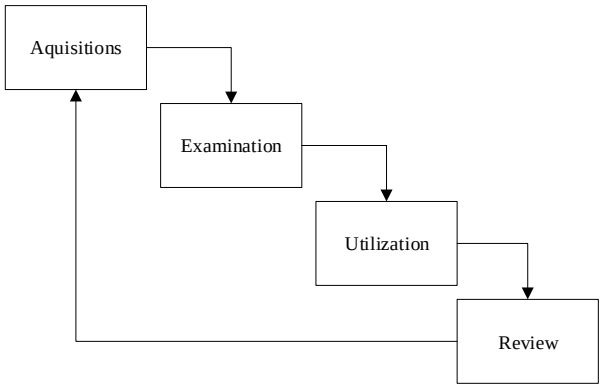


Fig. 6. NIST Methodology.

TABLE. II. EQUIPMENT FOR SUPPORT VULNERABILITY SCANNING PROCESS

No	Equipment	Description
1	E-voting Computer Server	Intel Pentium CPU, 4GB RAM, SSD 128GB
2	Computer Tester	Intel i5 CPU, 12 GB
3	Apache Web Server	Version 2.2
4	MySQL database	Version 10.1.16
5	Switch	Tp-link TL-SG1005D
6	ZAP	Version 2.8.0
7	Arachni Framework	Version 1.5.1
8	Arachni WebUI	Version 0.5.12

TABLE. III. SERVER, TESTER'S COMPUTER AND IP ADDRESS

No	Host	IP Address
1	E-voting Server (Windows 10)	192.168.130.248 (*static*)
2	Computer Tester (Windows 10)	192.168.130.90 (*dynamic*)

Fig. 7 explains the reporting process on vulnerability scanner tools [33]:

a) The tester runs penetration testing tools.

b) Design API report on Penetration testing tools.

c) APIs Report inserted in Penetration testing tools.

d) Penetration testing tools produce XML documents.

e) XML documents and Report APIs are combined to produce formatted reports.

4) Review: The analysis is repeated to improve the process and practice in the context of the current task to help with policy problems, procedural problems, and other problems that need fixing. Regular updating of skills through courses, workplace experience, and academic resources helps ensure that the person performing the data analysis follows technology developments and rapidly changing job responsibilities. Regular reviews of policies and procedures also help to ensure that organizations stay abreast of the latest technology and change laws or rules.

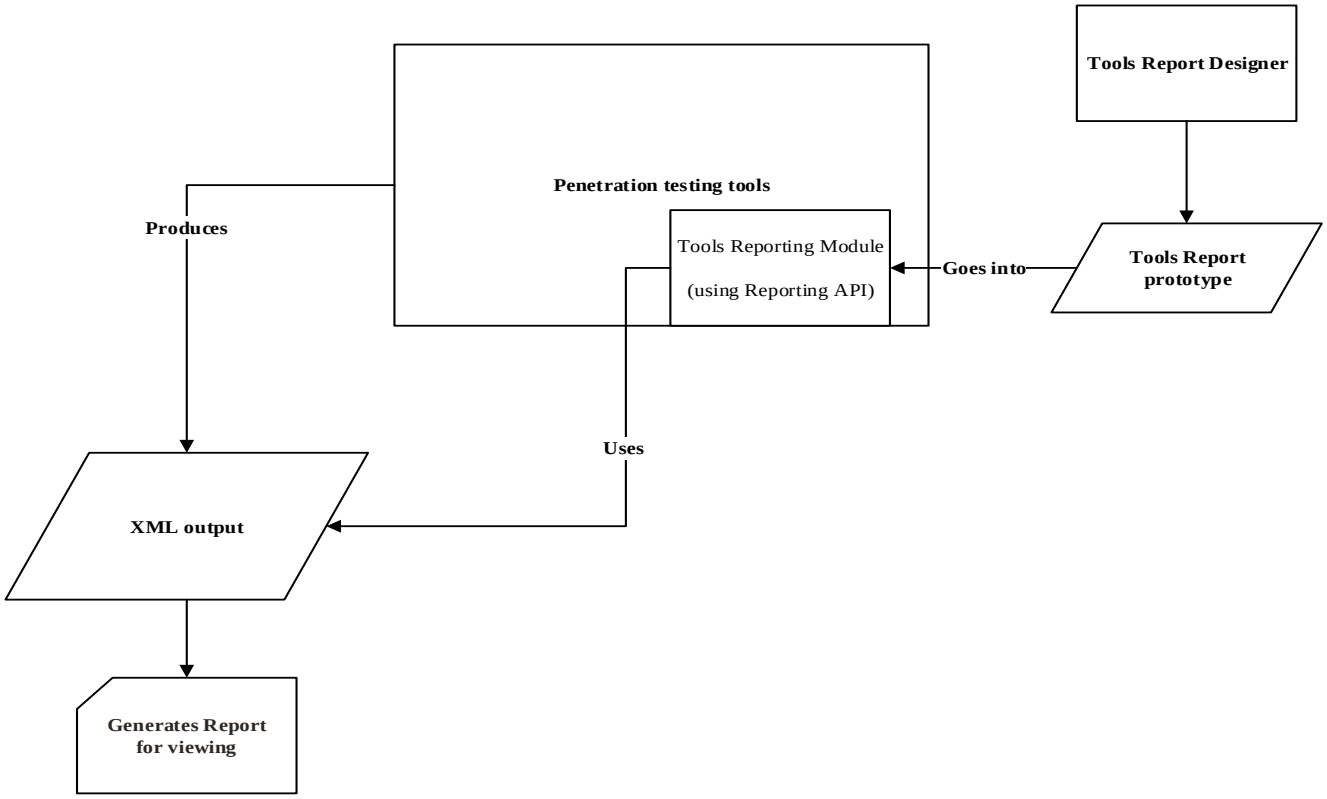


Fig. 7. Reporting Process on Vulnerability Scanning Tools.

III. Result and Analysis

Based on the results of the research and analysis carried out, there are criteria scanning analysis as a parameter to explain the expected results related to the most critical web application security risks in Table IV.

Identifying and analysing the vulnerability scanning process from this research can following steps will be taken to obtain results in the form of digital evidence as reports from ZAP and Arachni vulnerability scanning tools.

A. Acquisition

The acquisition of this research is to run a vulnerability scanner on the client to obtain vulnerabilities on the E-voting web server. In Fig. 8, we can see that the E-voting scan on the web server is ready to work.

TABLE. IV. Severity Level and Vulnerability Parameters

No	Severity Level	Vulnerability Parameters	Result
1	High	Cross-Site Scripting XSS	Yes
2	Medium	Source code disclosure / Application Error Disclosure	Yes
3	Medium	Common Directory / Directory Browsing	Yes
4	Low	Missing 'X-Frame-Options' header / X-Frame Options Header Not Set	Yes

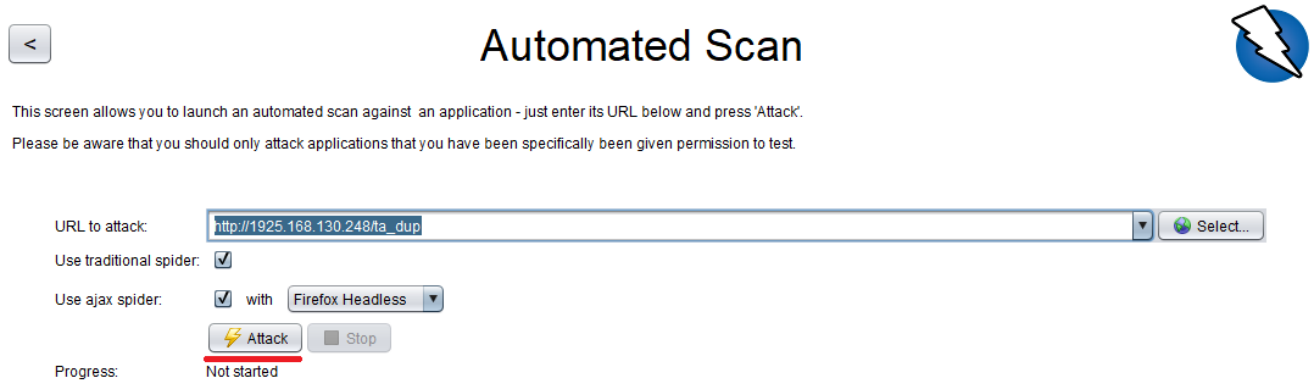


Fig. 8. Automated Scan Facilities.

Requirements must be worked out in preparation for this research.

1) All major components (Apache[11], MySQL[12], and web application[13]) in XAMPP[14] are ready to be scanned on the Server Computer.

2) All components of XAMPP is working on the server.

3) Vulnerability scanner installed on a computer.

a) ZAP and Arachni must load the Web Application URL as a target before scanning.

b) Then carry out scanning execution automatically by pressing the Attack button, so this application scans the target.

4) Finally, get a report from the vulnerability scanner.

The acquisition uses a vulnerability scanner as proof that it is scanning the target for the IP address by showing hexadecimal and the text in Fig. 9.

As shown in Fig. 10, when the scan is in progress, it can be seen in the Active Scan section with valuable information in the form of Request Timestamp, Response Timestamp, methods, and URL.

Depending on the device and server computer, this scanning process takes 15 minutes using ZAP, then 1 hour and 45 minutes using Arachni, because other factors influence this process. After finished scanning the server, the next step is to collect evidence and reports from Arachni as a vulnerability scanning tool, then the entire acquisition process using ZAP and Arachni is finished, we then proceed to the inspection process.

B. Examination

Previously we have made an acquisition, then an examination of the results of the acquisition will be carried out on both the vulnerability scanner and the application.

1) Examination of the results of the acquisition by scanning using ZAP has obtained reports. In this research, ZAP has conducted scanning for vulnerability to E-voting without constraints. The inspection process using ZAP can be measured because we can do the same thing in using other vulnerability scanning tools. ZAP scanning measures the level of vulnerability that exists in E-voting and application.

Fig. 11 explains ZAP has obtained and collected the results as evidence of response messages that come from scanning the Application, so the scanning process using ZAP will explain to us about valuable information relating to vulnerabilities.

2) Arachni reports as a whole that recorded on the results of the webserver scan with a report explaining starting from the base URL to the web application directory in Fig. 12.

The scan produced by Arachni is known to be the response message that comes from scanning the entire web server and web application. So it requires a re-examination of the scan results. However, the results reported by Arachni are relatively the same as ZAP.

C. Utilization

ZAP and Arachni scanners utilize response messages from the application in the form of reports that are easy to use. So in the process of utilization, we also get the convenience of the tools provided by exploring the Desktop Application interface on ZAP and WebUI on Arachni by merely clicking on the menu button available. This step will also provide appropriate information from both vulnerability scanning tools.

1) Utilization of ZAP: ZAP used to scan has many categories of results of the scan that have a level of risk in the results of the scan, one of its features is the ZAP Scanning Report, where ZAP can obtain classified evidence so that practitioners can utilize the information provided by ZAP Scanning Report. In this *research*, we can obtain vulnerability information that has classified and use it through this facility.

As shown in Fig. 13 shows the reporting results of the investigation results from ZAP. However, as we can see, the source comes from web applications and server computers with file extensions, and IP addresses http://192.168.130.248/ta_dup/in more detail.

2) Utilization of Arachni: Using Arachni can report the results of the scan, the scan results extracted, and risk levels have arranged in Fig. 14. Similar to ZAP, Arachni also explained the scan results in the form of IP address and direction.

D. Review

We have conducted investigations including acquisition, examination, utilization, and then the final step is to conduct a review. ZAP and Arachni have successfully obtained digital evidence using scanning through a computer network. Evidence in the form of reports obtained from a vulnerability scanner and then carried out an analysis, the details of the report contain descriptions, URLs, methods, parameters, information, and evidence.

[11] https://en.wikipedia.org/wiki/Apache_HTTP_Server
[12] https://en.wikipedia.org/wiki/MySQL
[13] https://en.wikipedia.org/wiki/Web_application
[14] https://en.wikipedia.org/wiki/XAMPP

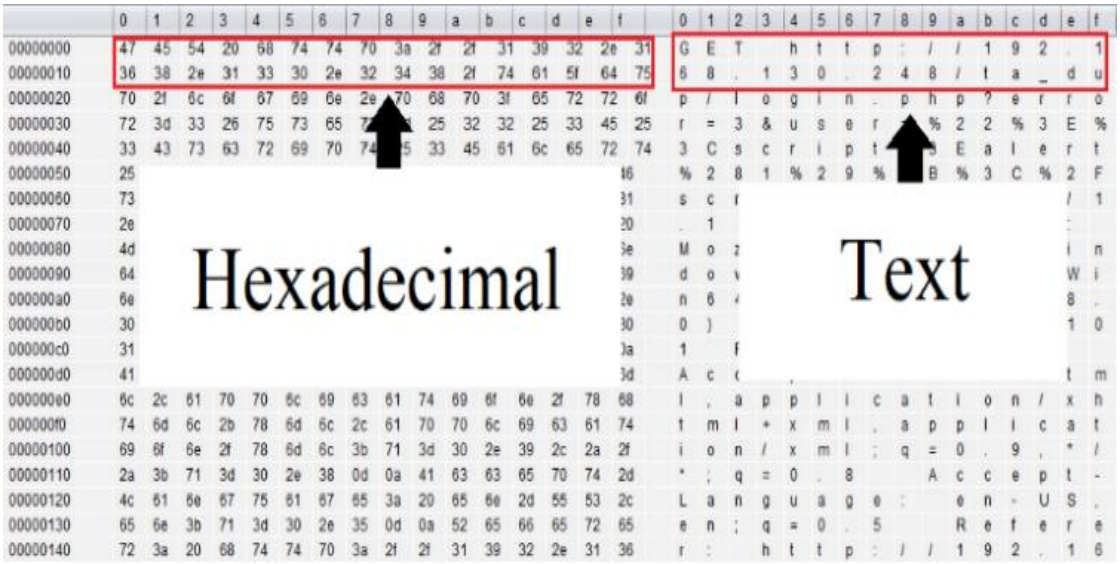


Fig. 9. Request Scan in Hexadecimal and Text Form.

Id	Req. Timestamp	Resp. Timestamp	Method	URL
363	8/29/19, 7:04:14 PM	8/29/19, 7:04:14 PM	GET	http://192.168.130.248/ta_dup/asset
364	8/29/19, 7:04:14 PM	8/29/19, 7:04:14 PM	GET	http://192.168.130.248/ta_dup/asset/bootstrap
365	8/29/19, 7:04:14 PM	8/29/19, 7:04:14 PM	GET	http://192.168.130.248/ta_dup/asset/bootstrap/css
366	8/29/19, 7:04:14 PM	8/29/19, 7:04:14 PM	GET	http://192.168.130.248/ta_dup/asset/css
367	8/29/19, 7:04:14 PM	8/29/19, 7:04:14 PM	GET	http://192.168.130.248/ta_dup/asset/Images
368	8/29/19, 7:04:15 PM	8/29/19, 7:04:15 PM	GET	http://192.168.130.248/ta_dup/login.php?error=c...
369	8/29/19, 7:04:15 PM	8/29/19, 7:04:15 PM	GET	http://192.168.130.248/ta_dup/login.php?error=.....

Fig. 10. Progress of Scanning.

URL	http://192.168.130.248/ta_dup/login.php?error=4&user=%22%3E%3Cscript%3Ealert%281%29%3B%3C%2Fscript%3E
Method	GET
Parameter	user
Attack	"><script>alert(1);</script>
Evidence	"><script>alert(1);</script>

Fig. 11. ZAP Scan Report.

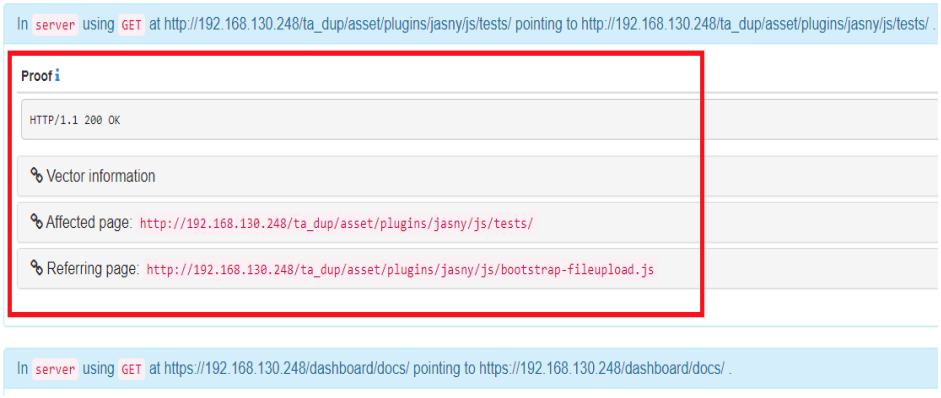
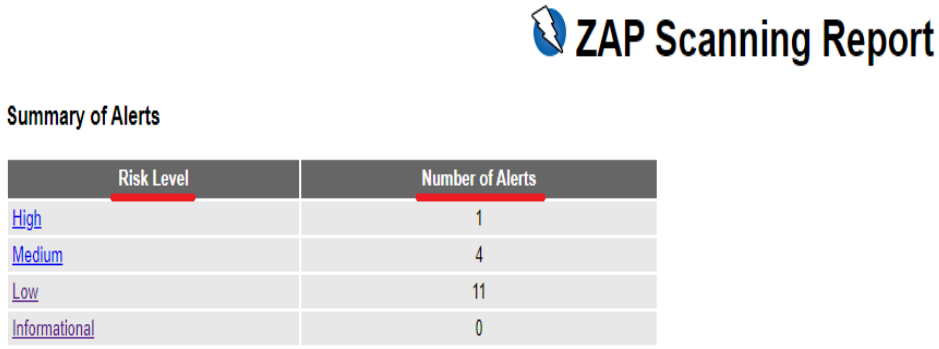


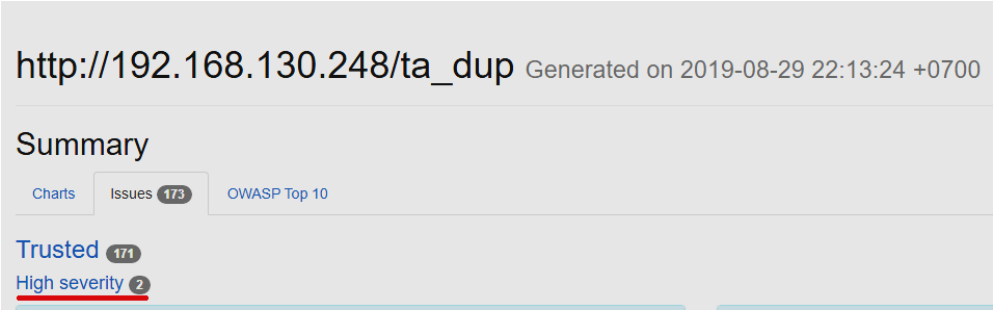
Fig. 12. Arachni Scan Report.


ZAP Scanning Report

Summary of Alerts

Risk Level	Number of Alerts
High	1
Medium	4
Low	11
Informational	0

Fig. 13. A Report of the Results by ZAP.


http://192.168.130.248/ta_dup Generated on 2019-08-29 22:13:24 +0700

Summary

Charts Issues 173 OWASP Top 10

Trusted 171

High severity 2

Fig. 14. A Report of the Results by Arachni.

IV. Conclusion

Based on the results of the research that has reviewed, ZAP and Arachni completed the scanning and analysis of reports that were measured and run through the client's computer. ZAP and Arachni managed to get evidence in the form of descriptions, URLs, methods, parameters, information, and evidence: (1) 1 high level in ZAP & 1 high level in Arachni, (2) 4 medium levels in ZAP & 3 medium levels on Arachni, (3) 11 low levels in ZAP & 2 low levels in Arachni. We hope that ZAP and Arachni can be developed to identify digital evidence of vulnerabilities in mobile applications.

References

[1] R. Krimmer, "The Art of Structuring," in The Art of Structuring, Springer International Publishing, 2019, pp. 421–426.

[2] B. Rexha and I. Murturi, "Applying Efficient Crowdsourcing Techniques for Increasing Quality and Transparency of Election Processes," Electron. Gov., vol. 15, no. 1, pp. 107–128, 2019.

[3] S. Djanali, D. P. Nugraha, H. Studiawan, and B. Adi Pratomo, "Vote identification and integrity of ballot in paper-based e-voting system," Electron. Gov., vol. 14, no. 3, pp. 240–254, 2018.

[4] T. D. Wagner, E. Palomar, K. Mahbub, and A. E. Abdallah, "A Novel Trust Taxonomy for Shared Cyber Threat Intelligence," Secur. Commun. Networks, vol. 2018, 2018.

[5] A. Fadlil, I. Riadi, and S. Aji, "Review of detection DDOS attack detection using naive bayes classifier for network forensics," Bull. Electr. Eng. Informatics, vol. 6, no. 2, pp. 140–148, 2017.

[6] A. R. Caesarano and I. Riadi, "Network Forensics for Detecting SQL Injection Attacks using NIST Method," Int. J. Cyber-Security Digit. Forensics, vol. 7, no. 4, pp. 436–443, 2018.

[7] R. Mardisalu, "14 Most Alarming Cyber Security Statistics in 2019," 2019. [Online]. Available: https://thebestvpn.com/cyber-security-statistics-2019/.

[8] The OWASP Foundation, "ZAP Proxy."

[9] I. Riadi, R. Umar, and W. Sukarno, "Vulnerability of Injection Attacks Against The Application Security of Framework Based Bebsites Open Web Access Security Project (OWASP)," J. Inform., vol. 12, no. 2, pp. 53–57, 2018.

[10] D. Pandya and N. Patel, "Owasp Top 10 Vulnerability Analyses in Government," Int. J. Enterp. Comput. Bus. Syst., vol. 6, no. 1, 2016.

[11] M. Volkova, P. Chmelar, and L. Sobotka, "Machine Learning Blunts the Needle of Advanced SQL Injections," Mendel, vol. 25, no. 1, pp. 23–30, 2019.

[12] A. Yassine, E. H. Ettifouri, J. Berrich, and B. Toumi, "Modeling The OWASP Most Critical WEB Attacks," in International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning, 2019, pp. 442–450.

[13] J. M. Dharmalingam and M. Eswaran, "An agent based intelligent dynamic vulnerability analysis framework for critical SQLIA attacks: Intelligent SQLIA vulnerability analyzer agent," Int. J. Intell. Inf. Technol., vol. 14, no. 3, pp. 56–82, 2018.

[14] J. Blue and E. Furey, "A Novel Approach for Protecting Legacy Authentication Databases in Consideration of GDPR," 2018 Int. Symp. Networks, Comput. Commun. ISNCC 2018, pp. 1–6, 2018.

[15] H. Sohoel, M. G. Jaatun, and C. Boyd, "OWASP Top 10 - Do Startups Care?," 2018 Int. Conf. Cyber Secur. Prot. Digit. Serv. Cyber Secur. 2018, no. 0102, 2018.

[16] M. Srokosz, D. Rusinek, and B. Ksiezopolski, "A new WAF-based architecture for protecting web applications against CSRF attacks in malicious environment," Proc. 2018 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2018, vol. 15, pp. 391–395, 2018.

[17] R. Wang, G. Xu, X. Zeng, X. Li, and Z. Feng, "TT-XSS: A novel taint tracking based dynamic detection framework for DOM Cross-Site Scripting," J. Parallel Distrib. Comput., vol. 118, pp. 100–106, 2018.

[18] G. S. Leite and A. B. Albuquerque, "An Approach for Reduce Vulnerabilities in Web Information Systems," in Proceedings of the Computational Methods in Systems and Software, 2019, pp. 86–99.

[19] C. Gil, L. Baquero, and M. Hernández, "A Conceptual Exploration for the Safe Development of Mobile Devices Software Based on OWASP," Int. J. Appl. Eng. Res., vol. 13, no. 18, pp. 13603–13609, 2018.

[20] The OWASP Foundation, "Top 10 2017 A10-Insufficient Logging & Monitoring," 2017. [Online]. Available: https://www.owasp.org/index.php/Top_10-2017_A10-Insufficient_Logging%26Monitoring.

[21] The OWASP Foundation, "OWASP: The Ten Most Critical Web Application Security R isks," 2017. [Online]. Available: https://www.owasp.org/index.php/Top_10-2017_Top_10.

[22] A. Kurniawan, I. Riadi, and A. Luthfi, "Forensic Analysis and Prevent of Cross Site Scripting in Single Victim Attack using Open Web Application Security Project (OWASP) Framework," J. Theor. Appl. Inf. Technol., vol. 95, no. 6, pp. 1363–1372, 2017.

[23] M. Alzahrani and L. Georgieva, "Structural modeling and verification of web applications," in Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, 2018.

[24] Kaspersky, "Adversary Attack Vectors 4," 2018.

[25] D. Saputra and I. Riadi, "Network Forensics Analysis of Man in the Middle Attack Using Live Forensics Network Forensics Analysis of Man in the Middle Attack Using Live Forensics Method," Int. J. Cyber-Security Digit. Forensics, vol. 8, no. 1, pp. 66–73, 2019.

[26] K. Pandey, "A Bug Tracking Tool for Efficient Penetration Testing," Int. J. Educ. Manag. Eng., vol. 8, no. 3, pp. 14–20, 2018.

[27] The OWASP Foundation, "OWASP Risk Rating Methodology," 2019. [Online]. Available: https://www.owasp.org/index.php/Threat_Risk_Modeling.

[28] The OWASP Foundation, OWASP Top 10 – 2017 The Ten Most Critical Web Application Security Risks. 2017.

[29] I. Riadi, R. Umar, and A. Firdonsyah, "Forensic tools performance analysis on android-based blackberry messenger using NIST measurements," Int. J. Electr. Comput. Eng., vol. 8, no. 5, pp. 3991–4003, 2018.

[30] I. Riadi, R. Umar, and A. Firdonsyah, "Identification Of Digital Evidence On Android's Blackberry Messenger Using NIST Mobile Forensic Method," Int. J. Comput. Sci. Inf. Secur., vol. 15, no. 5, pp. 155–160, 2017.

[31] A. Yudhana, I. Riadi, and I. Anshori, "Identification of Digital Evidence Facebook Messenger on Mobile Phone With National Institute of Standards Technology (Nist) Method," Kursor, vol. 9, no. 3, pp. 111–118, 2019.

[32] Sunardi, I. Riadi, and A. Sugandi, "Forensic analysis of Docker Swarm cluster using GRR Rapid Response framework," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 459–466, 2019.

[33] The OWASP Foundation, "GSoC2013 Ideas/OWASP ZAP Exploring Advanced reporting using BIRT," 2013. [Online]. Available: https://www.owasp.org/index.php/GSoC2013_Ideas/OWASP_ZAP_Exploring_Advanced_reporting_using_BIRT.

On the Prediction of Properties of Benzene using MP4 Method Executed on High Performance Computing with Heterogeneous Platform

Norma Alias[1], Riadh Sahnoun[2]
Ibnu Sina Institute for Fundamental Science Studies
Universiti Teknologi Malaysia, Johor, Malaysia

Nur Fatin Kamila Zanalabidin[3]
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, Johor, Malaysia

***Abstract*—High computational complexity problem, high computational cost and deal with a big data are the motivation to study the physical and chemical properties of benzene. Based on the limitation of memory system, processor speed and huge time step computation, we propose the implementation of parallel Gaussian suites of program, particularly the program dealing with high order Møller–Plesset perturbation theory, on high performance homogeneous computing platform (HPC) for predicting the physical and chemical properties of small to medium size molecules, such as benzene, the subject of the present work. Besides high accuracy of the geometrical parameters that can be offered by MP4 simulation, orbital shapes, HOMO-LUMO energy gaps and spectral properties of the molecule are among the properties that can be obtained with accurate prediction. In order to achieve high performance indicators, we need to execute the program in multiple instruction and data stream (MIMD) paradigm using homogenous processors architecture. At the end of this paper, it is shown that Parallel algorithm of Gaussian program using the Linda software can be executed and is well suited in both homogenous and heterogeneous processors. The performance evaluation is essentially based on run time, temporal performance, effectiveness, efficiency, and speedup.**

Keywords*—MP4; parallel computing; homogenous platform; properties of benzene; high performance computing

I. INTRODUCTION

Quantum chemists running electronic structure calculations seem to have an insatiable need for CPU resources in order to get their results faster, or run bigger jobs, or both. Performing high theoretical level calculation is always the most desired target for quantum chemists as performed in paper [1] and [2]. With the recent advancement in computing facilities, calculation of medium size molecules using high theoretical level has become affordable.

In the field of computational chemistry, one of a few quantum chemistry post-Hartree–Fock methods is Møller–Plesset perturbation theory, or MP. By adding electron correlation effects of Rayleigh–Schrödinger perturbation theory, Hartree–Fock method is enhanced [3]. Second (MP2), third (MP3) [4; 5], and fourth (MP4) [6; 7] order Møller–Plesset calculations are enforced in many computational chemistry codes and are common levels implied in calculating modest systems. MP2 imitates dispersion-corrected MP2 coupled model, which is outstanding in a framework with five global empirical parameters and new analogous uncoupled Hartree–Fock [8]. Also, MP2 is a preferable theory due to its relatively moderate computational cost and is known as one of the most extensively used methods for studying noncovalent interactions in quantum chemistry [9]. Paper [10] also stated that single point energy calculation proposes an accurate and fast compromise by combining a MP2 with a Hartree–Fock geometry optimization.

While single point energy calculation at the MP4 level is affordable, full geometry optimizations are rarely reported because of the computational cost, complexity and time-consuming calculations. A single point energy calculation only calculates the energy of the atoms in their current positions, that is, with absence of molecular vibration. Geometry optimizations is a method to determine the minimum energy structure of a system, by modifying the geometry at every stage until a lowest-energy arrangement is found and performing an iterative series of single point calculations. Geometry optimizations mainly rely on the gradient of the first derivative of the energy with respect to atomic positions. Instead of using only a single-point determination, an optimization inevitably requires a lot more computational resources.

Gaussian program package [11] is one of the most popular commercial quantum chemistry codes for which a wide variety of quantum chemistry calculations, including MPn (2-5), can be executed. Depending on the type of operating system and interface, the software can be executed either in single processor, or in parallel execution using multiple processors.

The Gaussian 09 (or simply G09) can be installed for parallel use (shared memory, multi-processor) within a single computer node, serial use (single processor), and distributed use (multi-processor, distributed memory) using TCP Linda. This model implicates a process involving master and a number of workers where in a master process, the programming runs on the current processor. Meanwhile, the programming can run on other nodes of the network in workers processes. In order to execute Gaussian 09/Linda parallel computing, one should fix the list of processors including the jobs that should be executed, job parameters, and the number of processors to use.

Apart from that, every nodes requires some access to the Gaussian 09 directory tree. In our custom configured HPC we used the suggested configuration, where each system will using

G09 for the parallel work. Note that the Linda binaries need to have the same path on each machine. Parallel performance evaluations are established on temporal performance, effectiveness, efficiency, speedup and time execution [12].

II. Gaussian Modeling

In calculating MP2 and beyond, some local disk must be inserted in each node so that Gaussian 09 can put temporary files. The details are in the GAUSS_SCRDIR environment variable, which should be set in the .cshrcor and .profile files for the user account on each node.

Our custom configured workstation is equipped with dual-processor (double threaded) quad-core nodes. In a single node, one can use a maximum of 8-processors (cores) to run parallel Gaussian Job. In MPn calculations it is important to request multiple processors, since the processing time for this type of calculation is very long. To perform parallel Gaussian 09 job in single-node multi-processor, the TCP Linda uses shared memory. The specifications of the number of processors using shared memory need to be thorough in the Gaussian input file. This can be specified by embedding the following line at the beginning of the Gaussian input file:

%NProcShared=n

The above line indicates the "n" processors that will share the memory within the computer node assigned to execute the job. In our calculation, the number of processors was varied from one (1) to the maximum number of eight (8) processors.

III. Sequential Algorithm

All calculations were performed using the Gaussian 09 suites of programs. The model system used was the benzene molecule, for which unconstrained geometry optimization was carried out employing the standard 6–311G(3df,3dp) basis set as implemented in Gaussian 09 program package. The fourth order Møller Plesset perturbation theory employing the space quadruple, double and single substitutions, MP4(SDQ) was used throughout the calculations.

Fig. 1 summarizes the sequence of the steps carried out during MP4(SDQ) geometry optimization.

Steps from "Berny optimizations to locate minimum and transition state" until the last step of "Process information for optimizations and frequencies" will be repeated iteratively until a local minimum is located in the potential energy surface (minimum energy is found). In our calculations, the local minimum was found after four iterations.

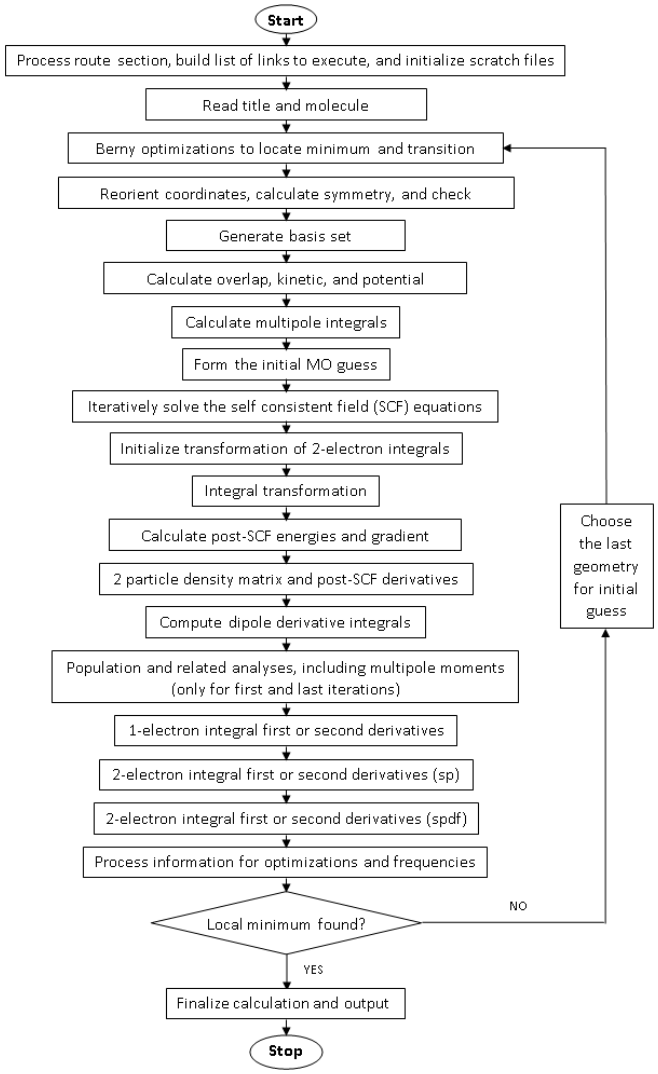


Fig. 1. Flowchart of Sequential Algorithm MP4 Calculation using Gaussian Program.

IV. Parallel Algorithm

Fig. 2 shows the flowchart of parallel algorithm MP$ calculations using Gaussian Method. The calculations were carried out in two types of machine: one is HP and the other one is custom-configured. Both machines have 16 GB of RAM, for which only 2 GB was specified in the input stream of the Gaussian job. As for the hard disk, the HP type has 3 x 1 TB of hard disk while the custom-configured has 3 x 2 TB.

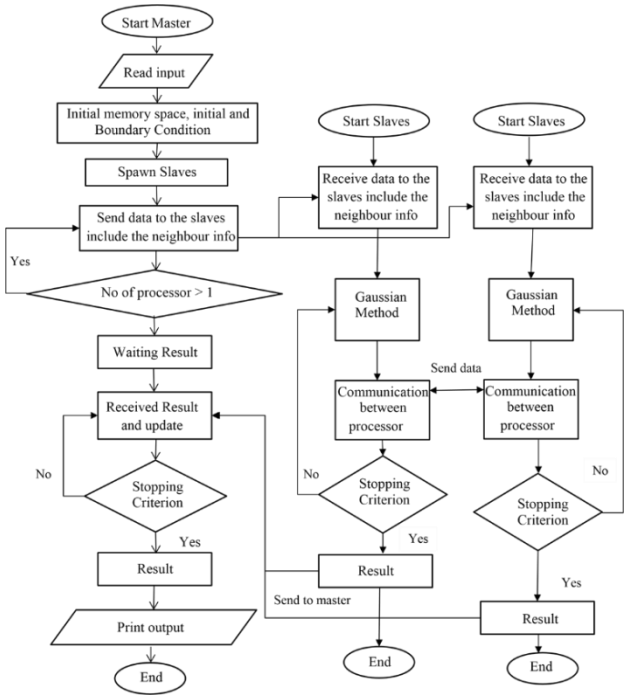


Fig. 2. Flowchart of Parallel Algorithm MP4 Calculation using Gaussian Method.

V. Performance Measure

Run Time, Speedup, Effectiveness, Efficiency, and Temporal Performance are used to measure the performances of parallel strategies.

Run time is the period of time obligated for a program routine to be completed and the time needed for a computer to interpret and carry out an accumulated command. Also, by measuring an algorithm's speedup we can determine its performance since we are occasionally concerned about the performance gains from the algorithm. Amdahl's law emphasized that Speedup can be described as the time taken for a computer to enact in serial divided by the time taken to enact in parallel. The measurements of Speedup are defined as follows;

$$\text{Speedup } (S_P) = t_1 / t_p \quad (1)$$

Now, a calibration of processor usage is called the Efficiency of a parallel algorithm and can be utilized to determine its effectiveness. It is easily explained that the Efficiency is the Speedup divided by the number of used processors. The measurements of Efficiency are defined as follows:

$$\text{Efficiency } (E_p) = \text{Speedup } (S_P) / P \quad (2)$$

Next, by evaluating the Speedup and Efficiency, Effectiveness of a method using parallel algorithm can be determined. The measurements of Effectiveness are defined as follows:

$$\text{Effectiveness } (F_p) = S_P / Pt_p = E_p / t_p \quad (3)$$

The Temporal Performance is a parameter to evaluate the performance of a parallel algorithm, defined as:

$$\text{Temporal} = 1/\text{Time } (p), \quad (4)$$

where Time (p) = time execution using p processor.

VI. Results and Discussion

To predict the physical and chemical properties of small to medium size molecules, such as benzene, the subject of this work, at MP4(SDQ) level of theory, we have implemented the Gaussian suite of program on high performance homogeneous computing platform.

The Gaussian software, particularly the program dealing with high order Møller–Plesset perturbation theory including MP4(SDQ), supports sending and receiving data activities between processors. Note that the parallel computing's performance was evaluated on the prospect of Temporal Performance, Effectiveness, Efficiency, Speedup and Run Time based on the numerical results attained. Here, Table I and Table II show the results of parallel performance evaluation.

Fig. 3 and Fig. 4 show the run time in seconds for 8 types of number of processors – 1, 2, 3, 4, 5, 6, 7 and 8. Observing the graph, execution time decreases with increasing number of processors. This might be attributed to the fact that the task from the master is split and shared between all slaves. It means that the higher the number of utilized processors, the higher the number of slaves the master can involve in the execution of the task which contributes in lowering the execution time. Hence, with the increasing number of processors used, the execution time decreases.

TABLE. I. Run Time, Speedup, Efficiency, and Effectiveness with Respect to the Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM)

No. of Processors	Run time	Speedup	Efficiency	Effectiveness	Temporal Performance
1	49466	1.00	1.00	2.02159E-05	2.02159E-05
2	23174	2.13	1.07	4.60548E-05	4.31518E-05
3	15459	3.20	1.07	6.89928E-05	6.46858E-05
4	11764	4.20	1.05	8.93625E-05	8.50069E-05
5	10867	4.55	0.91	8.37756E-05	9.20217E-05
6	9273	5.33	0.89	9.5884E-05	0.000107844
7	8271	5.98	0.85	0.000103298	0.000120904
8	7638	6.48	0.81	0.000105999	0.000130931

TABLE. II. Run Time, Speedup, Efficiency, and Effectivenesswith Respect to the Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM)

No. of Processors	Run time	Speedup	Efficiency	Effectiveness	Temporal Performance
1	41800	1.00	1.00	2.392E-05	2.392E-05
2	21156	1.98	0.99	4.670E-05	4.727E-05
3	14052	2.97	0.99	7.056E-05	7.116E-05
4	10584	3.95	0.99	9.329E-05	9.448E-05
5	9171	4.56	0.91	9.940E-05	1.090E-04
6	7765	5.38	0.90	1.155E-04	1.288E-04
7	7034	5.94	0.85	1.207E-04	1.422E-04
8	6558	6.37	0.80	1.215E-04	1.525E-04

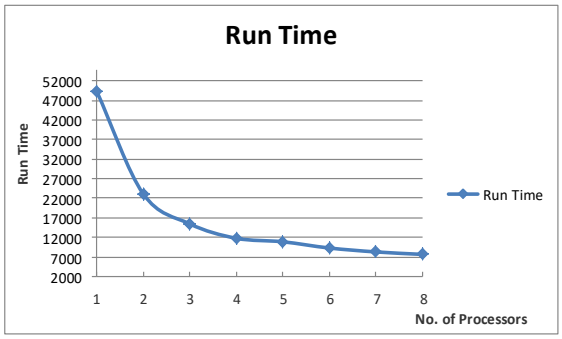


Fig. 3. Run Time vs. Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM).

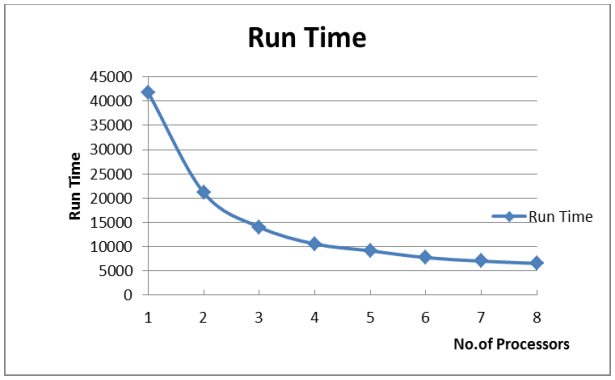


Fig. 4. Run Time vs. Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM).

Inspection of Fig. 5 and Fig. 6 show that the parallel algorithm's speedup increases with increase in the number of processors used. These results are expected because by increasing the number of processors, the calculation will, generally, perform faster.

From the graphs plotted in Fig. 7 and Fig. 8, one can conclude that, the parallel algorithm's efficiency decreases with the increasing number of processors. It is known that the efficiency is the ratio of speedup with respect to the number of processors. Therefore, the efficiency is a performance that is closely related to the speedup.

In Fig. 9 and Fig. 10 are plotted the effectiveness against the number of processors. It can be noticed that as the processors used increases, the effectiveness also escalates. As shown in Equation (3) the effectiveness is proportional to the speedup.

Finally, the graphs plotted in Fig. 11 and Fig. 12 show that as the number of processors is increasing, the temporal performance increases. It is escalating with respect to the increase in the number of processors used. Overall, from these prospects mentioned above, it can be seen that by using more processors, the parallel algorithm can be substantially improved.

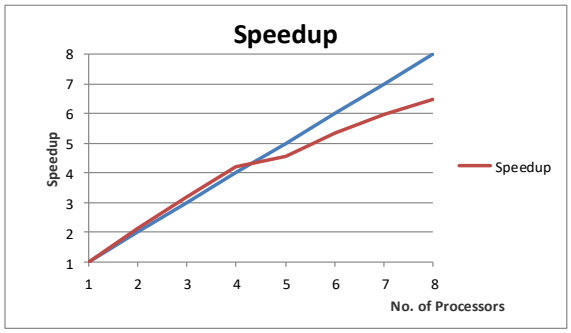


Fig. 5. Speedup vs. Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM).

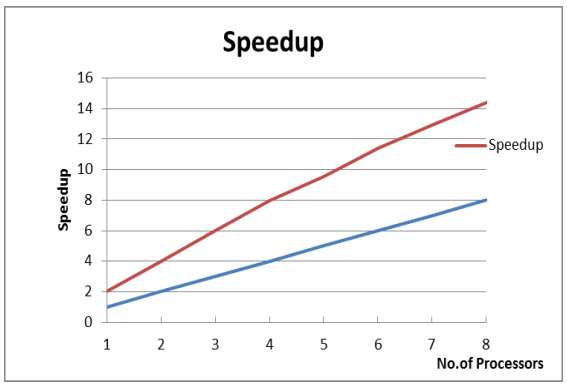


Fig. 6. Speedup vs. Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM).

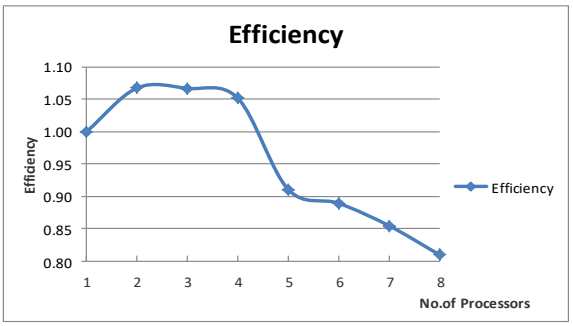


Fig. 7. Efficiency vs. Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM).

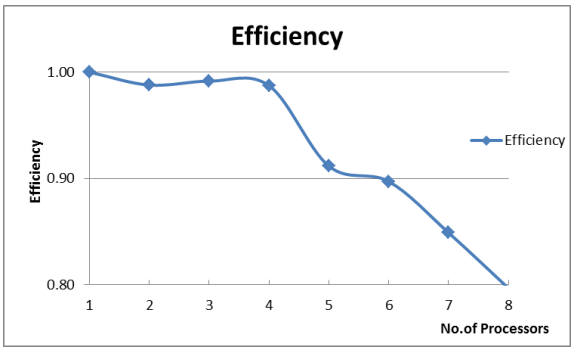


Fig. 8. Efficiency vs. Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM).

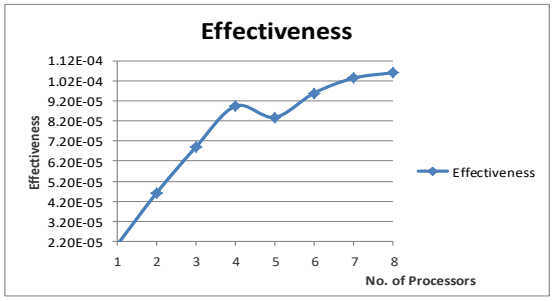


Fig. 9. Effectiveness vs. Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM).

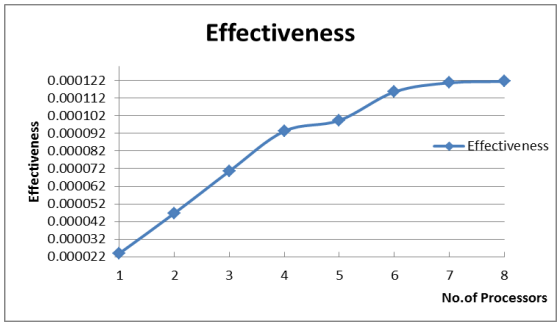


Fig. 10. Effectiveness vs. Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM).

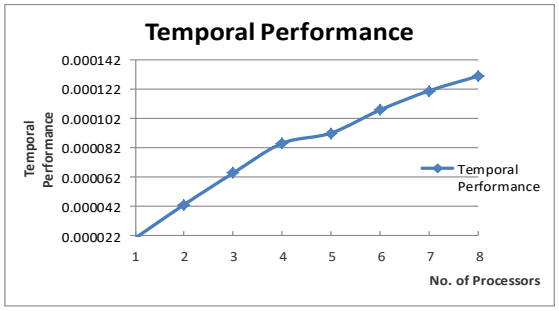


Fig. 11. Temporal Performance vs. Number of Processors (RSN003, 1X3TB, 1 Hitachi + 2 WD BLACK, 16GB RAM).

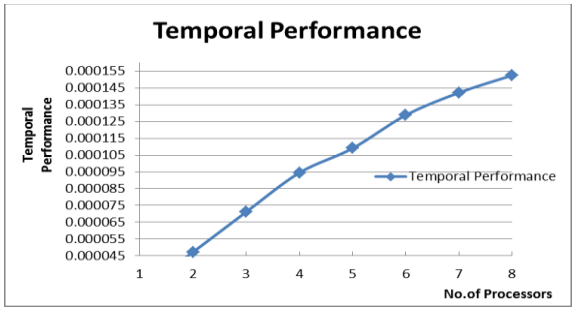


Fig. 12. Temporal Performance vs. Number of Processors (RSN015, 2X3TB, WD BLACK CAVIAR, 16GB RAM).

VII. Conclusion

In conclusion, the parallel algorithms of Gaussian program are well suited for solving the large sparse problem of physical and chemical properties of benzene based on MP4 method. The program is available executed on high performance computing heterogeneous computing platform with different operating systems. Some parallel performance measurements show that the proposed model is accurate prediction of physical and chemical properties of benzene. For a future research, it is recommended to implement the Gaussian program based on MP4 method in solving other large sparse problem.

References

[1] P. K. Barkoutsos, J. F. Gonthier, I. Sokolov, N. Moll, G. Salis, A. Fuhrer, M. Ganzhorn, D. J. Egger, M. Troyer, & A. Mezzacapo. (2018). Quantum algorithms for electronic structure calculations: Particle-hole Hamiltonian and optimized wave-function expansions. Physical Review A, 98(2), 022322.

[2] A. Kuc, & T. Heine. (2015). The electronic structure calculations of two-dimensional transition-metal dichalcogenides in the presence of external electric and magnetic fields. Chemical Society Reviews, 44(9), 2603-2614.

[3] K. Simeonov. (2009). Higher fullerenes: isolation, halogenation and structural studies.

[4] H. Yamada, Y. Mochizuki, K. Fukuzawa, Y. Okiyama, & Y. Komeiji. (2017). Fragment molecular orbital (FMO) calculations on DNA by a scaled third-order Møller-Plesset perturbation (MP2. 5) scheme. Computational and Theoretical Chemistry, 1101, 46-54.

[5] J. A. Pople, J. S. Binkley, & R. Seeger. (1976). Theoretical models incorporating electron correlation. International Journal of Quantum Chemistry, 10(S10), 1-19.

[6] P. Goel, & J. F. Stanton. (2018). Semiclassical transition state theory based on fourth order vibrational perturbation theory: Model system studies beyond symmetric Eckart barrier. The Journal of chemical physics, 149(13), 134109.

[7] J. Z. Gong, D. A. Matthews, P. B. Changala, & J. F. Stanton. (2018). Fourth-order vibrational perturbation theory with the Watson Hamiltonian: Report of working equations and preliminary results. The Journal of chemical physics, 149(11), 114102.

[8] J. Rezac, C. Greenwell, & G. J. Beran. (2018). Accurate Noncovalent Interactions via Dispersion-Corrected Second-Order Møller–Plesset Perturbation Theory. Journal of chemical theory and computation, 14(9), 4711-4721.

[9] R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Hargus, K.-H. Law, J. L. Klepeis, & D. E. Shaw. (2017). Improving the accuracy of Møller-Plesset perturbation theory with neural networks. The Journal of chemical physics, 147(16), 161725.

[10] B. D. Sellers, N. C. James, & A. Gobbi. (2017). A comparison of quantum and molecular mechanical methods to estimate strain energy in druglike fragments. Journal of chemical information and modeling, 57(6), 1265-1275.

[11] J. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, & O. Kitao. Gaussian 09, Revision C. 01. In. 2010. Gaussian, Inc.: Wallingford, CT.

[12] N. Alias, H. F. S. Saipol, A. Ghani, A. Che, & M. N. Mustaffa. (2014). Parallel performance comparison of alternating group explicit method between parallel virtual machine and matlab distributed computing for solving large sparse partial differential equations. Advanced Science Letters, 20(2), 477-482.

A Framework for Traceable and Transparent Supply Chain Management for Agri-food Sector in Malaysia using Blockchain Technology

Kok Yong Chan[1], Johari Abdullah[2], Adnan Shahid Khan[3]
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

***Abstract*—This paper presents a framework for traceable and transparent supply chain management (SCM) system for the agri-food sector using blockchain technology in Malaysia. Numerous researchers believed that the current SCM system consists of several weak points, especially when multiple enterprise resource planning (ERP) system utilizing centralized SCM. Thus, data transparency and traceability are limited. This study hypothesized that if blockchain technology correlates with transparency and traceability of SCM, the above limitation can be minimized, as blockchain technology works in a distributed manner. This research uses "pepper" as an agri-food domain. The research also recommends that permissioned blockchain is a better fit as compared to permissionless blockchain.**

Keywords*—Supply chain; blockchain; consensus algorithm; traceability; transparency

I. Introduction

Supply chains are essentially a series of linked suppliers and customers until products reach the ultimate customer [1]. A good supply chain management should have a high degree of traceability and transparency. Transparency of a supply chain is the extent to which all its stakeholders have a shared understanding and access to the product-related information that request without loss, noise, delay and distortion [2]. Aung and Chang define traceability in terms of what, how, where, why and when aspects of underlying product along a supply chain [3]. Other than managing food quality and safety risks, an effective food traceability system also promotes the development of effective Food Supply Chain management [4]. In recent years, the supply chain industry shows some interest in blockchain technology and some suggest that it might have certain positive impact on the traceability and transparency of supply chain management.

Blockchain is a shared and distributed ledger that used to store transaction records in a business network [5]. Every record stored in the blockchain is automatically encrypted and is decentralized and trackable; hence it meets the requirement of a transaction which is fast and secure. There are a few characteristics in a blockchain network. First is consensus, all parties in the blockchain network must follow a certain protocol to validate a transaction. Second is provenance, the source of information of transaction in a blockchain network is trackable. The third is immutability, there is no chance that a participant in a blockchain network can tamper the information of a transaction when it is recorded and stored in the blockchain. Its usage is best known as the backbone of Bitcoin which is currently one of the most popular cryptocurrencies. However, the ability of blockchain is far beyond the creation of cryptocurrency. Theoretically it can be applied to supply chain management to overcome certain issues.

The business of the agri-food sector, also known as agribusiness, is one of the most important economic sectors in Malaysia, Koo, Othman, Moy, and Khor said that it contributed around 8.6 percent of Gross Domestic Product [6]. Pepper as one of the agri-food in Malaysia, known as the king of Spices, has been exported 12,199mt of pepper worth RM490.2mil to Japan, China, Hong Kong, South Korea, Taiwan, and Singapore in 2016 [7].

However, such a large economy contributor is facing certain challenges. Fletcher reported that up to one hectare of land in the Samunsam Wildlife Sanctuary Sarawak, which is near the border with the Camar Bulan village at West Kalimantan, has been turned into pepper farms by the Indonesians who live there [8]. Mah mentioned that rural pepper farmers from Sarawak still depend on this crop as their main source of income [7]. In Malaysia, the pepper industry has provided employment and income for about 67,000 farming families and households. This incident is highlighting the importance of traceability and transparency in supply chain management to ensure that all production of agri-food is from valid and authorized sources as it is a matter for both the income of the farmer and economy of the country.

A. Problem Statement

The background of the study reflects how important the agribusiness sector is playing the role in the economy of Malaysia. As for that, it is important for us to identify the business friction that impedes the productivity of agribusiness. Improvement of the supply chain is one of the suggestions that could overcome the business friction experienced by the agribusiness industry.

Conventional software solution used in industry is in a centralized model as well as supply chain management solution. With conventional methods for recording transactions and tracking assets, participants on a network keep each own ledgers and other records [5]. This could cause information friction that results in imperfect information and information risks. Imperfect information exists when all the parties in the business network do not have access to the same information

as each of the party is having each own copy of information. This is where the missing of transparency in supply chain management as the information altered accidentally or purposely without acknowledgment of relevant party in the supply chain network.

Most of the existing solutions possess strong traceability, but almost neglect transparency of supply chain management. This lack of transparency is often intentional as some participants want to keep the competitive advantages from other competitors such as an inexpensive supplier who delivers quality products on time. This also implies another issue that its scale may lead to delays and defaults in the delivery of goods [9].

B. Objective

The aim of this research is to design and develop a framework for traceable and transparent supply chain management solution for the agri-food sector in Malaysia using blockchain technology.

Other objectives include:

- To study which kind of blockchain is more suitable to apply to the selected case study (agri-food supply chain)
- To find out how blockchain technology correlates with traceability and transparency of the supply chain.

C. Scope

Blockchain technology remains in infancy, but it is gaining momentum within supply chains, trust being the predominant factor driving adoption. The value of such technologies for supply chain management lies in four areas: transparency, traceability, supply chain digitalization and disintermediation, improved data security and smart contracts [10]. This study will focus on the first two areas which are transparency and traceability.

Other than that, the exploration of blockchain for the purpose of this study will also an emphasis on permissioned or private blockchain technology Hyperledger. Although there might some other private blockchain platform other than Hyperledger, this will just adopt Hyperledger for the solution since the whole point of the research is to compare private and public blockchain platforms and not among private blockchain platforms.

From the domain of supply chain management, this study will focus on the supply chain process of the agri-food sector which is pepper. The environment for implementation and testing the solution will model this supply chain process as much as possible based on the information gathered.

II. Literature Review

Supply chain management is an important piece of enterprise resource planning (ERP). Supply chain management is the oversight of funds, raw materials, components, and finished products, as these items move from suppliers to manufacturers, to wholesalers, to retailers, to consumers. The figure below illustrates the overview of supply chain management of agri-food (pepper) in Malaysia.

Based on Fig. 1, each participant in the supply chain network is linked together with a physical flow of certain products. The whole network of supply chain network of agri-food consists of 5 participants. The role of the farmer is most understandable, which include in charge of farming, harvesting the product and send to the processor. Processor role could be varied depends on the type of food or product in the supply chain network. For instance, the product of pepper will undergo the process of fermentation, washing, drying, and grinding. Once these processes are done, it will be packaged and certified by JAKIM for the halal certificate. The next participant in the supply chain network is distributor, who in charge of storage and warehousing of all the packaged product to be distributed to retailer, wholesaler or other channels so that product is available for the consumer.

The representation of the supply chain management from Fig. 2 seems a simple process however a comprehensive of logistic take place behind the scene when a product flow from one party to another party. Take example of farmer transfers harvested crops to a processor.

In real-world scenario, it usually involved another participant (a middleman) in the supply chain network known as a trusted third-party logistic provider as shown in Fig. 2. According to Chetak Logistic first party logistic (1PL) is referring a firm to make certain shipments from location A to location B by the firm itself; the firm itself has control of all the logistic processes [11]. Second party logistic (2PL) refers to a firm rent the transportation service from a service provider to get the shipment done, while the firm still has control of all the logistic processes. Third party logistic (3PL) is a firm outsource part or all the logistic work to a service provider. One of the roles of 3PL that distinguish itself from 2PL and 1PL is facilitating trust in the supply chain process such as certifying raw materials, components, or finished products, as these items travel through the supply chain. The reason for 3PL is involved in the supply chain network is due to systems operating in a centralized model, where all transactions are executed through a trust-based system making all participants dependent on the third party [12]. The implication of the modern systems operating in a centralized model is that the execution of transactions is dependent on a third party i.e. both the supplier and consumer rely on the third party for the exchange of goods or services [13].

Laney (2018) point out that the current supply chain management system consists of a certain weak point [14]. These weak points happen when there is multiple enterprise resource planning (ERP) systems in use across different organizations as data does not flow well via interface points between systems or individual ledgers during the transference of ownership or change in status between two parties. Thus, data transparency is limited at the hand-off points of raw materials, components, funds, and finished products.

Fig. 1. A Typical Supply Chain in the Retail Industry in Malaysia.

Fig. 2. Third Party Logistic (3PL).

A. Supply Chain Transparency and Traceability

Traceability and transparency are often used as interchangeable terms during the discussion of supply chain management. In order to develop a useful solution for supply chain management, it is important to find out differences between these terms as both of it is two different indicator of a good supply chain management system. The detailed concept of transparency can be visualized in Fig. 3.

Transparency has been defined as the disclosure of information [15] that enables fair competition [16], profitable business ventures [17], and company fulfillment regarding sustainability efforts [18]. According to Tsai, supply chain transparency captures high-level information about a supply chain a product such as product components, names of suppliers, locations, associated certifications, etc. [19].

In other words, transparency decides the breadth of the information or disclosure of information that can be exposed to a certain party. Based on Fig. 4, supply chain transparency enables participants to identify all the suppliers for all the components in a product, down to the provenance. Supply chain transparency also enables participants in the supply chain network to identify and collect previously unknown product information and/or communicate and establish specific requirements. Thus, all the participants can gain better visibility of their global supply chains and ensure compliance with safety, sustainability, and social responsibility requirements.

Tsai mentioned that supply chain traceability captures much more granular information about a product component, such as batch-lot data (harvest date, field data, etc.), purchase order data, and other operational information [19]. Unlike supply chain transparency, the focus of traceability is less on mapping the entire end-to-end supply chain but on following the flow of product components or purchase orders as it moves through the supply chain. The granularity of the data used in traceability allows more targeted recalls, reducing scale and cost.

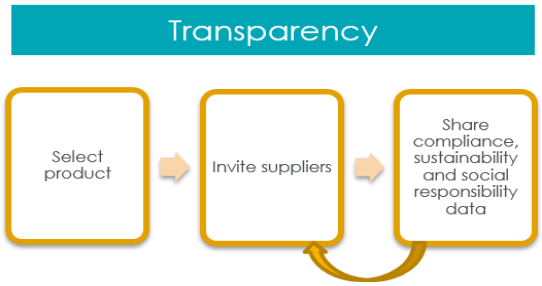


Fig. 3. Supply Chain Transparency

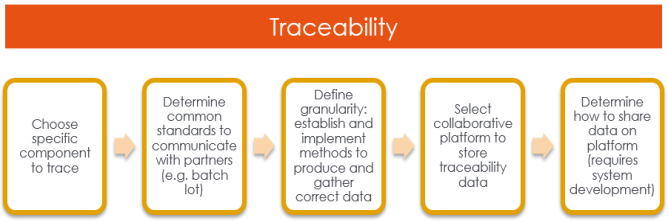


Fig. 4. Supply Chain Traceability.

Fig. 5 depicts the relationship between supply chain transparency and traceability that is traceability relies on transparency [19]. Transparency must come first before traceability data can be collected. Without transparency, it is not practical to implement traceability. For example, granular information such as batch, operational data, purchase order data is almost impossible to gather without information of the supplier. Successful supply chain management requires traceability that is built upon complete transparency.

B. Blockchain

According to Hyperledger, a blockchain is a peer-to-peer distributed ledger forged by consensus, combined with a system for "smart contracts" and other assistive technologies [19]. Blockchain can be categorized as public blockchain and private blockchain, due to these difference public blockchain and private blockchain operate under different consensus algorithms for different purposes.

1) Public Blockchain vs Private Blockchain: In public blockchains (permissionless) anyone can join as regular user or miner. All participants can perform transactions or operate under conditions preset in smart contracts [20]. Bitcoin is the pioneer of the public blockchain platform. It's the first cryptocurrency unveiled using blockchain as the backbone of the infrastructure to store transaction data in an untampered manner with the hash algorithm make it even more secure. A public blockchain network like bitcoin is accessible to anyone around the globe, anyone can make transactions with Bitcoin and anyone could be the miner in the bitcoin blockchain network under the Proof of Work consensus protocol. Miners are those who verify the new transaction and create the new block [21].

Ethereum is the most popular blockchain alternative of Bitcoin which borrows heavily from the Bitcoin architecture. It is a public blockchain with a built-in cryptocurrency called Ether. What distinguishes Ethereum from Bitcoin is that Ethereum is an open source platform that enables developers to build and deploy decentralized applications (dApps). On the other hand, Ethereum has smart contracts that define rules and penalties around an agreement and also enforces those obligations. The role of the smart contract is to keep conditional records and upon occurring specific events [22]. The smart contract allows the performance of transactions without involving any third parties [23].

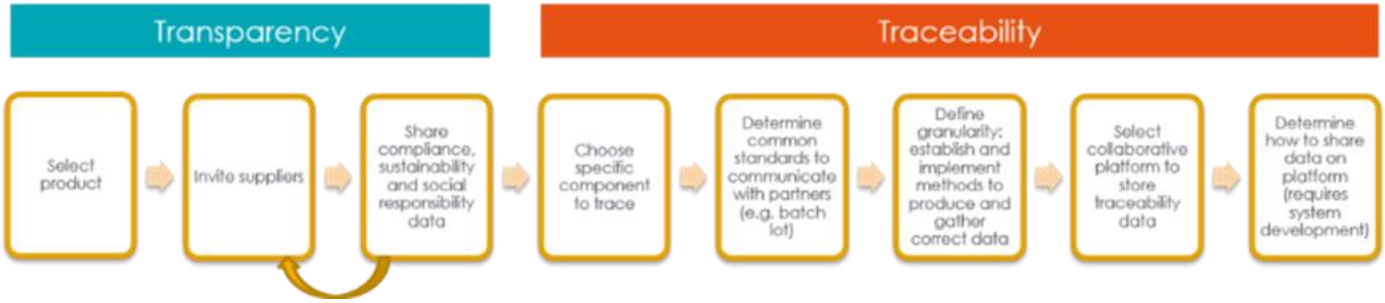


Fig. 5. Supply Chain Transparency and Traceability.

Instructions embedded in Ethereum contracts are paid for in ether (or more technically "gas") and can be implemented in a variety of scripting languages. Ethereum currently uses a proof of work (PoW) protocol but the plan is to update its network to proof of stake (PoS). In a subset of nodes from the Ethereum network, the winner from the miners validates and appends the new block in their ledger and broadcast it to other nodes [24]. Ethereum provides a cryptocurrency called Ether, which can be transferred between accounts and used as reward for miner who successfully approves the transaction [25].

Permissioned blockchain only allows authorized nodes that are pre-registered by authorities to join the blockchain network [26]. Hyperledger Fabric is a permissioned blockchain infrastructure providing a modular architecture with a delineation of roles between the nodes in the infrastructure, execution of Smart Contracts and configurable consensus and membership services. Consensus is optional for Hyperledger Fabric depending application specific requirements various algorithms can be used hence no mining is required.

The most essential distinction between Hyperledger and Ethereum is the intent of both technologies is designed for (see Fig. 6). Ethereum runs the Smart Contracts on the EVM for applications that are attributed to being decentralized and are for mass consumption. On the other hand, Hyperledger leverages blockchain technology for business. It is designed to support pluggable implementations of components delivering high degrees of confidentiality, resilience, and scalability. Hyperledger has a modular architecture and provides a lot of flexibility in how you want to use it. Its extensible architecture provides futuristic solutions for enterprise blockchains.

2) Hyperledger: Hyperledger is an open source collaborative effort created to advance cross-industry blockchain technologies. It is a global collaboration, hosted by The Linux Foundation, including leaders in finance, banking, Internet of Things, supply chains, manufacturing and Technology [27].

There are currently five blockchain frameworks under the umbrella of Hyperledger, which are Fabric, Sawtooth, Iroha, Indy, and Burrow. Each of these frameworks established at different time and has unique features of its own. Among them Fabric is the earliest and most mature framework when comes to stability and popularity adoption, Sawtooth comes close to Fabric as second most mature, the rest of it still under the incubation stage.

In the exploration of Hyperledger Fabric and Hyperledger Sawtooth, there are a few key differences that are clarified in Table I which are used as references to determine which of it is best adopted for the framework in this study.

Both consensus algorithm of the framework has its own built-in consensus algorithm. Fabric is using Kafka, which is one kind of crash fault tolerant, but it does not prevent the system from reaching an agreement in the case of malicious or faulty nodes like Byzantine Fault Tolerant (BFT). Sawtooth is using Proof of Elapsed Time (PoET), each of the node that involved in approving the transaction is randomly given a waiting time, the node that with the shortest waiting time will be the one that approves the transaction.

Characteristics	Ethereum	Hyperledger Fabric
Founded Date	July 2015	July 2017
Currency	Ether	None
Governance	Ethereum Developers Enterprise Ethereum Alliance	Linux Foundation IBM
Consensus	POW (Proof of Work)	No-op (no consensus needed) PBFT (Practical Byzantine Fault Tolerance)
Network	Permissionless, Public or Private	Permissioned, Private

Fig. 6. Ethereum vs Hyperledger Fabric (Melissa, 2017).

TABLE. I. DIFFERENCES OF HYPERLEDGER FABRIC AND HYPERLEDGER SAWTOOTH

Differences of components	**Hyperledger Fabric**	**Hyperledger Sawtooth**
Consensus algorithm	Kafka	Proof of Elapsed Time (PoET)
Network Infrastructure	Permissioned	Permissioned or Permissionless
Governance	More restricted with Membership Service Provider (MSP) feature	More open approach on adding roles to the network
Transparency	Comes with “Channel” feature that only show related ledger (information) to relevant party in the network	All the party in the network has read access to all the ledger in the joined network

One of the unique infrastructure characteristics of Sawtooth is the network can be permissioned or permissionless. Hence Sawtooth has a flexible approach using roles and permissions. Fabric has a built-in Membership Service Provider (MSP) feature which promotes better governance than Sawtooth. Any party has to undergo MSP system for identity verification to join the network. A digital certificate will be issued to each of them for further verification to perform every transaction.

Fabric has a feature, "Channel" which serves the purpose of organizing the information in need to know basis. With Channel, only related ledger is appending on the relevant parties.

C. Related Works of Blockchain on Supply Chain

Blockchain is believed being able to solve the weak points of current supply chain management by eliminating the need for a 3PL to account for transactions and good among other data or at least reduce its role to 2PL or 1PL because all participants have access of untampered past transactions in the blockchain network. From recent literature, there are a few approaches to apply blockchain technology on the supply chain as shown as Table II.

Most of the solutions adopted on the above studies use public blockchain (i.e. Ethereum) solutions that are permissionless blockchain. Some other solution is just merely determining the feasibility or potential of applying blockchain technology on supply chain rather than provide a usable solution for the supply chain management. There is more similar research have been done but most of it also consists of the mentioned research gap hence only some of the research without showing redundant information

TABLE. II. RELATED WORKS OF BLOCKCHAIN ON SUPPLY CHAIN

Researches	Research Gap
Blockchain for securing sustainable transport contracts and supply chain transparency	• The solution is merely to explore the potential of blockchain in providing transparency to supply chain system
Blockchain Technology in Supply Chain Traceability Systems Developing a Framework for Evaluating the Applicabilit	• This solution is aimed to improve traceability of a supply chain management and neglect transparency • The solution is just to determine the applicability of blockchain in the supply chain and not providing an end solution
The Supply Chain Has No Clothes: Technology Adoption of Blockchain for Supply Chain Transparency	• The solution is merely used to predict the usefulness or feasibility or blockchain with supply chain
Blockchains Everywhere - A Use-case of Blockchains in the Pharma Supply-Chain	• modium.io built on top of Ethereum which is a public blockchain network • Data is visible to the public
Towards an Ontology-Driven Blockchain Design for Supply Chain Provenance	• Solution built on top of Ethereum public blockchain network • Data is visible to the public

III. FRAMEWORK

A. Introduction on Conceptual Framework

The conceptual framework will comprise of a supply chain system using blockchain and all the relevant parties (i.e. farmer, processor, distributor, retailer, and consumer) in the supply chain join in the network. The supply chain system will be given a code name "Prochain". Prochain is a traceable and transparent supply chain system that builds on top of the blockchain that will cover the whole process of data management, such as data input, sharing and data monitoring of every activity in the supply chain network. Prochain allows these verified parties to have write access to create a transaction with each other and gain read access for the information whenever needed. Prochain is illustrated in Fig. 7.

Prochain consists of components like identity verification, consensus algorithm, and database. Prochain will first assign a digital certificate to each of the parties that are to be invited to join the supply chain network, this means that all the parties in Fig. 7 like farmer, processor, distributor, retailer will be issued a digital certificate to each of them. It is required for identity verification when needed to have write access to create transaction to other party. Prochain could allow second the party join the network if the party within the blockchain network agreed. Second party could be a transportation company, or warehouse company to gain read access to the transaction date when needed to carry on the operations.

B. Transaction flow of Prochain

In order to make the transactions happen, the transaction data captured by Prochain will undergo a consensus algorithm before appends the new transaction (block) to the supply chain network (chain). The prerequisite is the node will verify the identity of the party that submits the transaction request.

From Fig. 8, farmer will submit a transaction request to be appended to the blockchain database. Prochain will authenticate the farmer before allowing the transaction data to undergo the consensus algorithm.

Once the transaction data is signed after the consensus algorithm, it will append to the blockchain database, and Prochain will notify and show the results to farmer. Processor who is also in the same blockchain network will query the transaction information when in need. Processor will still need to be authenticated before access the transaction data.

C. Consumer

Consumer is a special party in Prochain. The way consumer gains the read access of a product can be in several ways, i.e. QR code, barcode and RFID tag. It is up to the retailer or other relevant parties who sold the product to consumer on how those parties want to disclose the information of a product to the consumer. Hence, the consumer does not need to authenticate identity to get the product information.

Based on Fig. 9, consumers request transaction information via QR code, barcode or RFID tag that will generate query through Prochain application. There is no identity authentication in the process and transaction result will straight away present to the consumer.

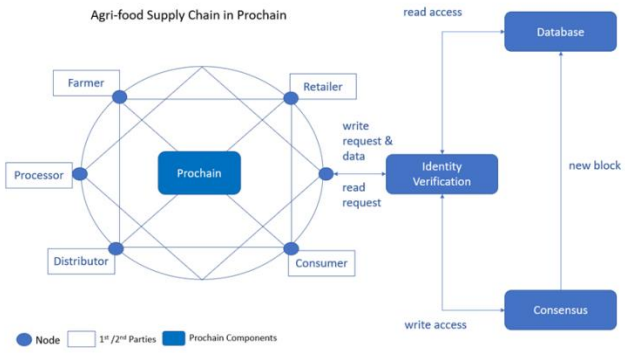


Fig. 7. Illustration of Conceptual Framework of Agri-food Supply Chain in Prochain.

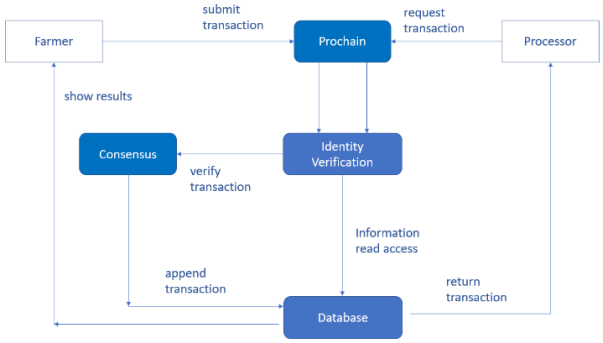


Fig. 8. Transaction Flow of from Farmer to Processor.

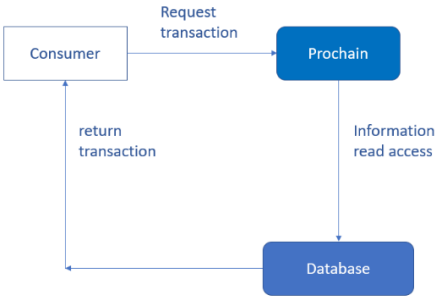


Fig. 9. Read Access of Consumer.

D. Transparency and Traceability of Prochain

In a centralized supply chain system, all the information is revolving from supply chain parties to the third-party layer service providers.

Fig. 10 depicts how are the parties in the supply chain manage the transactional data with each other via 3PL. There is a possibility that the data in such a system is mutable from other non-relevant parties that for suspicious purposes. Moreover, the third party like itself is having control over the data in the system where all the transactional data is supposed belong to the supply chain party. Hence the credibility of the data in the system has affected the transparency and traceability of the data.

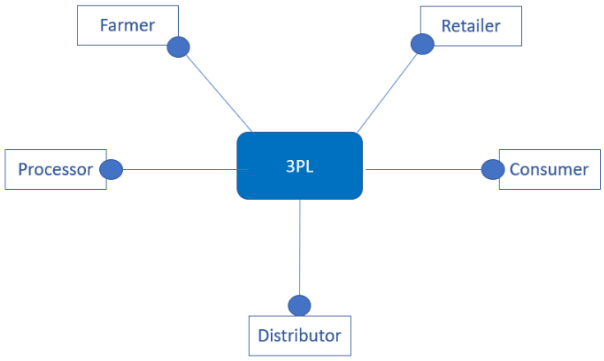


Fig. 10. Conceptual Framework for Centralized Supply Chain System.

Looking back the transaction flow of Prochain, none of the transaction has undergone to non-relevant party. The data in the blockchain is signed and encrypted by the default before appending to the blockchain database to ensure the credibility of the data. All the transaction flow is transparent among supply chain parties. Meanwhile the information in the blockchain data can be accessed in need to know basis. Hence it also maintained the confidentiality of the data. With the transparency of Prochain, the traceability of the system will work as it should be.

IV. Implementation and Analysis

A. Prochain Hyperledger Sawtooth Version

Based on the transaction flow of conceptual framework as shown in Fig. 8, it depicts the general blockchain component interact with each other when a transaction is made. Fig. 11 presents a modified version of Prochain with Hyperledger Sawtooth blockchain applied.

Based on Fig. 11, farmer or processor can have read-write access via the Sawtooth client. Either one of the access will first notice the validator in the blockchain network. The validator will first examine the access right of the user, if the user is doing a transaction submission, the validator will communicate with the transaction processor which is the backend service of Sawtooth, if the transaction is valid it will undergo the PoET consensus for the creation of the new block.

Waiting time is randomly assigning to all the nodes (validators) that in charge of validating the transaction, regardless of the power or type of hardware of that node possesses. Then the validator with the shortest wait time will have to append the new block and update the global state, which is the database that consists of all and the latest record of the ledger and being shared across all nodes of that blockchain network.

B. Prochain Hyperledger Fabric Version

Things work differently in Hyperledger Fabric of blockchain as it has different components such as Membership Service Provider (MSP), the consensus algorithm and World State.

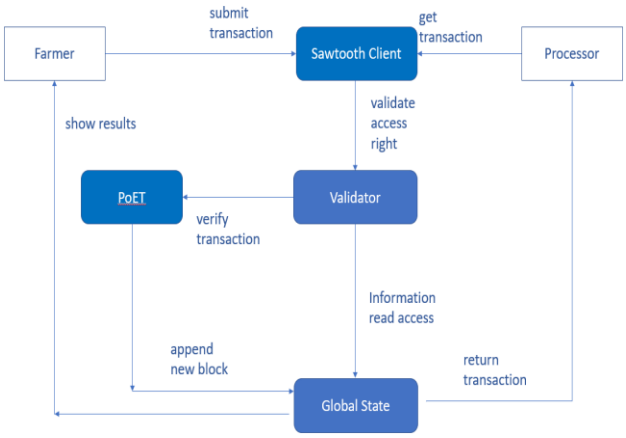


Fig. 11. Prochain Hyperledger Sawtooth Version

Based on Fig. 12, the user can have read write access via Fabric Client. The MSP will validate the access right of the user. If the user is submitting transaction it will pass to peers and undergo Kafka consensus algorithm before appending the new block. A much details visualization of the process is shown in Fig. 13.

The Peers in Fig. 12 consists of Endorsing Peer and Committing Peer. The role of Endorsing Peer is passing a set of transaction data that proposed by the user to be sorted under Kafka. There is a need of Kafka (an ordering mechanism) to sort out the set of transaction data then will pass to the Committing Peer that in charge of creating the new block into the World State. World state is a database that store all and the latest record of ledger that being shared across all nodes in that blockchain network.

C. Transparency and Traceability of Fabric and Sawtooth

Although the consensus algorithm of Hyperledger Sawtooth and Hyperledger Fabric are not the same, it will not have any impact on the transparency nor the traceability of Prochain, as it just affects the steps on approving the creation of the new block.

However, transparency and traceability of Prochain are affected by the variation of the building block of Sawtooth and Fabric. One notable of that variation is the feature of Channel provided from Fabric that allows the stakeholders in the network carry out transaction privately with a subset of other stakeholders that are not visible to another subset of stakeholders.

In Fig. 14, the data of transaction A is visible to Farmer A, Processor A and Distributor A as all of them are on the same channel. However, the data of transaction B is only visible among Farmer B and Processor A but not Farmer A and Distributor A, note that both channels co-exist in the same blockchain network.

Based on the steps of establishing the supply chain transparency and transparency in Fig. 3, a subset of data of suppliers (Farmer and Processor) and transaction data (pepper) is disclosed as shown in Fig. 15 under the agreement agreed among stakeholders in the supply chain network. Both the Sawtooth and Fabric version of Prochain is using the same data in the analysis.

Fig. 16 shows that Sawtooth can achieve 100 percent transparency and traceability whereas Fabric is a little bit more complicated. In the same channel of stakeholders, it can have the same results as Sawtooth, but stakeholder that is out of that channel has zero visibility of the data.

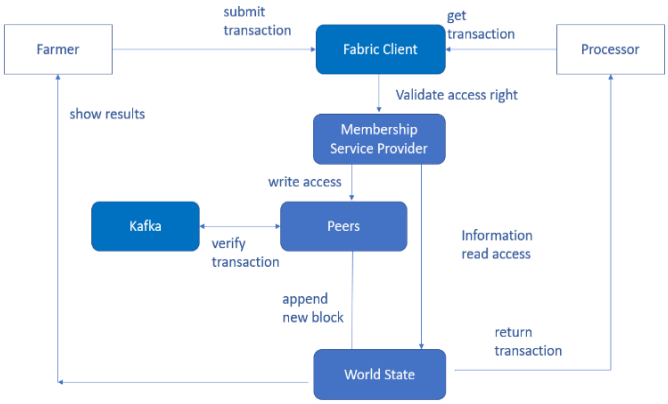


Fig. 12. Prochain Hyperledger Fabric Version.

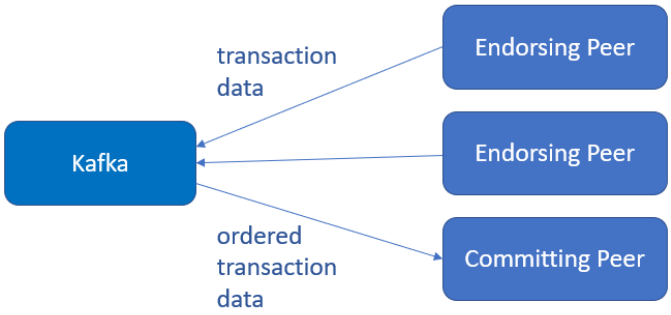


Fig. 13. Transaction flow in Ordering Service.

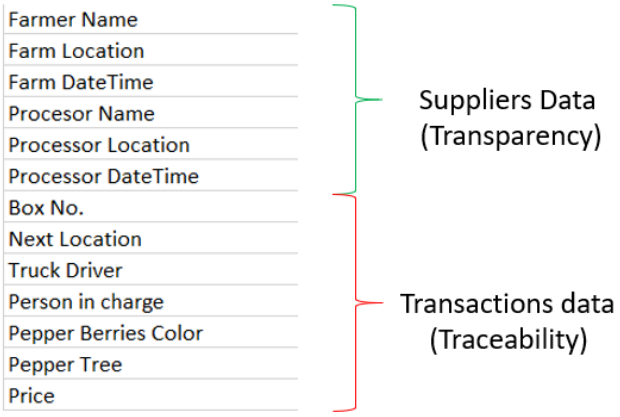


Fig. 14. 2 Channels in the Same Blockchain Network.

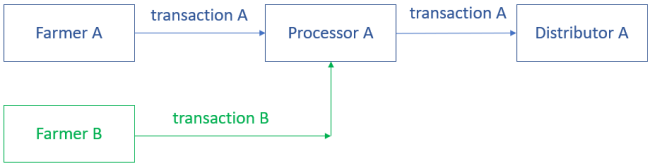


Fig. 15. Suppliers and Transactions Data from Farmer to Processor.

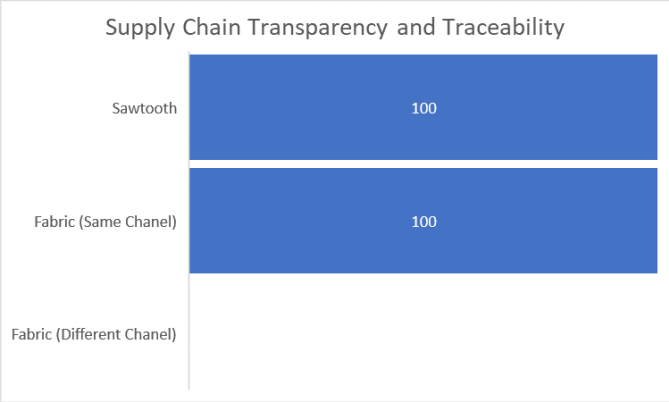


Fig. 16. Transparency and Traceability.

V. Conclusion and Future Work

This study explained how permissioned blockchain a better fit in supply chain management is compared to a permissionless blockchain. Among the permissioned blockchain Hyperledger Sawtooth and Fabric, the framework of Sawtooth has a more technological conditions that inclined with to better transparency than Fabric due to the MSP feature makes it have more central control mechanism. However, there might exist some scenario where supply chain player might favor Fabric over Sawtooth if supplier value data privacy more than data transparency.

There is still much work to be done in this research. Prochain framework can further extend with other permissioned blockchain technology like Hyperledger Indy, Burrow or other non-Hyperledger blockchains such as Quorum.

References

[1] R. B. Handfield and L. N. J. Ernest, in Supply Chain Redesign: Converting Your Supply Chain into an Integrated Value Stream, New York, Financial Prentice Hall, 2002.

[2] C. Hofstedel, H. Schepers, L. Spaans-Dijkstra, J. Trienekens and A. Beulens, in Hide or Confide: The Dilemma of Transparency, Gravenhage., Reed Business Information BV's, 2005.

[3] M. M. Aung and Y. S. Chang, "Traceability in a food supply chain: Safety and quality perspectives," Food Control, vol. 39, pp. 172-184, 2014.

[4] R. A. R. Manzini, "The new conceptual framework for food supply chain assessment," Journal of Food Engineering, vol. 115(2), pp. 251-263, 2013.

[5] M. Gupta, Blockchain For Dummies®, IBM Limited Edition, Hoboken, NJ: John Wiley & Sons, Inc, 2017.

[6] A. Koo, A. L. Othman, P. Y. Moy and P. Khor, "Agribusiness," 2017. [Online]. Available: https://gettingthedealthrough.com/area/93/jurisdiction/55/agribusiness-malaysia/.

[7] S. K. Mah, "Bright future for Sarawak pepper," 2017. [Online]. Available: Retrieved from https://www.thestar.com.my/news/nation/2017/10/25/malaysia-aims-to-be-worlds-top-supplier-of-the-premium-king-of-spices/.

[8] N. Fletcher, "llegal cross-border pepper, rubber farming uncovered," 2017. [Online]. Available: Retrieved from http://www.theborneopost.com/2017/04/03/illegal-cross-border-pepper-rubber-farming-uncovered/.

[9] R. Casado-Varaa, J. Prietoa, F. D. Prietaa and J. M. Corchado, "How blockchain improves the supply chain: case study alimentary supply chain," Procedia Computer Science, vol. 134, pp. 393-398, 2018.

[10] W. Yingli, H. Jeong Hugh and B.-D. Paul, "Understanding blockchain technology for future supply chains: a systematic literature review and research agenda," Supply Chain Management: An International Journal, vol. 24, no. 1, pp. 62-84, 2019.

[11] Chetak Logistics, "Difference between 1PL, 2PL, 3PL, 4PL & 5PL," 2015. [Online]. Available: https://www.linkedin.com/pulse/difference-between-1pl-2pl-3pl-4pl-5pl-chetak-logistics/.

[12] H. H. Lee, "How a "3-D" supply chain process system could revolutionize business," 2013. [Online]. Available: http://www.supplychainquarterly.com/topics/Strategy/20130621-how-a-3-d-supply-chain-process-system-could-revolutionize-business/.

[13] ChainThat Limited. , "Simple introduction to smart contracts on a blockchain," 2015. [Online].

[14] J. Laney, "How Blockchain can revolutionise the Supply Chain," 2018. [Online]. Available: https://jameslaney.com/blockchain-revolutionise-supply-chain/.

[15] N. Egels-Zandén and N. Hansson, "Supply chain transparency as a consumer or corporate tool: The case of Nudie Jeans Co," Journal of Consumer Policy, vol. 39(4), pp. 377-395, 2015.

[16] G. A. Akerlof, "The market for "Lemons": Quality uncertainty and the market mechanism," Quarterly Journal of Economics, vol. 84(3), pp. 488-500, 1970.

[17] D. Tapscott and D. Ticoll, The naked corporation: How the age of transparency will revolutionize business, Free Press, 2003.

[18] T. Horst, "The impact of the blockchain on the supply chain: a theory-based research framework and a call for action," Supply Chain Management: An International Journal, vol. 23, no. 6, pp. 545-559, 2018.

[19] K. Tsai, "Transparency vs. Traceability: What's the Difference?," 2018. [Online]. Available: https://www.transparency-one.com/transparency-vs-traceability-whats-the-difference/.

[20] F. Casinoa, T. K, Dasaklisb and P. Constantinos, "A Systematic Literature Review of Blockchain-based Applications: Current Status,Classification and Open Issues," Telematics and Informatics, vol. 36, pp. 55-81, 2019.

[21] T. Surasak, N. Wattanavichean, C. Preuksakarn and S. C.-H. Huang, "Thai Agriculture Products Traceability System usingBlockchain and Internet of Things," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, pp. 578-583, 2019.

[22] A. Ahmad , J. Salman , K. Shah , A. Toqeer , A. Yazed and Y. Muhammad , "A Blockchain-based Value Added Tax (VAT)System: Saudi Arabia as a Use-Case," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 708-716, 2019.

[23] N. Tien Thanh Le, Q. Nghiep Nguyen, N. Ngoc Phien, N. Duong-Trung, T. Tam Huynh, T. Phuc Nguyen and H. Xuan Son, "Assuring Non-fraudulent Transactions in Cash onDelivery by Introducing Double Smart Contracts," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 677-684, 2019.

[24] Y. Sobia , K. Muhammad Murad , T. Ramzan , B. Arslan Dawood, S. Sohaib , A. Fatima and N. Amna , "Use of Blockchain in Healthcare: A SystematicLiterature Review," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 644-653, 2019.

[25] N. Duong-Trung, X. Son Ha, T. Tai Phan, P. Nam Trieu, Q. Nghiep Nguyen, D. Pham, T. Tam Huynh and H. Trieu Le, "Multi-Sessions Mechanism for Decentralized Cashon Delivery System," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, pp. 533-560, 2019.

[26] R. Asad , K. Muhammad Murad, T. Ramzan , B. Arslan Dawood, H. Noman , A. Sultan and R. Muhammad Razeen, "Use of Blockchain in Governance: A SystematicLiterature Review," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 685-691, 2019.

[27] Hyperledger, "About Hyperledger," 2018. [Online]. Available: https://www.hyperledger.org/about..

A Key-Ordered Decisional Learning Parity with Noise (DLPN) Scheme for Public Key Encryption Scheme in Cloud Computing

Tarasvi Lakum[1]
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, A.P, INDIA

B.Thirumala Rao[*, 2]
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, A.P, INDIA

***Abstract*—The variation of decisional learning parity with noise (DLPN) named as key-Ordered DLPN based security algorithm is presented in this work. The proposed scheme uses DLPN by extending it to an even-odd-order scheme, depend on the value of probability distribution of odd and even bits for encryption, where odd and even bits are the input integer values for key generation algorithm. This states that the probability distribution of odd and even bits are ordered based on the key generation, the process of odd and even bits resolving is the solution of DLPN attacker problems, thus, the proposed scheme provides more correctness and security proof. Through the learning parity with noise (LPN), DLPN and RSA algorithms, the proposed system is evaluated, to measure the encryption time, public key and ciphertext bits.**

Keywords—LPN; DLPN; RSA; key-ordered; time

I. INTRODUCTION

To provide data security, in a confidential and in an authorized encryption, there is a need of protection to the information from an unauthenticated user [1-3]. The information required should be made available to the authorized users and to be protected from unauthorized users by creating it unavailable. Through which the availability, confidentiality and integrity of data become necessary for the security of data.

Recently, an encrypted security [4] [20-21] became an ideological research area, with a process of keeping the data in a server and encrypted form of data is communicated, in a way for the purposed users can have access and process. Cryptography broadly made in [1], [5-8] to a symmetric and a public-key. In asymmetric cryptography, public key is used for encryption process and private key is used for decryption process. The prior is more important and secure than the later for cryptography, which depends on the length of the key used [4-9] and work of cryptography made during the computations.

From the survey works, motivated with the challenges in cryptography, a variation of DLPN with two order bits has been proposed, where the keys are dependent of LPN variables and is possible to enhance the scheme by odd and even bits with newly computed bits during the process of encryption and decryption [9-15]. Increased key generation time can be reduced by increasing the process of coding, it made a big-task in the implemented method, to provide security by the process, from attacks and made secure [16-28].

A. LPN

LPN [19] computational version is an analogue of linear codes decoding through random numbers, which is an NP-complete problem. The improvement in the efficiency is made through the sparse Fourier spectrum, that is LPN solver through $2^{O(n/\log n)}$ constant term with $\mu = O(1)$, is ≤ 0.5 & independent of secret size for all the values of n, is represented as n. Here $q = poly(n)$ is the number of training samples, with this the time complexity reaches to the training samples, for $q = O(n)$, and for $q = n + O(1)$ the time complexity goes to $2^{O(n^{1-e})}$; based on these, is the main drawback of the LPN, has to be reduced.

B. DLPN

In this, the public key becomes small, having a random public and private key vectors $a \in Z_2^n$ and $s \in Z_2^n$ respectively, based on these if an attacker gets $(a, \langle a, s\rangle + e)$, where $e \leftarrow Ber_\tau$, occurring only between $0 < \tau < 1$. But from the noise rate of LPN the distribution is $0 < \tau < 0.5$, through which attacker is able to differentiate the random $r \leftarrow Z_2$ and sampling $\langle a, r\rangle$ elements, which should be solved through DLPN [29], which is a public-key encryption scheme to improve the security.

A DLPN attacker with $(a, \langle a, s\rangle + e)$ set, where $e \leftarrow Ber^{nxn}{}_\tau$, remaining parameters at DLPN based. To distinguish between the random $r \leftarrow Z^{nxn}{}_2$ and sampling $\langle a, r\rangle$ elements, with a new sample $(a, \langle a, s\rangle + e)$, taking $\tau = O(1/\sqrt{n})$ noise rate and randomly selected public and private key vectors $a \in Z_2^{nxn}$ and $s \in Z_2^{nxn}$ respectively, by assuming the DLPN probabilistic polynomial time (PPT) parameters (n, τ) is negligible.

C. Contributions

The proposed work in this paper is an approach through the tradition public key cryptography scheme RSA [26-28] and current public key cryptography schemes LPN [29] and DLPN [29], so we restrict our discussions and contributions among RSA, LPN and DLPN only. This paper provides new constructions of encryption schemes from a variant of DLPN.

*Corresponding Authors

First contribution is to introduce a DLPN variety problem with $S \leftarrow Ber^{nxn}{}_{\tau}$ within the assumptions of normal DLPN problem. As a second contribution a key-bit is constructed into vector-bit through cryptographic operations. During the cryptographic decryption process, the n dimensional vector is having the hamming weight of n/2, with the plaintext-bit as even or odd or vice-versa. The probability order of odd and even plaintext-bits is monitored to decay their exponential exceeding expectations, through by reducing the error probability. In the third contribution, the odd and even plaintext-bits are ordered in a multi-bit level based on the encryption and decryption algorithm of the public key. Unlike the previous schemes, the proposed scheme is a minimization to the LPN and DLPN problem. Proposed authentication scheme is efficient as the surveyed schemes.

II. Related Works

A. Algorithms

Many schemes have been proposed for public key cryptography. In this paper, the contributions are on RSA algorithm, LPN and DLPN.

In RSA algorithm [22-25], it becomes difficult to find the decryption key under the large integer's factors. An enhanced RSA algorithm is proposed by factorizing and deriving the key variable and considering the third prime number by making the complexity more and robust. A new factor should be replaced to increase the complexity at cryptography process to reduce the track back difficulty in the product of three prime numbers, by achieving the increased time complexity.

In LPN [19],[29], the problems available are made in to two non-trivial solving methods, one is a type of method which intends for all possible noise vectors to be intended and the other which has a sub index time complexity $2^{O(n/\log n)}$. This complexity is increased further in to $2^{O(n/\log\log n)}$ with the sampling time of $n^{1+\varepsilon}$. A further improvement in the algorithm with less running time is to be made, and there is a need of polynomial time algorithms to solve the variety of LPN problems. So there is a need of a design for LPN based cryptographic applications, through symmetric encryption in public key scheme. Here a LPN based on public key encryption with the noise ratio of $\tau \approx 1/\sqrt{n}$ is considered. However, in all the variants of, an encoding error prevails which is a non-negligible. To solve these, in this paper, a matrix LPN problem is considered to solve the encoding error problem through Damgård's scheme.

In DLPN [29], the problem is to vary between the uniform distribution over the Z value and the number of samples given by the oracle LPN. It can be formulated by an optimization solution i.e., by using random matrix A with a random column vector c over Z, to find the vector v to maximise the equations of the scheme is Av=c. This illustrates a problem of decrypting a NP-hard, which is a random linear code. To solve these variants of LPN problems, require a sub-exponential query during the sub-exponential time. The DLPN is a variant of LPN_1 problem, with a distributed secret s is a uniform random variable and through Ber_{τ}^{k}. Here noise parameter made non-constant and it depends on the value of k, through a linear number of queries which are arbitrarily polynomial and matrix version of LPN_1. A public key cryptosystem based on LPN_1 is given as $\left|Pr_{s,A,e}\left[D(A, A\cdot s\Theta e) = 1 - Pr[D(A,r)=1]\right]\right| \geq \epsilon$, where $A \leftarrow B_{\mu}^{q\times n}, s \leftarrow Z_2^n, e \leftarrow B_{\mu}^{q}$ *and* U_q is a uniform distribution over Z_2^q. The $LPN_{n,\mu,n+q}$ problem is hard and makes the problem of $Knapsack - LPN_{n,\mu,n+q}$ problem becomes hard. The DLPN problem is hard compare to LPN problem defined above, which leads to more complex results in public encryption key schemes, to make it available the design is made in black-box manner from the available DLPN problem identified, which is made for noise of $\mu = \omega(1)/\sqrt{n}$.

B. Mathematical Explanation

1) LPN: To make PKE correct, the PKE should be PKE= (KeyGen,Enc,Dec) for all the messages m ϵ M, the equation is given by: $Pr[Dec_{sk}(Enc_{pk}(m)) \neq m|(pk,sk) \leftarrow KeyGen(1^z)] \leq negl(z)$, where $negl(z)$ is negligible function. To discuss the LPN problem scenario, let us look in to the LPN oracle which is given by: $\{(v,b)|v \leftarrow Z_2^k, b =< v, s > \oplus\epsilon, \epsilon \leftarrow Ber_{\tau 0}\} \in Z_2^{k+1}$, where $s \leftarrow Z_2^k$, $\tau \in]0,\frac{1}{2}[$ is a constant noise parameter and $Ber_{\tau}$is the Bernoulli distribution with τ parameter. For the LPN search problem, depending on the distribution of As,τ the LPN oracle has the output independent random samples of $A_{s,\tau}^{LPN}$. Consider $LPN_{k,\tau}$as an instance of LPN with a secret key of size k and the noise parameter as τ. The algorithm which solves the $M(q,t,m,\theta)$ LPN search problem if $Pr\left[M^{A_{s,\tau}^{LPN}}(1^k) = s | s \leftarrow Z_2^k\right] \geq \theta$. It distinguishes between the distribution over the Z_2^{k+1} and the samples given by an LPN oracle. The search LPN problem is formulated as a optimization problem, through a random matrix A, a random column matrix b over Z_2, to find the vector s which maximises the number of equations of the system As=b.

2) DLPN: The decisional LPN problem is defined by the parameters $n \leftarrow N, \tau \epsilon R, \tau = \Theta(1/\sqrt{n})$ and randomly selected matrix $A \leftarrow Z_2^{n\times n}, S \leftarrow Z_2^{n\times n}$ as random selected matrix. The sample set of key can be obtained by the attacker in the form of (A,AS+E) by $E \leftarrow Ber_{\tau}^{n\times n}$, with the database sample (A,R), $R \leftarrow Z_2^{n\times n}$ having a non-negligible probability after getting the enough sample sets, which makes the DLPN problem to be solved. With the noise rate τ, the DLPN assumptions are defined with the probabilistic polynomial time (PPT) attacker including the parameter n which is negligible and is defined as $\tau = \Theta(1/\sqrt{n})$. The bit level encryption of DLPN is:

a) Choose $A \leftarrow Z_2^{n\times n}, S \leftarrow Ber_2^{n\times n}, E \leftarrow Ber_{\tau}^{n\times n}$, *compute* $B = AS + E$ *and* $KeyGen(1^n,\tau)$ *and returns a public key* $pk = (A,B)$ *and a private key* $sk = (S)$.

b) Choose encryption $Enc(pk,m)$, public key pk and user message $m\,\epsilon\, Z_2$, compute $c_1 = r^T A + e_1^T, c_2 = r^T B + e_2^T + mI$ and returns to a ciphertext $c = (c_1, c_2)$.

c) Choose decryption $Dec(sk,c)$, the private key sk and a ciphertext $c = (c_1, c_2)$, compute $d = c_1 \times S + c_2$ and returns m=0 for $h(d) \ll n/2$, and m=1 for $h(d) \gg n/2$.

To provide correctness of DLPN problem, define $X \sim Bin_{n,\tau}$ with an even variable with variations of $\frac{1}{2} + (1-2\tau^2)^2/2$, and

define hamming weight of each column of odd variable with variations of $n(1-(1-2\tau^2)^n/2)$, with the selected scheme of $h(d) \ll n/2$ to meet its reduced decryption error.

III. Proposed Key Encryption Scheme

A. Proposed Scheme

Proposed scheme uses two prime messages in an order with the increased size. The message noise of these two orders generate the public key (P), a variable (O) and private key (Q). P and Q are generated with O parameter considered. Random messages Even() and Odd() are required to create the prime messages. It takes a time in generation of secure key using Even() and Odd() messages and find the noise in the message and time taken is also less by dividing them in to two categories, which makes the reduction in complexity of the algorithm. The value of O is generated in a random way and these values are transmitted through a sequence of newly generated that is, O_EO as a public key. It continues with the P and Q through the regenerative O and O_EO sequences, so it becomes difficult to the attackers to enter the system which is encrypted, which helps system to improve the security. Proposed key encryption scheme algorithm is described as follows.

B. Mathematical Representation of Proposed DLPN with two order bits Algorithm

It includes:

1) kO-DLPN of kO_KeyGen(1^n,τ) takes n as integer and τ as noise rate, by choosing a random matrix $A \leftarrow Z_2^{n\times n}, S \leftarrow Ber_\tau^{n\times n}, E \leftarrow Ber_\tau^{n\times n}$.

2) kO-DLPN of kO_Enc(pk, m) is divided in to two parts.

First is Even(), where m is converted to an even-square matrix $M^{Even} \in Z_2^{n\times n}$, if m_e=1, for M^{Even}e-th column of is 1 and similarly at each entry of the e-th column is 0, e.g., m=$(0,1,1,1)^T$, then $M^{Even} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$, by choosing $E^{Even} \leftarrow Ber_\tau^{n\times n}$.

And second is Odd(), where m is converted to a odd-square matrix $M^{Odd} \in Z_2^{\overline{n\times n}}$, if m_e=0, each entry of the e-th column of M^{Odd} is 0 and similarly at each entry of the e-th column is 1, e.g., m=$(1,0,0,0)^T$, then $M^{Odd} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, by choosing $E^{Odd} \leftarrow Ber_\tau^{\overline{n\times n}}$.

3) kO-DLPN of kO_Dec(sk, c) with ciphertext $C = (C_{Even} \oplus C_1, C_2 \oplus C_{Odd})$.

C. Solution Equations for Proposed DLPN with two order bits Algorithm

It includes:

1) kO-DLPN of kO_KeyGen(1^n,τ): Compute $B = AS + E$, to return pk with key $sk = (S)$.

2) kO-DLPN of kO_Enc (pk, m) is divided in to two parts.

First is Even(), returns to ciphertext C=(C_{Even} and C_2) where the computations are C_{Even}=RA+E_{Even} and C_2=RB+E_2+M^{Even} with $E_{Even} \leftarrow Ber_\tau^{1\times n}$ and $E_2 \leftarrow Ber_\tau^{n\times n}$.

And second is Odd (), returns to ciphertext C=($C_1$and C_{Odd}) where the computations are C_1=RB+E_1+M^{Odd} and C_{Odd}=RA+E_{Odd} with $E_{Odd} \leftarrow Ber_\tau^{n}$ and $E_1 \leftarrow Ber_\tau^{n\times n}$.

3) kO-DLPN of kO_Dec(sk, c), returns $(C_1 \times C_{Even}) \times S + (C_2 \times C_{Odd})$.

IV. Application is a Web Model

Proposed scheme is applicable for secure data encryption and decryption to eliminate the noise words between the users. This scheme improves the security information transmission between multi-users utilizing web benefit as an admin middle person. It identifies the noise information in the users' message which is in encrypted unscrambled by user side and by admin encryption the noise words are recognized and identified.

Separate models are made - 1) A web-application (user and admin support) and 2) A web-server (server support). Web-application program encrypt user message received by admin through noisy message elimination during the decryption scheme and makes a web server benefit with noise-free message. The web-server receives the data processed from web applications, warns and removes user message information utilizing both sides of encryption and decryption. After admin information approval, web server restores the user message by eliminating the noise words and results the blocked user authentication for noise words presence and providing the message without any noise through the web application, which is illustrated in the below figures (Fig. 1 and 2).

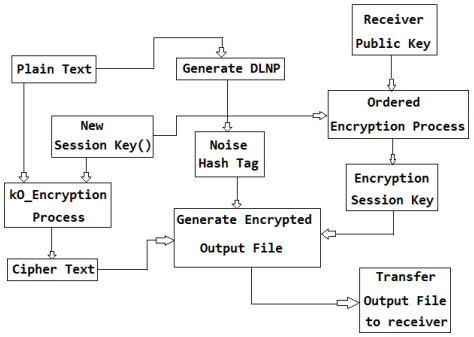


Fig. 1. Flowchart at the Sender Side.

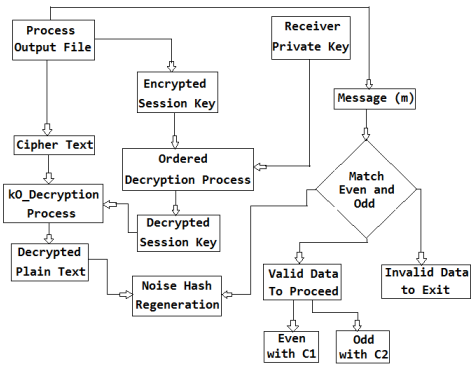


Fig. 2. Flow Chart at the Receiver Side.

V. Results and Discussions

The proposed scheme uses Client layer through HTML and CSS, Business layer through java and jdbc and Persistence through web server. All these three are integrated on Apache NetBeans platform. Here 128-bit security is used with n=29000 through 3072-bit security levels. This work is reviewed by other methods through the following parameters.

A. Key Generation Time

Table I list the comparison between proposed scheme with Damgård schemes [18] and PPKE based LPN [29] in computational efficiency.

All the above are with respective to LPN, the multiplications and additions have made the computational time to reduce. Proposed work is similar to PPKELPN but proposed work increases slightly in public key and ciphertext in both the scenarios and decryption error can be neglected. From Table II, the key generation time of proposed work is better than the reviewed works. From the experiments it is proved that proposed work key generation time is higher than RSA, Table II shows these results.

B. Encryption Time and Decryption Time

Comparing the performance of proposed work with PPKELPN , RSA(not padding) and Damgård's scheme, the proposed work illustrates better than RSA, Table III shows these results.

Comparatively the method proposed scheme is better than the surveyed works.

TABLE. I. Comparison between Proposed Work and Damgård Schemes and PPKE based LPN

Method	Public key size(bit)	Ciphertext size(bit)	Encoding error
Damgård for bit=1	$2n^2+2n$	n+1	Yes
PPKE based LPN for bit=1	$2n^2$	2n	No
Proposed for bit=1	n^2	N	No
Damgård's for bit=multi	$4n^2$	2n	Yes
PPKE based LPN for bit=multi	$2n^2$	$2n^2$	No
Proposed for bit=multi	$2n^2$	n^2+1	No

TABLE. II. Comparison between Proposed Work and RSA, ERSA, HRSA in Key Generation Time

Security level (128 bits)	Key generation time (ms)
RSA[24]	0.127
ERSA [27]	0.112
HRSA [28]	0.241
Proposed Work	0.352

TABLE. III. Comparison between Proposed Work and Reviewed in Encryption Time and Decryption Time

Security level (128 bits)	Encryption time(ms)	Decryption time (ms)
PPKE based LPN [29]	102.10	0.258
Damgård's scheme [18]	241.70	0.128
RSA(not padding) [12]	0.060	2.890
Proposed Work	99.85	0.119

VI. Conclusions

A cryptography scheme under public key through DLPN assumptions is an important research work, carrying many advantages comparatively. Due to decryption errors the existing systems are still having problems, which have to be corrected.

Through DLPN variant problem, a key-Ordered DLPN is proposed in this paper. There is a drastic change in the computing overhead of the proposed work compared to the PPKELPN, Damgård's scheme and RSA. Proposed work can withstand with the practical security like quantum attacks. A comparative result shows the proposed work gives high security.

VII. Future Scope

Further in future, through this work, design of public and private key cryptography to implement as a CCA level security, which can be made possible.

References

[1] X. Sun, B. Li, X. Lu, and F. Fang, ``CCA secure public key encryption scheme based on LWE without Gaussian sampling," in Proc. Int. Conf. Inf. Secur. Cryptol., vol. 9589, 2015, pp. 361378.

[2] Z. Liu, Y. Huang, J. Li, X. Cheng, and C. Shen, ``DivORAM: Towards a practical oblivious RAM with variable block size," Inf. Sci., vol. 447, pp. 111, Jun. 2018.

[3] C.-Z. Gao, Q. Cheng, P. He, W. Susilo, and J. Li, ``Privacy-preserving Naive Bayes classiers secure against the substitution-then-comparison attack," Inf. Sci., vol. 444, pp. 7288, May 2018.

[4] P. Li et al., ``Multi-key privacy-preserving deep learning in cloud computing," Future Generat. Comput. Syst., vol. 74, pp. 7685, Sep. 2017.

[5] G. Liu, H. Li, and L. Yang, ``A topology preserving method of evolving contours based on sparsity constraint for object segmentation," IEEE Access, vol. 5, pp. 1997119982, 2017.

[6] N. Koblitz, A. Menezes, and S. Vanstone, ``The state of elliptic curve cryptography," Des., Codes Cryptogr., vol. 19, nos. 23, pp. 173193, 2000.

[7] Z. Huang, S. Liu, X. Mao, K. Chen, and J. Li, ``Insight of the protection for data security under selective opening attacks," Inf. Sci., vols. 412413, pp. 223241, Oct. 2017.

[8] D. Boneh and A. Silverberg, ``Applications of multilinear forms to cryptography," J. Contemp. Math., vol. 324, no. 1, pp. 7190, 2003.

[9] J.-S. Coron, T. Lepoint, and M. Tibouchi, ``Practical multilinear maps over the integers," in Advances in CryptologyCRYPTO (Lecture Notes in Computer Science), vol. 8042. Berlin, Germany: Springer, 2013, pp. 476493.

[10] Y. Hu and H. Jia, ``Cryptanalysis of GGH map," Cryptol. ePrint Arch., Int. Assoc. Cryptol. Res., Tech. Rep., Feb. 2016. [Online]. Available: http://eprint.iacr.org/2015/301.pdf

[11] C. Gentry, C. Peikert, and V. Vaikuntanathan, ``Trapdoors for hard lattices and new cryptographic constructions," in Proc. 40th Annu. ACM Symp.Theory Comput., 2008, pp. 197206

[12] J. Li, Y. Li, X. Chen, P. Lee, and W. Lou, ``A hybrid cloud approach for secure authorized deduplication," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 12061216 May 2015.

[13] W. Chen, H. Lei, and K. Qi, ``Lattice-based linearly homomorphic signatures in the standard model," Theor. Comput. Sci., vol. 634, pp. 4754, Jun. 2016.

[14] S. Garg, C. Gentry, S. Halevi, M. Raykova, A. Sahai, and B. Waters, ``Candidate indistinguishability obfuscation and functional encryption for all circuits," in Proc. Annu. IEEE Symp. Found. Comput. Sci., Oct. 2013, pp. 4049.

[15] A. Blum, A. Kalai, and H. Wasserman, ``Noise-tolerant learning, the parity problem, and the statistical query model," J. ACM, vol. 50, no. 4, pp. 506519, 2003.

[16] P. Kirchner, ``Improved generalized birthday attack," Cryptol. ePrint Arch., Int. Assoc. Cryptol. Res., Tech. Rep., Jun. 2016. [Online].Available: http://eprint.iacr.org/2011/377.pdf

[17] A. Juels and A. Stephen, ``Authenticating pervasive devices with human protocols," in Proc. Annu. Int. Cryptol. Conf., vol. 3621, 2005, pp. 293308.

[18] J. Katz, J. S. Shin, and A. Smith, ``Parallel and concurrent security of the HB and HBC protocols," J. Cryptol., vol. 23, no. 3 pp. 402421, 2010.

[19] I. Damgård and S. Park, ``How practical is public-key encryption based on LPN and ring-LPN?" Cryptol. ePrint Arch., Int. Assoc. Cryptol. Res., Tech. Rep., Jun. 2016. [Online]. Available: http://eprint.iacr.org /2012/699.pdf

[20] W. Stallings, “Cryptography and network security: principles and practice”, sixth edition, 2014, ISBN: 0-13- 335469-5, pp. 9-60, 253-285.

[21] R. Rivest, A. Shamir and L. Adleman, “A method for obtaining digital signatures and public key cryptosystems”, Communications of the ACM vol. 21 (2), pp.120–126, 1978.

[22] R S Dhakar, A K Gupta and P Sharma, “Modified RSA encryption algorithm (MREA)”, 2nd ICACCT, IEEE, pp. 426-429, 2012.

[23] M.Thangavel, P. Varalakshmi, M. Murrali and K.Nithya, “An enhanced and secured RSA key generation scheme” Journal of Information Security and applications, Elsevier, vol 20, pp.3-10, 2015.

[24] F.Kong, J. Yu and L. Wu, “Security analysis of an RSA key generation algorithm with a large private key”, Springer- Verlag Berlin Heidelberg, PP-95-101, 2011.

[25] L. H. Encinas, J. M. Masqu´e and A. Q. Dios, “An algorithm to obtain a RSA modulus with a large private key”, Cryptology ePrint Archive: Report 2003/045.

[26] B.R. Ambedkar, A. Gupta,P. Gautam and S.S.Bedi, "An Efficient Method to Factorize the RSA Public Key Encryption." Communication Systems and Network Technologies (CSNT), 2011 International Conference on.

[27] R. Minni, K. Sultania and S.Mishra, “An algorithm to enhance security in RSA” , 4th ICCCNT, IEEE , pp.1-4, 2013.

[28] Jayraj Gondaliya at all.,Hybrid Security RSA Algorithm in Application of Web Service,2018 1st International Conference on Data Intelligence and Security, ©2018 IEEE, pp-149-152.

[29] Zhimin Yu at all.,A Practical Public Key Encryption Scheme Based on Learning Parity With Noise, Special Section On Information Security Solutions For Telemedicine Applications, IEEE Access, VOLUME 6, 2018, pp-31918-31923.

A Proposed Course Recommender Model based on Collaborative Filtering for Course Registration

Norazuwa Binti Salehudin[1]
Kolej Komuniti Bentong
28600 Karak, Pahang
Malaysia

Hasan Kahtan[2], Hael Al-bashiri*, [4]
Faculty of Computing
Univirsity Malaysia Pahang
Gambang, Kuantan, Malaysia

Mansoor Abdullateef Abdulgabber[3]
Faculty of Computer & Cyber Sciences
University of Prince Mugrin
Madinah, Saudi Arabia

***Abstract*—Students face issues and challenges in making decisions for course registration. Traditionally, students rely on suggestions from academic advisers prior to course registration. Therefore, students spend a considerable amount of time waiting for advisers to help them register for the right subjects. However, the number of students rises yearly, thereby increasing the responsibilities of lecturers. Moreover, academic advisers experience constraints in analysing data during consultations for course registration. Therefore, this study proposes a course recommender model based on collaborative filtering. Collaborative filtering is adopted because it provides recommendations based on students' performance in previous subjects. A dataset from the Information & Communication Technology Centre (ICT) of the University Malaysia Pahang is used to evaluate the proposed model. The evaluation is conducted based on two experiments. The first experiment is performed by calculating the difference between actual and predicted scores to verify prediction accuracy. Results show that the average of the mean absolute error of the proposed model is 0.319, which is highly accurate. The second experiment is conducted by comparing the recommendations of the proposed model with those of experts to validate the course recommendation accuracy of the proposed model. Results of the second experiment show that the proposed model has a 91.06% accuracy rate with an error rate of 8.94%. In addition, average precision is 0.68 and recall is 0.724, which are considered accurate. Therefore, the proposed model can play a vital role in assisting students and academic advisers to recommend the right courses during registration, thereby overcoming the limitations of academic advising.**

***Keywords*—Course registration; recommender system; collaborative filtering; academic advisory**

I. Introduction

Numerous students have made wrong decisions in terms of course selection during registration, which can have a negative effect on their education [1]. Therefore, academic advisers play an important role in advising students regarding such matters. Academic advisers ensure that students will make correct decisions in course registration. An academic adviser must monitor a student's academic history to provide accurate and effective recommendations. Consequently, academic advising requires a considerable amount of patience, commitment and ingenuity. Given that academic advisers lack time sufficient knowledge about students, effective advising is rarely achieved. [2-5].

Moreover, knowledge on students' backgrounds, academic plans and goals is required to provide effective recommendations. Academic advisers likewise require certain skills to analyze students' academic history to make appropriate recommendations for students' course registration. Thus, academic advising has become an added responsibility for academic staff. Academic advisers face limitations in analyzing relevant data for student course registration, and academic advising requires psychological as well as people management skills [3, 4, 6].

Moreover, academic advisers are responsible for a large number of students. Owing to the increasing number of students, academic advisers handle a substantial number of redundant cases. Therefore, they experience a tedious process of solving redundant cases and answering repetitive questions. Consequently, the academic-advising process has become a time-consuming endeavor [3, 7]. Academic advisers need a tool that can facilitate their advising tasks and responsibilities. Thus, the functions of a recommender system are appropriate to overcome such issues for several reasons. a) A recommender system is a software that provides recommendations for users based on past preferences [8-11]. b) A recommender system helps users make decisions and select appropriate options. c) A recommender system eases users' tasks by filtering and identifying their preferred options [12, 13]. d) Finally, a recommender system assists users to process data, thereby saving time by filtering and finding suitable options [14, 15].

This research aims to provide a course recommender model to overcome traditional academic-advising issues. Hence, the essence of the course recommender model is not to replace academic advising but to support students and academic advisers during course registration by providing a set of recommendations that can facilitate the implementation of their tasks and responsibilities effectively.

The rest of this paper is divided as follows. Related works on academic advising are discussed in Section 2, and Section 3 describes the proposed course recommender model based on collaborative filtering. Section 4 explains the evaluation process and results, and Section 5 concludes the paper and discusses limitations and future work.

II. Related Works

Automating the traditional academic-advising process is necessary to help students enroll in the right subjects during registration. Similarly, smart academic advising can reduce the

*Corresponding Author

workload of academic advisers in terms of time and effort in providing guidance to a large number of students. Several advising systems have been proposed to personalize and thus facilitate the registration process.

Academic advising plays a crucial role in the development of education to achieve a university's vision and mission [16]. Academic advisers are lecturers who use knowledge and experience to advise students regarding academic requirements. Specifically, academic advisers assist students in determining their study plan. Hence, academic advising is initiated during the early stages of education [17].

The role of academic advisers is to ensure that students make the right decisions during course registration. Advisers are responsible for advising students based on the students' abilities. Academic advisers must have knowledge of a student's academic history to provide accurate and effective recommendations. Hence, they require a considerable amount of patience, commitment and ingenuity. Given the limitations of academic advisers, effective advising is rarely achieved [3-5, 18]. Moreover, knowledge of students' backgrounds, academic plans and goals is required to provide effective recommendations. Academic advisers likewise need excellent decision-making skills to analyze students' academic records. Thus, they experience limitations in analyzing relevant data for course registration. In addition, academic-advising tasks are added responsibilities for academic advisers who should possess psychological and people management skills [3, 4, 19].

The number of students increases yearly; thus, academic advisers encounter issues and challenges regarding consultation time. Moreover, academic advisers cannot become fully committed to students given their other responsibilities as lecturers [3, 5, 7, 18-20]. Academic advisers are responsible for a large number of students, and the increasing number of students generates a substantial number of redundant cases for them to handle. Therefore, they face a tedious process of solving such cases and answering the same questions repeatedly. As a result, academic advising has become a time-consuming process [3, 7].

Table I presents common academic advising issues and challenges. These issues render the advising process extremely difficult and increase the responsibilities of academic advisers. Therefore, developing a tool that can facilitate academic advising tasks and responsibilities is of vital importance.

The works of [3, 5, 7, 18-20] identified academic advising as a time-consuming process. Meanwhile, [3-5, 18] stated commitment and patience as issues and challenges in academic advising. Knowledge and experience were highlighted by [3, 4, 19], and [5, 19, 20] reported the increasing number of students as the main issue and challenge related to the process. Furthermore, redundant cases were also emphasized by [3, 7].

Various techniques have been applied by researchers to overcome these issues and challenges. For example, Mostafa, Oately [19] utilized a case-based reasoning method for an academic advising system in Egyptian educational institutions and used the historical cases of students to provide recommendations. The authors utilized a survey to evaluate their proposed system. Daramola, Emebo [18] implemented case- and rule-based reasoning methods in an expert course advisory system and employed rule and historical information to generate recommendations for students. The authors likewise adopted a survey to evaluate their proposed system. Meanwhile, Rajput [7] proposed a multilayer neural network method for an intelligent advisory system and used rule and content features to provide recommendations. Henderson and Goodridge [3] proposed a rule-based reasoning method in an intelligent web-based application for academic advising and used the rule feature to generate recommendations. Shatnawi, Althebyan [5] adopted the association rule-mining method for a smart academic advising system and employed the historical feature to provide recommendations. Furthermore, Abdelhamid, Ayoub [20] implemented an agent-based method in an intelligent academic advisor system and used the rule feature to provide recommendations.

Table II shows that rule and historical features are the most common features used in research. However, most researchers did not employ evaluation methods for their proposed systems. Thus, the present research utilizes rule and historical features in the proposed system by employing collaborative filtering techniques. Mean absolute error (MAE), precision and recall methods are adopted to evaluate prediction accuracy and the accuracy of the proposed system.

TABLE. I. ACADEMIC ADVISING ISSUES AND CHALLENGES

Issues and Challenges	[19]	[18]	[7]	[3]	[5]	[20]	[4]
Time consuming	✓	✓	✓	✓	✓	✓	
Commitment and patience		✓		✓	✓		✓
Knowledge and experience	✓			✓			✓
Increasing number of students	✓				✓	✓	
Redundant cases			✓	✓			

TABLE. II. ACADEMIC ADVISING METHODS

	Techniques	**[19]**	**[18]**	**[7]**	**[3]**	**[5]**	**[20]**	**Current research**
Features	Rule		✓	✓	✓		✓	✓
	Content			✓				
	Historical	✓	✓			✓		✓
Method	Rule-based reasoning		✓		✓			
	Case-based reasoning	✓	✓					
	Multilayer neural network			✓				
	Association rule mining					✓		
	Agent based						✓	
	Collaborative filtering							✓
Evaluation	Survey	✓	✓					
	MAE							✓
	Precision							✓
	Recall							✓

III. Course Recommender Model based on Collaborative Filtering

In this research, the collaborative filtering technique is applied to the academic advising process to provide recommendations for students during course registration. A collaborative filtering engine will search for suitable courses for students based on a program structure and transcript file. The input of the proposed model is the case data of a specific student. The case data consist of a program structure containing a course code and a transcript file with student ID, course code, points and status. The output of the course recommender model is recommended courses for the subsequent semester's registration. The proposed model, which is based on the collaborative filtering method, involves six steps, as shown in Fig. 1.

A. Gather Course Score

Course code and student ID, points and status are collected from the program structure and transcript file in the 'gather course score' step. This information is stored in the matrix, which is illustrated in Fig. 2, where the columns represent courses and the rows represent students. The values in the cells represent student scores for the courses. A cell contains no value if a student has not registered for a course in the columns.

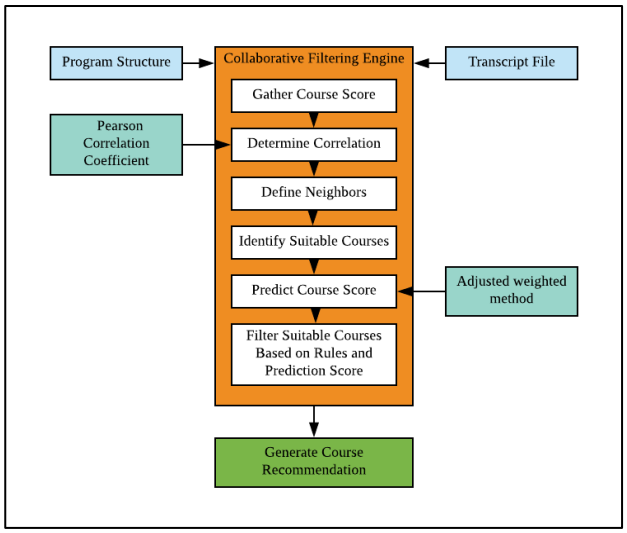


Fig. 1. Model Workflow.

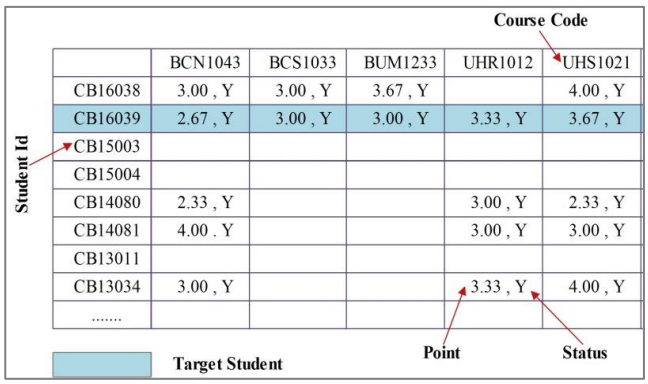


Fig. 2. Matrix for Gather Course Score Step.

B. Determine Correlation

The next step involves determining the correlation. The correlation for each student is determined by the Pearson correlation coefficient (PCC) formula [21], as shown in Equation (1). The PCC formula is described in Table III.

$$S(x,y)^{PCC} = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \overline{r_x})(r_{y,i} - \overline{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \overline{r_x})^2} \sqrt{\sum_{i \in I_{xy}} (r_{y,i} - \overline{r_y})^2}} \quad (1)$$

The correlation is determined to identify similarities between two students. If student *x* and *y* have similar scores for a course, then they have the same learning level for that course. However, if student *x* and *y* have different scores for a course, then they have different learning levels for that course. The correlation is stored in a similar matrix and illustrated in Fig. 3, where the columns and rows represent students. The values in the cells represent the correlation between two students.

C. Define Neighbours

In this step, students are sorted based on their correlation with target students, and only 30 students are defined as neighbours of a target student. This process is illustrated in Fig. 4. The neighbours are used to identify suitable courses for a target student and to predict the course scores of the target student.

D. Identify Suitable Courses

Courses that have been taken by neighbours but have not been taken by a target student are listed as suitable courses. Fig. 5 shows a sample list of suitable courses.

$$P_{x,i} = \overline{r_x} + \frac{\sum_{y \in G_{x,i}} s(x,y) * (r_{y,i} - \overline{r_y})}{\sum_{y \in G_{x,i}} s(x,y)} \quad (2)$$

TABLE. III. PCC Formula Description

Formula	Description
$S(x,y)^{PCC}$	Correlation between student *x* and student *y*
I_{xy}	Set of subjects taken by student *x* and student *y*
$\overline{r_x}$	CGPA of student *x*
$\overline{r_y}$	CGPA of student *y*
$r_{x,i}$	Score of subject *i* taken by student x
$r_{y,i}$	Score of subject *i* taken by student y

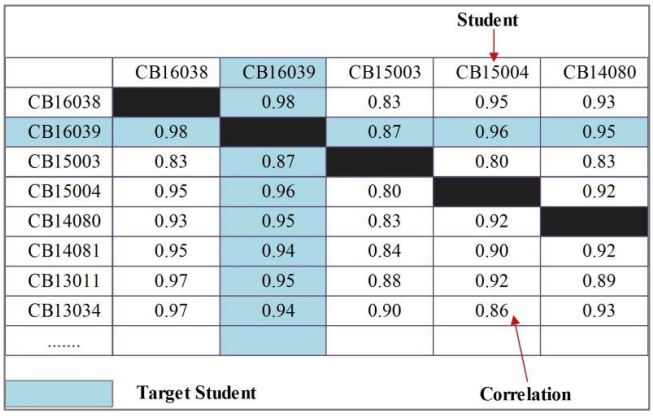


Fig. 3. Matrix for Determine Correlation Step.

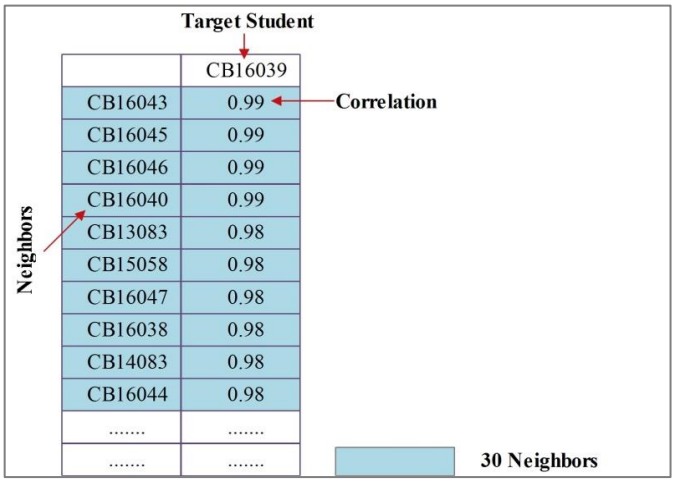


Fig. 4. Matrix for Define Neighbours Step.

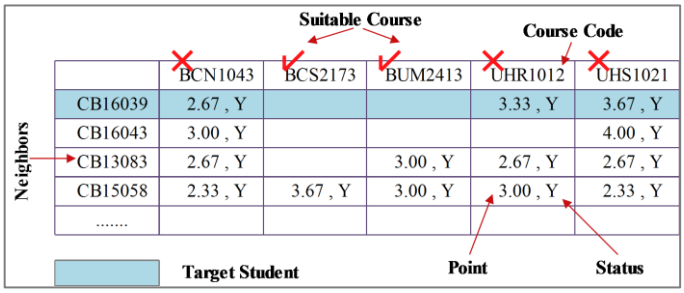


Fig. 5. Matrix for Identify Suitable Course Step.

E. Predict Course Score

In this step, the scores for each suitable course are predicted (see Fig. 6) by using a prediction formula (i.e. adjusted weighted method), as shown in Equation (2). The description of the prediction formula is provided in Table IV.

TABLE. IV. PREDICTION FORMULA DESCRIPTION

Formula	Description
$P_{x,i}$	Prediction score of course *x*
$\overline{r_x}$	CGPA of student *x*
$\overline{r_y}$	CGPA of student *y*
$r_{y,i}$	Score of the subject taken by student *y*
$G_{x,i}$	Set of students who are neighbours of student *x* and have taken subject *i*
$s(x,y)$	Correlation between student *x* and student *y*

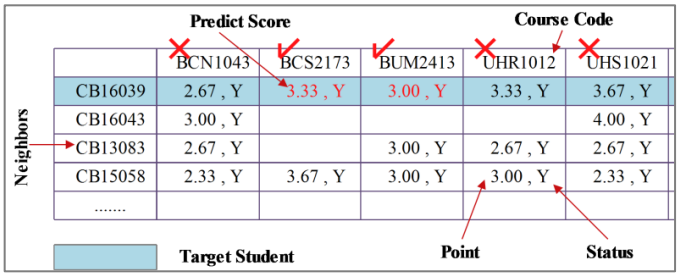


Fig. 6. Matrix for Predict Course Score Step.

F. Filter Suitable Courses based on Rules and Prediction Score

Fig. 7 illustrates the process of filtering suitable courses. Several weights have been added based on the prediction score and priority rules. The priority rules are shown in Table V. Subsequently, the courses are sorted to identify the most relevant ones for a student. Suitable courses are compared with a list of available courses based on faculty rules, as shown in Table VI. Course recommendations are generated after the filtering process.

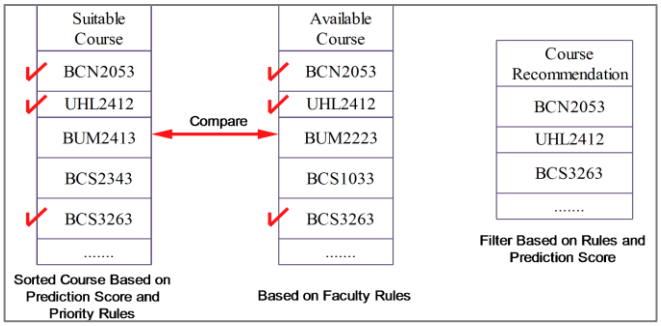


Fig. 7. Matrix for Filter Suitable Courses based on Rules and Prediction Score Step.

TABLE. V. PRIORITY RULES

Rule	Description
Rule 1	Failed courses have high priorities
Rule 2	Courses that are prerequisite to a large number of courses are given higher priority than those that are prerequisite to a small number of courses.
Rule 3	Courses in the high-level category in the curriculum are given higher priority than those in the low-level category.

TABLE. VI. FACULTY RULE

Rule	Description
Rule 1	A student must pass the prerequisites of a course before registering for that particular course.

IV. RESULTS AND DISCUSSIONS

This research used a dataset gathered from the Information & Communication Technology (ICT) Centre of the Universiti Malaysia Pahang (UMP) in Pekan, Malaysia, namely, the dbCSR database, to evaluate the proposed model. The dbCSR database consists of information of 500 students as well as the program structures and transcript files covered in the tertiary education of the Faculty of Computer Systems & Software Engineering. Moreover, the database includes 14,286 records of the scores of 500 students for 43 courses.

An evaluation of the proposed model was conducted based on two experiments to verify prediction accuracy and to validate course recommendation accuracy. Fig. 8 shows the experimental design for the model evaluation.

The prediction score generated by the proposed recommender model via collaborative filtering was compared with an actual score from a transcript file to verify prediction accuracy. Moreover, the recommendation generated by the

proposed model was compared with an expert recommendation to validate the course recommendation accuracy of the proposed model.

A. Verify Prediction Accuracy

The first experiment was performed by calculating the difference between an actual score and a prediction score. MAE was used to verify prediction accuracy based on the number of neighbours by calculating the difference between an actual score and a prediction score. A low MAE value represents high predictive accuracy. The lowest MAE value was 0.0, which meant that the prediction and actual scores had the same values. The purpose of this experiment was to determine the number of neighbours that should be selected. Prediction accuracy was improved by determining the suitable number of neighbours. Seven different numbers of neighbours were assigned, and the neighbours with the lowest MAEs were selected.

This experiment was carried out for each of the following number of neighbours: 1, 10, 20, 30, 40, 50 and 60. Fig. 9 presents the results of the experiment. The selection of few neighbours causes similarities with the original meaning to be lost, as students have similarities with more than one student [22]. Thus, the highest MAE value, that is, 0.415, was generated by selecting only one student as a neighbours. Unfortunately, accuracy decreased when a large number of neighbours was selected owing to the differences and similarities among the students [22]. Hence, the MAE value continuously increased as the number of neighbours increased. Therefore, 30 neighbours were a suitable number to select, because it generated the lowest MAE value, which was 0.319. From the first experiment, it can be concluded that prediction accuracy was high when 30 neighbours were selected. Thus, the prediction score generated by the proposed model was close to an actual score when 30 neighbours were selected.

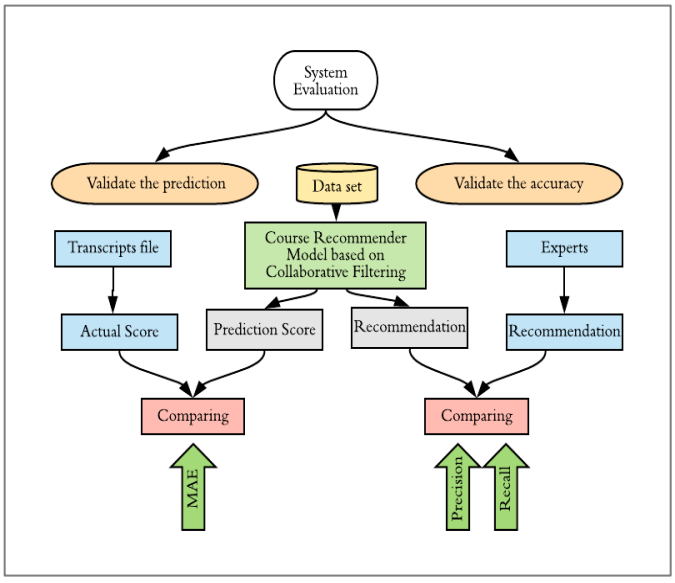


Fig. 8. Experimental Design.

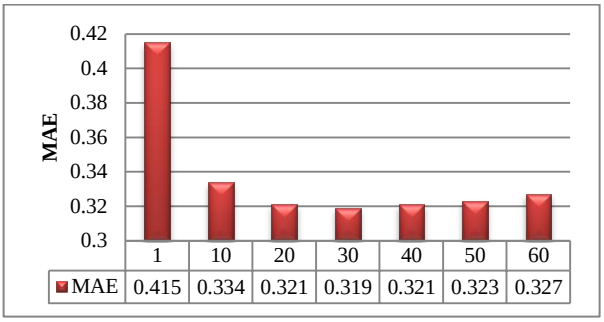


Fig. 9. MAE Results.

B. Validate the Accuracy of the Course Recommendation by the Proposed Model

The second experiment was conducted by comparing the recommendations of the proposed model with those of an expert. This experiment was time consuming and required substantial effort; thus, five different cases based on a discussion with expert lecturers were selected to represent student performance, as shown in Table VII. The purpose of this experiment was to validate the course recommendation accuracy of the proposed model.

Experiment contingency and confusion matrices were generated, as shown in Table IX, by comparing the recommendations of the proposed model with those of experts. The courses were classified as either relevant or not relevant and recommended or not recommended. Contingency and confusion matrices were adopted in this experiment owing to their capabilities to validate accuracy [23]. The proposed model made 214 correct predictions and 21 incorrect predictions in the confusion matrix. Thus, the proposed model exhibited a 91.06% accuracy rate with an error rate of 8.94%.

TABLE. VII. LIST OF CASES

Case	Student CGPA
Case 1	1.83
Case 2	2.56
Case 3	3.05
Case 4	3.55
Case 5	3.91

TABLE. VIII. EXPERIMENTAL CONTINGENCY MATRIX

	Case 1	Case 2	Case 3	Case 4	Case 5
Relevant and recommended	3	5	5	5	6
Relevant but not recommended	3	0	3	3	1
Not relevant and not recommended	39	40	37	36	38
Not relevant but recommended	2	2	2	3	2

TABLE. IX. CONFUSION MATRIX

	Recommended	Not Recommended
Relevant	24	10
Not relevant	11	190

The results of each case in Table VIII are illustrated in Fig. 10. For example, in Case 1, the proposed model recommended three relevant courses. However, compared with the recommendations of experts, the proposed model did not recommend three courses that the experts considered relevant to students. In addition, the proposed model recommended two courses that were considered not relevant by experts. Nevertheless, the prediction score and priority rules (based on the collected data) stated that the recommended courses were highly relevant compared with other courses. Thus, the proposed model recommended those courses to students. However, validity is threated, as experts have different backgrounds and the experiments generated a variety of opinions in terms of decisions for the best plans for students.

The corresponding precision and recall values were obtained from the contingency matrix in Table VIII, which is shown in Table X. High precision and recall values represent high accuracy, and 1 is the highest precision and recall value. Table X shows that average precision was 0.68 and recall was 0.724. Thus, from the second experiment, it can be concluded that the accuracy of the proposed model was high.

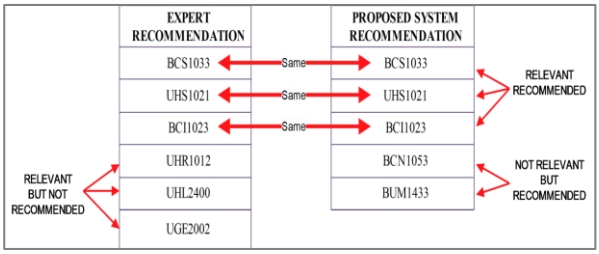


Fig. 10. Case 1 Evaluation.

TABLE. X. PRECISION AND RECALL VALUES FOR ALL CASES

	Case 1	Case 2	Case 3	Case 4	Case 5
Precision	0.6	0.71	0.71	0.63	0.75
Recall	0.5	1	0.63	0.63	0.86

V. CONCLUSION

The issues and challenges in academic advising have motivated the researchers of this study to carry out a preliminary investigation to develop a course recommender model to assist students and academic advisers during course registration. This research has achieved its second objective by developing a course recommender model using collaborative filtering for the UMP open registration process. Moreover, the developed model is capable of providing course recommendations for the subsequent semester.

The evaluation of the proposed model is conducted based on two experiments. The first experiment is performed by calculating the difference between actual and prediction scores to verify prediction accuracy. The results show that the average of the MAE of the proposed model is 0.319, which is highly accurate. Meanwhile, the second experiment is conducted by comparing the recommendations of the proposed model with those of experts to validate the course recommendation accuracy of the proposed model. The results demonstrate that the proposed model has a 91.06% accuracy rate with an error rate of 8.94%. Moreover, average precision is 0.68 and recall is 0.724, which are considered highly accurate. Therefore, the proposed model can play a vital role in assisting UMP students and academic advisers to recommend the right courses during registration. Furthermore, this research has achieved its third objective by evaluating the accuracy of the proposed course recommender model.

The proposed model offers recommendations based on students' scores. Generally, UMP accepts a new student every semester. Thus, the limitation of the proposed model involves new user cold-start problems. New user cold-start problems mean that the proposed model does not have adequate information on the historical record of a new student. Hence, it cannot identify which students have similarities with the new student. As a result, it cannot provide recommendations for the new student. Moreover, the program's structure is frequently revised, thereby adding to the proposed model's new item cold-start problem. This problem means that the proposed model cannot predict the scores for new courses until several similar students have registered in those courses. Moreover, the increasing number of students causes an increase in student scores. Thus, computation slows, thereby affecting the scalability of the proposed model.

Numerous improvements can be implemented to overcome the limitations of the proposed model. In this research, the proposed model only employs the traditional collaborative filtering approach. Thus, future studies can utilise a hybrid approach to solve the limitations of the collaborative filtering approach.

ACKNOWLEDGEMENT

This research is supported by the Department of Research and Innovation of University Malaysia Pahang under RDU190365 grant.

REFERENCES

[1] Hatch, D.K. and C.E. Garcia, Academic Advising and the Persistence Intentions of Community College Students in their First Weeks in College. The Review of Higher Education, 2017. 40(3): p. 353-390.

[2] Daramola, O., et al., Implementation of an intelligent course advisory expert system. IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2014. 5(4).

[3] Henderson, L.K. and W. Goodridge, AdviseMe: an intelligent web-based application for academic advising. International Journal of Advanced Computer Science and Applications, 2015. 6(8): p. 233-243.

[4] Laghari, M.S., Automated course advising system. International Journal of Machine Learning and Computing, 2014. 4(1): p. 47.

[5] Shatnawi, R., et al., Building A Smart Academic Advising System Using Association Rule Mining. arXiv preprint arXiv:1407.1807, 2014.

[6] Mostafa, L., et al. A Case based Reasoning System for Academic Advising in Egyptian Educational Institutions. in 2nd International Conference on Research in Science, Engineering and Technology (ICRSET'2014) March. 2014.

[7] Rajput, A.C., Intelligent counselor: an intelligent advisory system. IJSTE–Int. J. Sci. Technol. Eng, 2015. 1(9): p. 1-6.

[8] Bhatia, N. and R. Malhotra, Study on Collaborative Methods for Recommender System. Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org, 2016. 6(5).

[9] Jain, S., et al. Trends, problems and solutions of recommender system. in Computing, Communication & Automation (ICCCA), 2015 International Conference on. 2015. IEEE.

[10] Al-Bashiri, H., et al., Collaborative Filtering Recommender System: Overview and Challenges. Advanced Science Letters, 2017. 23(9): p. 9045-9049.

[11] Sharma, L. and A. Gera, A survey of recommendation system: Research challenges. International Journal of Engineering Trends and Technology (IJETT), 2013. 4(5): p. 1989-1992.

[12] Shinde, S. and M.M. Potey, Survey on Evaluation of Recommender Systems. International Journal Of Engineering And Computer Science, 2015. 4(2 February 2015).

[13] Al-Bashiri, H., et al., A Developed Collaborative Filtering Similarity Method to Improve the Accuracy of Recommendations under Data Sparsity. International Journal of Advanced Computer Science and Applications(IJACSA), 2018. 9(4): p. 135-142.

[14] Polatidis, N. and C.K. Georgiadis, A multi-level collaborative filtering method that improves recommendations. Expert Systems with Applications, 2016. 48: p. 100-110.

[15] Al-bashiri, H., et al., An improved memory-based collaborative filtering method based on the TOPSIS technique. PloS one, 2018. 13(10): p. e0204434.

[16] Pellegrino, J.L., et al., Leveraging institutional knowledge for student success: promoting academic advisors. Perspectives: Policy and Practice in Higher Education, 2015. 19(4): p. 135-141.

[17] Khalil, A. and J. Williamson, Role of Academic Advisors in the Success of Engineering Students. Universal Journal of Educational Research, 2014. 2(1): p. 73-79.

[18] Daramola, O., et al., Implementation of an intelligent course advisory expert system. IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2014. 3(5): p. 6-12.

[19] Mostafa, L., et al., A Case based Reasoning System for Academic Advising in Egyptian Educational Institutions. Proceedings of the 2nd International Conference on Research in Science, Engineering and Technology (ICRSET'2014), 2014: p. 21-22.

[20] Abdelhamid, Y., A. Ayoub, and M. Alhawiti, Agent-based intelligent academic advisor system. International Journal of Advanced Computer Technology, 2015. 4(2): p. 1-6.

[21] Al-bashiri, H., et al. Collaborative Filtering Similarity Measures: Revisiting. in Proceedings of the International Conference on Advances in Image Processing. 2017. ACM.

[22] Chen, X., et al. Trend prediction of internet public opinion based on collaborative filtering. in Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on. 2016. IEEE.

[23] Anwar, S.M., et al. A game recommender system using collaborative filtering (GAMBIT). in Applied Sciences and Technology (IBCAST), 2017 14th International Bhurban Conference on. 2017. IEEE.

Optimization of Cúk Voltage Regulator Parameters for Better Performance and Better Efficiency

Walid Emar[1]
Faculty of Engineering/Electrical Engineering Department
Isra University
Amman, 11622 Jordan

Zakariya Al-omari[2]
Faculty of Engineering/Renewable Engineering Department
Isra University
Amman, 11622, Jordan

Omar A. Saraereh[3]
Faculty of Engineering/Electrical Engineering Department
The Hashemite University Zarqa
13115, Jordan

***Abstract*—This paper discusses the harmonic distortion and voltage-current ripple minimization of a Cúk regulator based on the design optimization of its parameters using multichannel connection with uncoupled smoothing filters. The main attention is focused on the analysis and simulation of the fundamental and two-phase parallel connection of Cúk regulator with uncoupled smoothing filters. A detailed analysis has been done to show the benefits of uncoupled smoothing filters and their positive impact on balancing the energy compensation between the capacitors and inductors of the double-phase Cúk regulator as compared to conventional one. As a result of that the dc source current values do not go from positive to negative and vice versa as in the case of a fundamental connection which does not cause any saturation problem for the regulator. In general, multichannel parallel connection of Cúk regulators with uncoupled smoothing filters has ingrained benefits such as eminent current distribution characteristics, sacredness to component tolerance, reduction of parasitic effects and relief in current control complexity. Specifically, by employing double-phase connection with uncoupled smoothing filters for these regulators, overall current fluctuation can be effectively reduced by more than 25% after introducing the double-phase connection, compared to that of simple connection. Moreover, it is proved that the output voltage ripple of the double-phase connection is also reduced by more than 25% from that of the fundamental connection. Computer simulations using Simplorer 7 or Matlab and Excel have been done to validate the concepts.**

Keywords*—Cúk regulator; smoothing filters; double-phase connection; overall current ripple

I. INTRODUCTION

Due to environmental problems, the development of renewable energy has received much attention. Moreover, the green-house emissions of the generation system should be reduced according to international regulations [1]. One of the alternative energy resources is fuel cells that have attracted a great deal of attention.

To improve the performance of the fuel cell system, a fuel cell DC-DC converter with an appropriate controller is needed to regulate the energy flow and automatically adjust the output voltage of the converter [2]. The Dc-Dc chopper must be highly efficient and with less ripple.

A Cúk converter is a common type of converter which is using both buck and boost circuits converter will Change the voltage to a low or higher voltage level based on the characteristics that are designed for it [2].

Over the past few decades, the research of power electronics has focused on the development of multichannel parallel connections of DC-DC regulators in order to increase the ability of energy 2 processing and improve the reliability of the performance electronic system. The advantages of the regulator design using interlaced parallel regulators are the elimination of the ripple in both the input and output voltage and current curve waveforms to the maximum extent [3-5]. This will cause a lower value of the amplitude of the ripple and a high frequency in fluctuation in the resulting input and output waves. In addition, multichannel parallel power regulator connections reduce maintenance, increase reliability and tolerate faults. The insertion technique generally consists of shifting the control signals of several regulatory cells connected in parallel, and operating at the same switching frequency. The maximum benefits of interlacing can be achieved at a specific operating point. In general, for the Cúk regulator having one filter, the topology of one switch gives satisfactory performance in most applications, where the output voltage is greater than the input voltage but with a negative sign.

However, in a number of applications such as voltage correction circuits, energy saving and storage, and transformation of distributed power and solar systems, the performance of the Cúk regulator can be improved by implementing a Cúk with multiple channels in parallel and several smoothing filters.

The time-varying transfer function and parasitic effects of energy storage elements of a Cúk regulator leads to a positive/negative variation in the value of the source current and in a distortion in the output voltage and current. However, the current sourcing behavior of Cúk regulator helps for the use of parallel connection which can help to solve previously mentioned problems and can help paralleling many PV panels on the same bus of Common Coupling [4-7].

This paper proposes a new two-phase connection topology with uncoupled smoothing filters based on Cúk regulator with two-switch two-diode connected in parallel. Multichannel Cúk regulators operating in continuous mode of the filter current have better use of power equipments and devices, increasing power efficiency, lower losses and lower overall current

fluctuations. However, in some cases, it is also acceptable that a regulator operating in a discontinuous current mode of the filter is required. For example, multi-channel DC regulators operating in a discontinuous current mode have a lower diode loss in the case of a reverse-recovery regime and a lower loss of switching on the transistor [6-9].

Moreover, the main problem of the design between parallel channels is the simultaneous distribution of currents into individual channels. It can be pointed out that if two similar but independently controlled Cúk regulators with the same input and output voltage, the regulator with 3 a larger duty cycle can operate in the continuous mode of the filter current, while the other will work in the discontinuous current mode of the filter. In these conditions, the regulator in the continuous current mode will take over any additional loading current. So, the current sharing is very sensitive to the mismatch in the duty cycle. The design of the control circuits for equal current sharing was discussed in some literature [9-12]. Conventional multi-channel Cúk regulator is typically used for applications with input current and voltage to output current and voltage conversion. An example is that the interlaced Cúk regulator is sometimes used in applications with high power to eliminate the losses of reverse – recovery of the rectifier via the control of Cúk regulators on the boundary of continuous or discontinuous current, so that the switches are switched on, When the current over the corresponding rectifier is zero. In addition, the Cúk regulator is also used in solar systems for the production of large DC voltages from PV modules.

II. Simple Cúk Regulator

The one-channel Cúk regulator is illustrated in "Fig. 1". The circuit model of such a regulator comprises a DC input voltage source, input filter L_1, controllable switch S_1, forward capacitor C_1, diode D_1, filter L_2, output filter capacitor C_o and load resistance R. Its principle of operation explained in "Fig. 2" is currently known.

The Cúk voltage regulator provides an output voltage with a value smaller or larger than the input voltage value, but it is opposite to the input voltage. Its operation can be divided into two regimes: Regime 1 starts when the S_1 switch is on during the t_1 time, and its freewheeling diode is off. The current through filter L_1 raises linearly while simultaneously the C_1 capacitor voltage powering diode D_1 in the reverse direction which keeps it in the off time [17].

The C_1 capacitor is charged with energy through the circuit created by C_o, load and L_2. Regime 2 starts with the main switch S_2 in the off regime; the diode leads the currents of the L_1 and L_2 filters, while the C_1 capacitor is charged with a current of the L_1 filter and the input power, V_s. The energy stored in the L_2 filter is now transferred to the load. The steady voltage and current curve waveforms for $1/2 \leq k \leq 1$ are shown in "Fig. 3".

A positive property of this connection is the continuous current at the input and output of the regulator. The disadvantages of the Cúk regulator represent a large number of reactive components and high current voltage stresses on the switch, diode and capacitor C_1 and as a result of that the source current flowing through the DC source, i_s may have negative values for a short time during $0 \leq k \leq 1/2$, as shown in "Fig. 4".

To obtain the DC voltage transfer function, we use the principle that the average current through the capacitor is zero for steady-state operation. Suppose the L_1 and L_2 filters are large enough to allow their current fluctuation to be decreased. The C_1 capacitor is at steady state if

$$I_{L1}t_1 = I_s t_1 = I_{L2}t_2 \Longrightarrow I_s DT = I_{L2}(1-D)T \quad (1)$$

For a lossless regulator, the input power is equal to the output power

$$P_s = P_o \Longrightarrow V_s I_s = -V_o I_{L2} \quad (2)$$

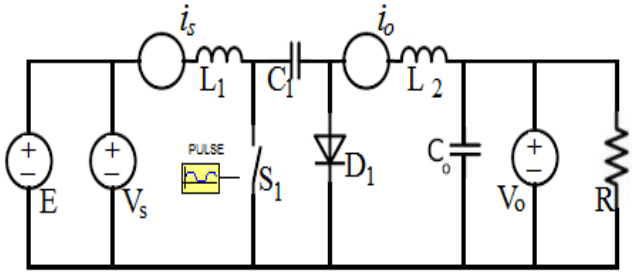


Fig. 1. Cúk Regulator Equivalent Circuit.

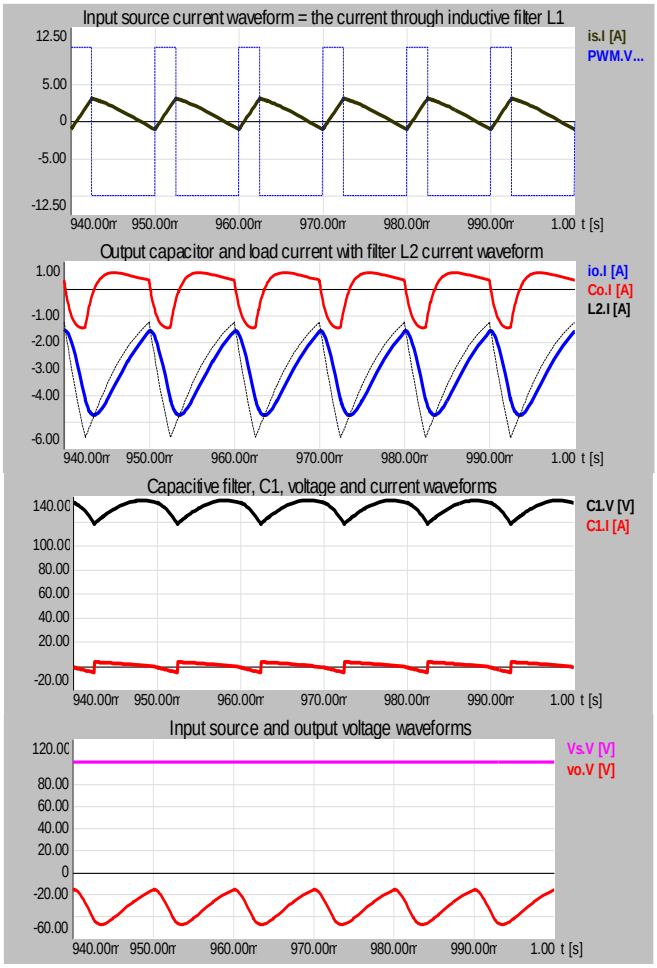


Fig. 2. Cúk Regulator with its Voltage and Current Waveforms for $0 \leq k \leq 1/2$.

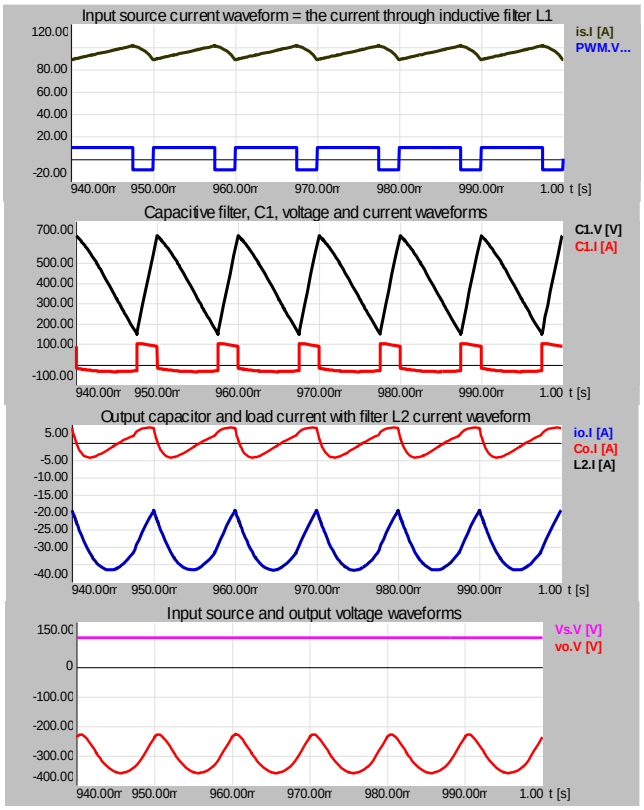


Fig. 3. Cúk Regulator with its Voltage and Current Waveforms for $1/2 \leq k \leq 1$.

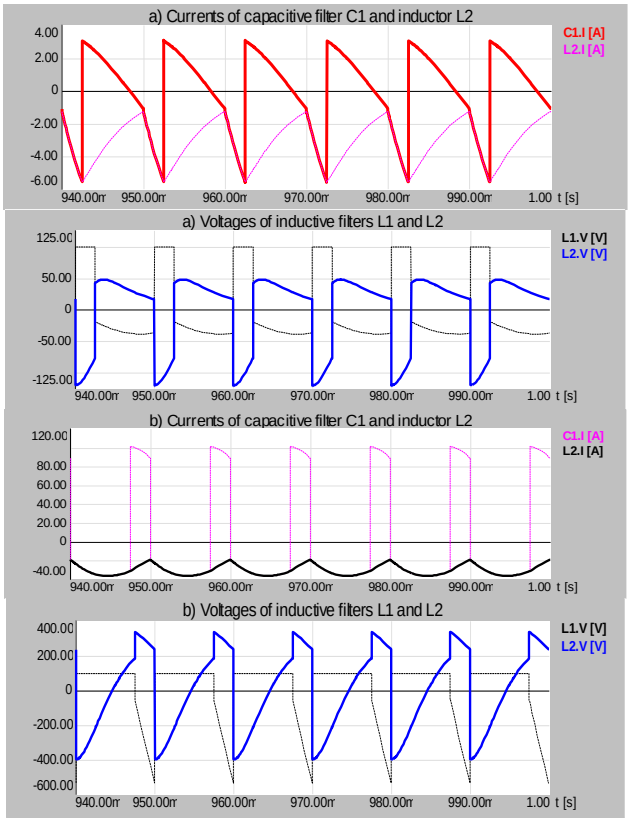


Fig. 4. Aditional Waveforms of Cúk Regulator for: a) $k \leq 1/2$ and b) $1/2 \leq k \leq 1$.

Combining these two equations, the dc voltage transfer function of the Cúk regulator is

$$V_o = -\frac{k}{1-k} V_s \tag{3}$$

When transistor S_1 is off, capacitor C_1 is charged by the input current for $t = t_2$. The average charging current of capacitor C_1 is $I_{C1} = I_S$ and the peak-to-peak ripple voltage of the capacitor C_1 is:

$$\Delta V_{c1} = \frac{1}{C_1}\int_0^{t_2} I_{c1}\,dt = \frac{1}{C_1}\int_0^{t_2} \frac{\Delta i_s}{4}\,dt = \frac{\Delta i_s\, t_2}{4\,C_1} \tag{4}$$

Substituting for $t_2 = \frac{1-k}{f}$ and for $\Delta i_s = \frac{\mathrm{kV_s}}{fL}$, Eq. (4) becomes:

$$\Delta V_{c1} = \frac{\mathrm{V_s}}{4f^2 L C_1} k(1-k) \tag{5}$$

When the main switch is on, it carries currents of filters, L_1 and L_2, which results in a high peak current through this transistor. And because the capacitor provides the energy transfer, the fluctuation of the capacitor C_1 current is also very high. Therefore, the peak-to-peak ripple of the currents flowing into both inductors i_1 and i_2 for a closed or open switch is:

$$\Delta i_1 = \frac{k\,V_s}{f\,L} \tag{6}$$

$$\Delta i_2 = \Delta i_o = \frac{V_{c1} - V_o}{f\,L} k = \frac{-V_o}{f\,L}(1-k) = \frac{k\,V_s}{f\,L} \tag{7}$$

Since the peak-to-peak ripple of the load current, Δi_o, is the same as the ripple of inductor current, Δi_2, we get for the peak-to-peak ripple of voltage $\Delta \mathrm{V_{co}}$:

$$\Delta V_{co} = \frac{1}{C_o}\int_0^{T/2} I_{co}\,dt = \frac{1}{C_o}\int_0^{T/2} \frac{\Delta i_2}{4}\,dt = \frac{\Delta i_2}{8 f C_o} = \frac{k V_s}{8\,f^2 C_o L} \tag{8}$$

III. Multichannel Cúk Regulator with Uncoupled Filters

A circuit diagram for a Conventional two-channel connection of Cúk regulator with uncoupled filters L_1, L_2 is shown in "Fig. 5". Such regulator has two legs connected in parallel, S_1, D_1 and S_2, D_2. They are supplied from a dc source $\mathrm{V_S}$ and they are switched on at different instants of time with a time displacement of T/n (T is the operating period of each switch and n is the number of channels of the regulator). The load comprises the following elements: resistor R, uncoupled smoothing filters L_1, L_2, L_3 and L_4, which are inactive electric components created from a wire coil wound around an eliminated magnetic core to have their magnetic flux within all turns of the coil winding. The magnetic field generated from them is much stronger than the one generated by the single smoothing filter coil due to the induced voltage produced at their terminals as a result of the current flowing into both of them.

The current in filter L_1 is assumed to increase during its on-time, i_1 linearly from $I_{L.a}$ to $I_{L.b}$. During the off-time of switch S1 and on-time of D_1, the forward capacitor C_1 is discharging and therefore, current i_1 decreases linearly from I_{Lb} to I_{La}, and:

$$V_s - V_{c1} = L\frac{\Delta i_1}{t_2} \Longrightarrow t_2 = L\frac{\Delta i_1}{V_s - V_{c1}} \tag{9}$$

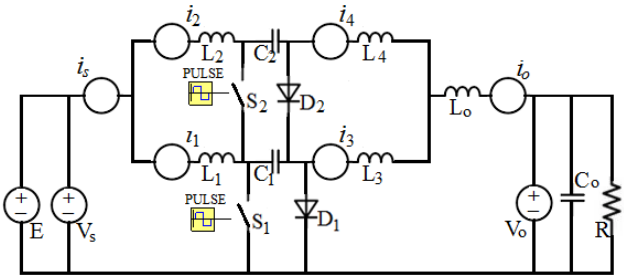


Fig. 5. Conventional Cúk Regulator with uncoupled Filters.

The peak-to-peak ripple of the phase current, Δi_1 under steady-state conditions should be same whether the switch is on or off, and it is obtained for the whole operating period of the regulator, $\boldsymbol{T}$, as shown in "Fig. 6".

Peak-to-peak ripple in regime of$1/2 \leq k \leq 1$

Traditionally, the multichannel Cúk regulator having smoothing filters magnetically uncoupled perfectly decreases the overall voltage-current peak-to-peak ripples at the input and output of the regulator. It also significantly improves as said before the power efficiency as compared to the basic connection of Cúk regulator and helps improving the balance of energy transfer between energy storage elements. Therefore, the regulator never gets into a saturation problem and its dc input current never goes from positive to negative values. These investigations are recorded for typical and linear design conditions of the filters, whereas $L_1 = L_2 = L_3 = L_4 = L$. Therefore, the averaged values of currents into both channels are divided between filters, L_1, L_2 equally by regulating the duty cycle of every switch in each leg with a time displacement of $T/2$ [17-19].

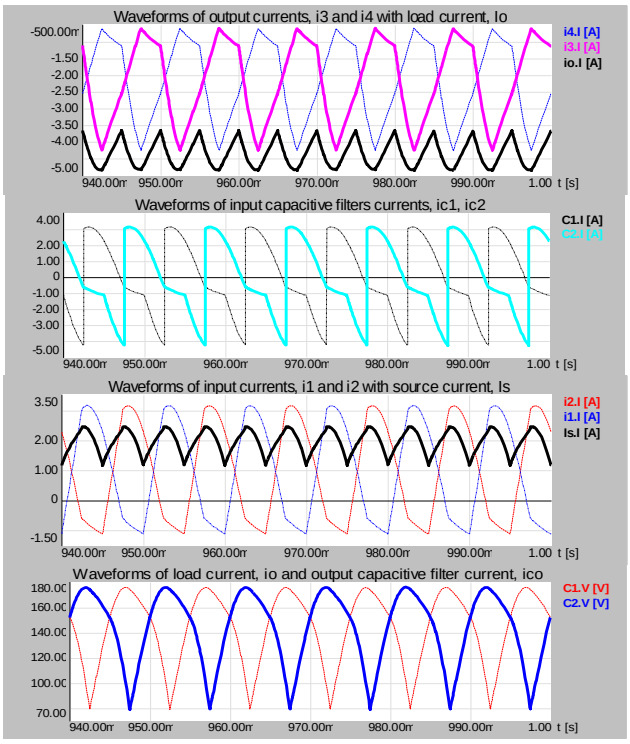


Fig. 6. Current and Volatge Steady state Waveforms of Two-Channel Connection of Cúk Regulator for 0≤k≤1/2.

Same currents but with opposite directions are flowing into filters, L_3 and L_4. So, considering a continuous operating mode of conduction, then the mathematical model representing the system when S_1 is switched on, for a time t_1 is as follows:

$$V_s = L\frac{I_{La}-I_{Lb}}{t_1} \Longrightarrow t_1 = L\frac{\Delta i_1}{V_s} \Longrightarrow \Delta i_1 = \frac{kV_s}{fL} \quad (9)$$

The current in filter L_1 is assumed to increase during its on-time, t_1 linearly from I_{La} to I_{Lb}. During the off-time of switch S_1 and on-time of D_1, the forward capacitor C_1 is discharging and therefore, current i_1 decreases linearly from I_{Lb} to I_{La}, and:

$$V_s - V_{c1} = L\frac{\Delta i_1}{t_2} \Longrightarrow t_2 = L\frac{\Delta i_1}{V_s - V_{c1}} \quad (10)$$

The peak-to-peak ripple of the phase current, Δi_1 under steady state conditions should be same whether the switch is on or off, and it is obtained for the whole operating period of the regulator, T, as follows:

$$T = t_1 + t_2 = L\frac{\Delta i_1}{V_s} + L\frac{\Delta i_1}{V_s - V_{c1}} \Longrightarrow \Delta i_1 = \frac{V_s(V_s - V_{c1})}{fL\,(2V_s - V_{c1})} \quad (11)$$

The averaged voltage across capacitor C_1, V_{C_1} is obtained from the equality of the fluctuation in the on and off-time as follows:

$$\Delta i_1 = \frac{V_s}{L}t_1 = \frac{V_s}{L}\,kT = \frac{V_s - V_{c1}}{L}t_2 = \frac{V_s - V_{c1}}{L}(1-k)T$$

$$\Longrightarrow V_{c1} = \frac{(1-2k)}{(1-k)}V_s \quad (12)$$

Similar thoughts may be applied for the second channel current, i_2. Therefore, because the smoothing filters $\boldsymbol{L_1 = L_2 = L}$ are identical, the peak-to-peak ripple of the regulator input current i_s may be determined for $\boldsymbol{0 \leq k \leq 1/2}$ as shown in "Fig. 7", when one main switch is on and the other is off as follows:

$$\Delta i_s = \frac{V_s}{L}t_1 - \frac{V_s - V_{c1}}{L}t_1 \Longrightarrow \Delta i_s = \frac{V_{c1}}{fL}k = \frac{k(1-2k)}{(1-k)fL}V_s \quad (13)$$

Concerning Δi_s in the region of $1/2 \leq k \leq 1$ as shown in "Fig. 7" when both switches are on, it may be obtained as follows:

$$\Delta i_s = 2\frac{V_s}{L}(t_1 - \frac{T}{2}) \Longrightarrow \Delta i_s = \frac{(2k-1)}{fL}V_s \quad (14)$$

Concerning the output currents, i_3 and i_4 through filters, L_3 and L_4, let's assume that current i_4 increases linearly in time i_1:

$$V_{c1} - V_o = L_4\frac{\Delta i_4}{t_1} \Longrightarrow t_1 = L\frac{\Delta i_4}{V_{c1} - V_o} \quad (15)$$

And in time t_2:

$$\boldsymbol{V_o = -L_4\frac{\Delta i_4}{t_2} \Longrightarrow t_2 = -L\frac{\Delta i_4}{V_o}} \quad (16)$$

Combining 14 and 15:

$$\boldsymbol{T = t_1 + t_2 \Longrightarrow \Delta i_4 = \frac{V_o\,(V_{c1} - V_o)}{fL\,(2V_o - V_{c1})} = \frac{V_s}{fL}k} \quad (17)$$

Equating the input power to the output power for a lossless regulator and Eq. (16) to Eq. (17), we get for the output voltage and current:

$$V_o = -\frac{kV_s}{1-k}, \qquad I_o = \frac{1-k}{k} I_s \tag{18}$$

The fluctuation of load current i_o may be determined for $0 \le k \le 1/2$ when one main switch is on and the other is off as:

$$\Delta i_o = \frac{V_{c1}-V_o}{L} t_1 + \frac{V_o}{L} t_1 \Rightarrow \Delta i_o = \frac{V_{c1}}{fL} k = \frac{(1-2k)kV_s}{(1-k)fL} \tag{19}$$

When both switches are on, $\Delta i_o \Delta i_o$ may be obtained for $1/2 > k > 1$ as:

$$\Delta i_o = 2\frac{V_{c1}-V_o}{L}(t_1 - \frac{T}{2}) \Rightarrow \Delta i_o = \frac{(2k-1)}{fL}(V_{c1} - V_o) = \frac{(2k-1)}{fL} V_s \tag{20}$$

The average value of capacitor C_o current, which flows for $T/2$, is $I_{co} = \Delta i_o/4$ and therefore, the peak-to-peak ripple voltage of capacitor C_o is:

$$\Delta v_{co} = \frac{1}{C_o}\int_0^{\frac{T}{2}} I_{co} dt = \frac{1}{C_o}\int_0^{\frac{T}{2}} \frac{\Delta i_o}{4} dt = \frac{\Delta i_o}{8fC_o} \tag{21}$$

Concerning the fluctuation of the voltage on capacitors, C_1, or C_2 is:

$$\Delta v_{c1} = \frac{1}{C_1}\int_0^{t_2} I_{c1} dt = \frac{1}{C_1}\int_0^{t_2} \frac{\Delta i_1}{4} dt = \frac{\Delta i_1 t_2}{4C_1} = \frac{\Delta i_1}{4fC_1}(1-k) \tag{22}$$

The peak-to-peak ripple, in this case, has the same value as that for the fundamental connection estimated in Eq. (5).

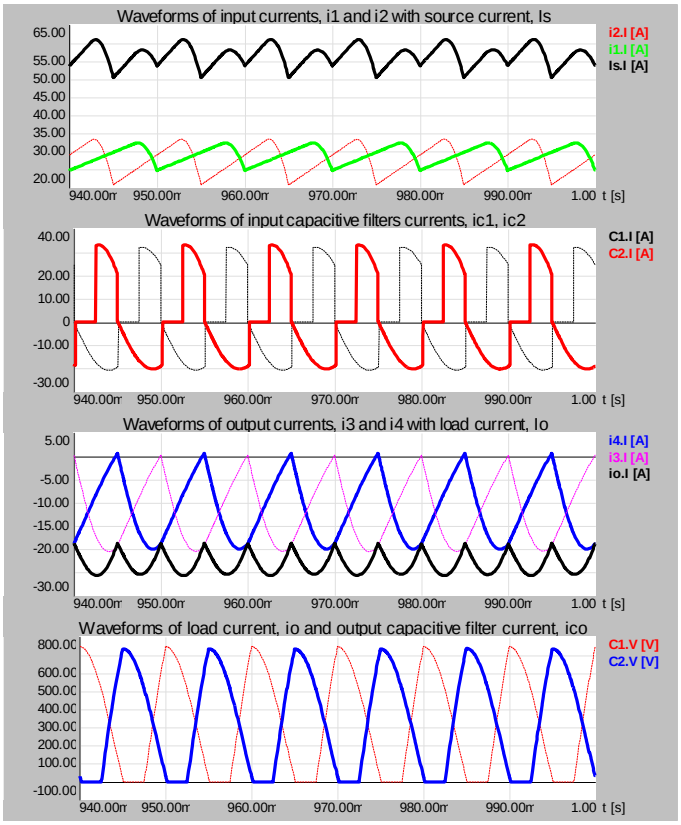


Fig. 7. Cúk Regulator with its Voltage and Current Waveforms for $1/2 \le k \le 1$.

The current and voltage fluctuations are undesired behavior which should be improved since they adversely affect the losses and heating effects in the system load above the acceptable limits and consequently decreases the expected life of the battery in which the energy created by PV systems is stored. It also makes the regulation of the regulator worse since it generates a distortion not only in the current and voltage waves at the output of the regulator but also at its input which requires further control processors and makes its control more sophisticated and the dynamic behavior of the PV supply system unstable and less efficient [13-16].

IV. Simulation Results

The peak-to-peak ripple of source and load current of the Cúk regulator with fundamental connection and double-phase connection with smoothing chokes is illustrated using previously derived equations as a function of duty time ratio, *k*, in "Fig. 8". The regulator simulation parameters for the current continuous regime are listed in "Table I" as [20-22].

The ripple of currents, i_s and i_o is directly proportional to the supply voltage V_s and inversely proportional to the chopping frequency of the regulator *f* and the filter inductance value *L*.

From "Fig. 8" and "Fig. 9" it is clear that the peak-to-peak source and load current ripple of double-phase connection of Cúk regulator with smoothing filters has smaller values than that for its fundamental connection. It should be mentioned that the number of smoothing filters for n-channel connections is n filters, but compared to the basic connection, each filter is designed to the value of $1/n$ of the source current.

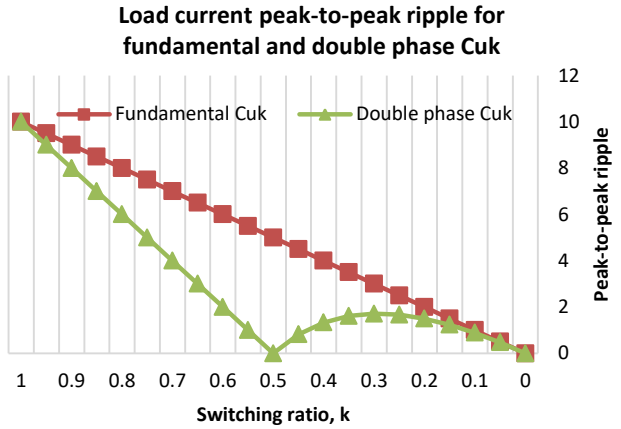


Fig. 8. Peak-to-Peak Ripple of Load Current Δ*io* Versus Duty Ratio *k*.

TABLE. I. Technical Parameters of the Simulated Regulator

Parameter	Symbol	Real value
Smoothing filters:	L_1=L_2= L_3=L_4	10mH
Regulator operating period:	T	10ms
Supply voltage	V_s	100V
Load resistance:	R	10Ω
Input capacitive filter	C_1=C_2	100μF
Output capacitive filter	C_o	100μF

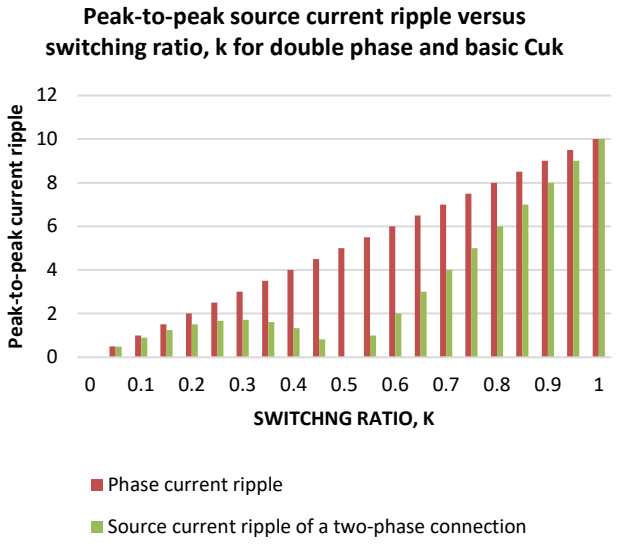


Fig. 9. Peak-to-Peak Ripple of SourceΔisversus Duty Ratio k.

In addition, "Fig. 10" shows that with an increase in switching frequency, current fluctuations in the two- channel connection are dramatically reduced compared to the underlying fundamental connection. The advantages of reducing fluctuations in the current waves due to multichannel connection of Cúk regulator result in the use of smaller filters and capacitors on the output and input of the regulator for the same requirements for the total peak-to-peak current and voltage ripple. Even the transient ripple of output voltage on the C_1 and C_2 channel capacitors can also be reduced due to smaller output coils, resulting in the use of smaller output capacitors. While the ripple on the channel capacitive filters, C_1 and C_2 does not change as shown in "Fig. 11" in blue, it is dramatically decreased on the output capacitive filter, C_o for double-phase connection (in green) as compared to that of the fundamental connection in red [9, 21-24].

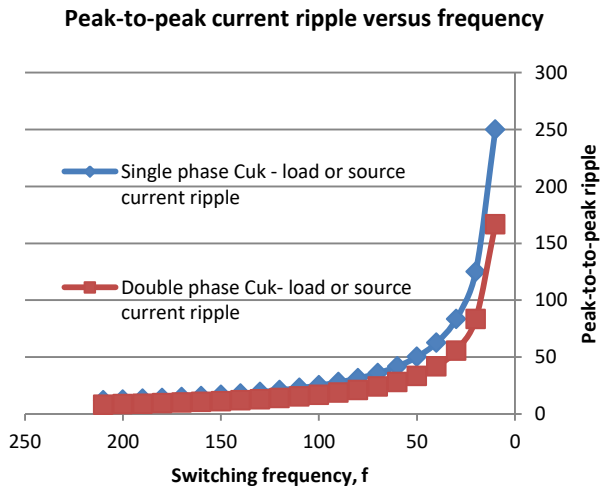


Fig. 10. Peak-to-Peak Ripple of Input forward and Output Capacitive Filter Voltages Versus Switching Ratio k.

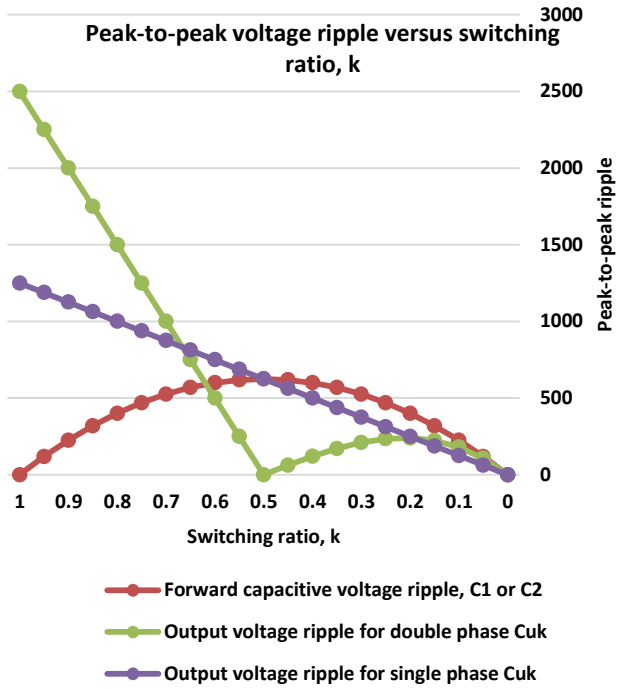


Fig. 11. Peak-to-Peak Ripple of Input forward and Output Capacitive Filter Voltages Versus Switching Ratio kfor Both Topologies of Cúk.

V. Conclusion

This paper focuses on the concept of total reduction of current and voltage ripple of the Cúk regulator using multichannel parallel connection with disconnected smoothing filters. A detailed analysis was performed, while simulations and experimental results were performed to verify the concept. In addition, it was found that the fluctuations in the channel currents will also be reduced compared to the basic connection of the regulator. This can increase efficiency and improve the system's dynamic performance. Furthermore, in the case of a single-phase Cúk regulator, the switching ratio should be carefully selected to achieve better results and to avoid parasitic effects of the capacitive and inductive filters. This is not a problem of the double-phase Cúk regulator with smoothing inductive filters that help to solve the parasitic effect of capacitive filters on the planar inductive filters.

- The simulations and experimental results obtained in this article lead to the following conclusion: Single-phase Cúk regulator may be given a priority when it is used for low power applications, for example, in telecom power supplies, smart home grid systems, and computer motherboards and power supplies where the parallel connection is not required.
- Two-phase connection with magnetically uncoupled/coupled filters is more favorable with respect to the present trend of high switching frequency, *f*. and better dynamic performance power electronic devices. These features of Cúk regulator results in a decrease of the demands to the material of smoothing inductive and capacitive filters used with it.

ACKNOWLEDGMENT

Authors acknowledge the Presidency of Isra University that supports this research study and to any person, institution or department that supported any part of the study.

REFERENCES

[1] Walid Emar, Omar A. Saraereh: Analytical and Comparative Study of Different Types of Two-Leg Chopping up Regulator, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019.

[2] Zakaria Al-Omari, A. Hamzeh, Sadeq A. Hamed, A. Sandouk, G. Aldahim, "A Mathematical Model for Minimizing Add-On Operational Cost in Electrical Power Systems Using Design of Experiments Approach", International Journal of Electrical and Computer Engineering (IJECE) Vol. 5, No. 5, October 2015, pp. 948~956.

[3] M. A. Z. A. Rashid, A. Ponniran, M. K. R. Noor, J. N. Jumadril, M. H. Yatim, A.N. Kasiran, "Optimization of PFC cuk converter parameters design for minimization of THD and voltage ripple", International Journal of Power Electronics and Drive System (IJPEDS) Vol. 10, No. 1, March 2019, pp. 514-521.

[4] Walid Emar, Omar A. Saraereh, Karim Aljebory: Three Phase Three Level Voltage Fed Inverter for a Doubly Fed Induction Motor, Transylvanian Review, Transylvanian Review: Vol XXV, No. 22, November 2017.

[5] Walid Emar: Anaslysis, Modelling and Simulation of step up Converter using Matlab-simulink and simplorer, International Journal of modeling, simulation and scientific computing, (IJMSSC) Vol. 7, Issue 3.

[6] Zoubir Belgroun, Mustapha Hatti, Salah Hanani, "Power interface efficiency evaluation for photovoltaic system used in hydrogen production", Przegląd Elektrotechniczny, ISSN 0033-2097, R. 95 NR 6/2019.

[7] M. Bildirici A. Karaarslan, "Analysis of Cuk Converter using PI and OCC Control Method", 13th International Conference on "Technical and Physical Problems of Electrical Engineering" 21-23 September 2017 Yuzuncu Yil University Van, Turkey.

[8] M. K. R. Noor, A. Ponniran, M. A. Z. A. Rashid, J. N. Jumadril, A. A. Bakar1, K. S. Muhammad, M. H. Yatim and A. N. Kasiran, "Optimization Parameter Design of SEPIC-Cuk Converter", International journal of Integrated Engineering Vol. 11, No. 1 (2019) 27-32.

[9] Walid Emar, Aqel Musbah, and I. M. M. El Emary: "Two-channel connection of switching Converters", European Journal of Scientific Research, Vol. 16, No. 1, January 2007.

[10] A. Bin Ponniran et al., "Fundamental Operation of Marx Topology for High Boost Ratio DC-DC Converter," IEEJ J. Ind. Appl., vol. 5, no. 4, pp. 329–338, 2016.

[11] A. N. K. and A. P. and M. A. H. and H. H. Hamzah, "A Study of 4-level DC-DC Boost Inverter with Passive Component Reduction Consideration," J. Phys. Conf. Ser., vol. 995, no. 1, p. 12062, 2018.

[12] A. Shahul, P. A. P. Oommen, and P. B. Cherian, "Self Lifted SEPIC-Cuk Combination Converter," vol. 3, no. 2, pp. 36–43, 2016.

[13] J. Marjani, A. Imani, A. Hekmati, and E. Afjei, "A new dual output DC-DC converter based on SEPIC and Cuk converters," in 2016 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 2016, pp. 946–950.

[14] A. Shahul, P. A. P. Oommen, and P. B. Cherian, "Self Lifted SEPIC-Cuk Combination Converter," vol. 3, no. 2, pp. 36–43, 2016.

[15] A. N. K. and A. P. and M. A. H. and H. H. Hamzah, "A Study of 4-level DC-DC Boost Inverter with Passive Component Reduction Consideration," J. Phys. Conf. Ser., vol. 995, no. 1, p. 12062, 2018.

[16] M. H. Yatim et al., "Symmetrical and asymmetrical multilevel inverter structures with reduced number of switching devices," Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, no. 1, pp. 144–151, 2018.

[17] Muhammad H. Rashid. Power Electronics, Circuits, devices, and Applications, Electrical and Computer Engineering, University of West Florida, Pearson Education International, third edition, 2004.

[18] M. Jang, M. Ciobotaru, and V.G. Agelidis, "A One-Channel Grid Connected Fuel Cell System Based on a Boost-Inverter," IEEE Trans. Power Electronics. Appl., vol. 28, no. 1, pp. 279–289, Jan. 2013.

[19] T. D. Rachmildha, Y. Haroen, A. Muqorobin, and E. Rijanto "One Channel Boost Inverter Using Hybrid Modelling Approach," IEEE Conference on Power Engineering and Renewable Energy 2012, pp. 1–6, Jul. 2012.

[20] Rosas-CaroJ, Mayo-Maldonado J. DC-DC multiplier Cúk converter with resonant switching, Electric Power, ELSEVIER, Systems ResearchVolume 119, February 2015; Pages 83–90.

[21] Chierchie, F. Paolini, E.E. Discrete-time modeling and control of a synchronous buck converter .Argentine School of Micro-Nanoelectronics, Technology and Applications, 2009. EAMTA 2009.1–2 October 2009, pp. 5 – 10. ISBN 978-1-4244-4835-7.

[22] Su, J.H.; Chen, J.J.; Wu, D.S.; "Learning feedback controller design of switching converters via Matlab/Simulink" Education, IEEE Transactions on, Volume: 45, Issue: 4, Nov. 2002 Page(s): 307 -315.

[23] N.Mohan, T. M. Undeland, and W. P. Robbins, "Power Electronics, Converters, Applications, and Design," John Wiley & Sons, 2003.

[24] KhadmunW, Subsingha W. High Voltage Gain Interleaved DC Cúk Converter Application for Photovoltaic Generation System, Energy Procedia, ELSEVIER, 10th Eco-Energy and Materials Science and Engineering Symposium, Volume 34, 2013; Pages 390–398.

AUTHORS' PROFILE

Walid Emar received his B.Sc. and his M.Sc. degree in power electronics in 1996/1997 and his Ph.D. degree in power electronics and control in 2002 from the University of West Bohemia, Czech Republic. He worked for different institutes in his field in Czech rep. as Škoda Company, Japanese air-conditioning Daikin and Fuji Koyo. Currently, He is at Isra University, Jordan as a full-time associate professor for teaching Energy Management for Master's degree students and control systems, electrical machines, in addition to power electronics and other subjects for undergraduate students. He is also engaged in research in control of power electronics and machinery control. He is also acting as a reviewer for Scopus and ISI Clarivate journals just like IJMSSC, INDERSCIENCE, ACTAPRESS.

Zakaria Al-Omari, was born in Irbid Jordan on June 3, 1966. He obtained his MSc degree (1991), in Electrical Engineering/Power from the Faculty of Electrical Engineering, Vinnytsia State Polytechnic Institute, Ukraine and his PhD degree from the Faculty of Electrical Engineering, Vinnitsa National Technical University, Ukraine in 1998. Currently he is an Associate Professor at Renewable Engineering Department at Faculty of Engineering, Isra University in Amman, Jordan. His main interests are minimizing of power system losses, renewable energy, Power Electronics, load forecasting, reliability and efficiency. He has published 15 technical papers in Journals and international conferences.

Omar A. Saraereh initially qualified as a Telecommunication Engineer 1999 from Mu'TAH University, Jordan; he then obtained a Master of Science Degree in Digital Communication Systems from England. In 2005 he completed his PhD in Electrical and Electronic Engineering/Mobile Communications from Loughborough University, England. Dr. Saraereh has Over 12 years of academic and practical experience in Electrical Engineering, Mobile Communications, Various Antennas Design, Fabrication & Measurements, Radiation Hazards and Health Effects, and Wireless Communications. Dr. Saraereh has published many papers in various international journals and conferences. Currently Dr. Saraereh is an associate professor in the Department of Electrical Engineering at the Hashemite University/Jordan.

Traditional Learning Problems of Computing Students

Saima Siraj[1], Akhtar Hussain Jalbani[2], Muhammad Ibrahim Channa[3]
Department of Information Technology
Quaid-E-Awam University of Engineering Science and Technology, Nawabshah, Pakistan

***Abstract*—This paper aims to report the traditional learning problems of computing courses students. To identify the problems a questionnaire was framed to focus on the problems and issues faced by students while interacting in a traditional learning environment. The study tested the respondent's attitude with five-point Likert Scale. The study was analyzed by using the NCSS program. Traditional learning problems were abridging by computing mean, median, mode, standard deviation and IRQ. This research highlights the problems of different computing courses particularly, problems of basic programming concepts, unable to write code, language barrier and confidence besides these, highlighted the various academic and non-academic problems. Reliability Analysis was achieved by Cronbach's Alpha and got encouraging results of an 80% reliability coefficient.**

Keywords*—Traditional learning problems; computing education; statistical analysis of questionnaire

I. Introduction

Education is knowledge that process learning, skills, values, and enlightening experiences. The rapid growth of technology, and new trends in computer technology and annually more than thousands of students getting a degree in a computer science discipline. But, the lack of knowledge and confidence in understanding the content, of course, has created the problems and confusion for Graduate students. The author tries to overcome the gap of past to current [1]. All students are not in command to learn the basic hypothesis of programming even they do not learn basic concepts of a program [2]. For every field to check the learning outcomes of the learners in different courses, especially computing the assessments is the process of judging the students, which helps students to improve their abilities and skills. There is a need for time to introduce the process of smart learning methodology for basic program ideas to enhance their approach [3].

This study explored the traditional learning issues and difficulties for undergraduate computing students.

II. Related Work

Education produces the resources, which leads to success for nations [4]. The educators try to deliver knowledge by their experiences, intelligence, style and mode of teaching. The advent of modern technologies and equipment has made education more and more easy, attractive and effective [5]. The smart devices with wireless connectivity to create ease and efficient education leads to a way of motivation and collective learning [6]. The learning is a continuous transformation process of obtaining the knowledge whether new or existing, behaviors and skills [7]. The learning is ability, whereas the skill and knowledge are acquired by the experiments and experiences. The transformation of new trends and technologies has a great impact on the learners, which helps them to enhance their ability and approach by the induction of smart devices [8]. There are many problems which are faced by the computing students when they are learning in a traditional environment. Most important problem is to increase the confidence level of students to create a perfect program, the proper practice is necessary for a student for the self-assessment to get good results [9, 10]. The students of the computer science discipline face failure to some extent. The Lack of interest in the importance of technology and course material [11]. The aim and objective of particular knowledge to examine the information or data and make a new solution, furthermore update the previous [12]. The traditional learning system works effectively but needs modification in theoretical and practical work which gives a path to learning methodology [13]. The author has reported that computer science courses and programming has been always a tough task, the results are discouraging for learners and educators respectively. The reason briefed by the author is the non-attractive traditional methods of learning. Furthermore, the author discussed the research and studies witnessed that encouraging results are achieved by incorporating the robotic and other visual environment techniques with the aim to build to deliver knowledge [14]. The prejudicial impact of computer games in our society regarding children idle time as quoted by the author. Later it was noticed that keen interest of children in computer gaming can be fruitful in creating motivation and driving force for learners to increase their interest. The author has developed a link between computer gaming and Software engineering ideas to bring innovative trends of teaching [15, 16]. The disappointing results of the student if DIT (Networking) regarding the simulation of practical work, there may be different reasons are non-achieving bright results. In order to dig- out the reason to improve and enhance the learning process by enforcing the theoretical to visual techniques. The author gives the idea of modifying and improving the laboratory environment with collaborative learning content of students to overcome the challenges faced by learners [17].

III. Research Objectives

The research study has two core objectives, which are

- To investigate the learning problems related to the traditional teaching methods in computing discipline with the help of designed questionnaire
- Data Analysis has been done with scoring matrices that provide the directions of traditional learning problems.

IV. Proposed Methodology

In order to achieve the objective of the research, a methodology with a number of steps has been designed as shown in Fig. 1; following steps are implemented as under.

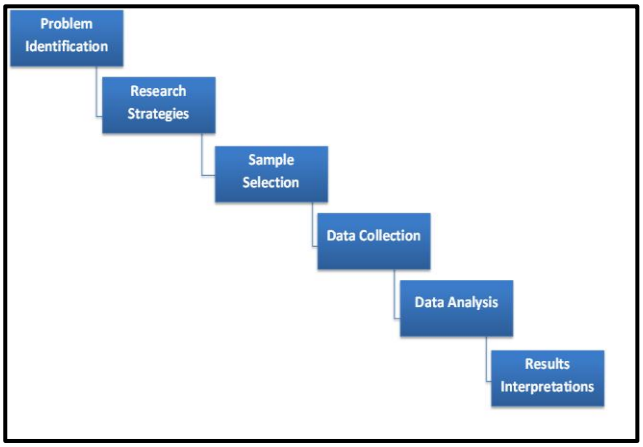


Fig. 1. Research Methodology.

- Problem Identification: This study aimed to extract the problems of the computing courses students in a traditional learning environment from the literature review as well as a pilot study.
- Research Strategies: In this step we analyze how to find out the problems of students, therefore, a questionnaire related to the problems of computing students in different subjects has been designed.
- Sampling: Final year students of the IT department are the subject of this study. The sample size of 28 has been used in this research who answered 33 questions.
- Data Collection: In this step, data has been collected using a questionnaire of 33 questions and collect the data in which students were asked to rate their problems with provided Likert scales of 5 points.
- Data Analysis: NCSS software has been used to analyze the data using descriptive statistics methods.
- Results Interpretations: Results have been interpreted using descriptive statistical analysis where central tendency mean, median, mode and standard deviation have been used. For the reliability analysis, Cronbach Alpha Coefficient method is used.

V. Data Description

The samples for this study have been collected from final year students of Information Technology department from Quaid-e-Awam University of Science and Technology. The Questionnaire has been designed to identify the problems which have been faced by the undergraduate students in traditional learning, for Gathering and assembling of information and reporting the consequence's in a systematic and well-formatted manner.

The comprehensive detail of the data description is shown in Table I. The students were asked to rate each item on a provided scale from 1-5. The score obtained were ranked as follows:

1= strongly disagree 2= Disagree 3= undecided 4=agree 5= strongly agree.

A Likert scale is one of the main frequently used methods as a psychometric tool in educational and social sciences research. In this study, a Likert scale has been used. The "Respondents" are asked with choices to show their level of agreement (from strongly disagree to strongly agree). The Participants may have a choice of five to seven or even nine pre-coded responses along with an "undecided" point of neither agree nor disagree [18,19].

Fig. 2 elucidates that from the questionnaire responses it has been revealed mostly students show discomfort in learning the computing courses. Where the x-axis shows the number of questions and the y-axis shows student's responses.

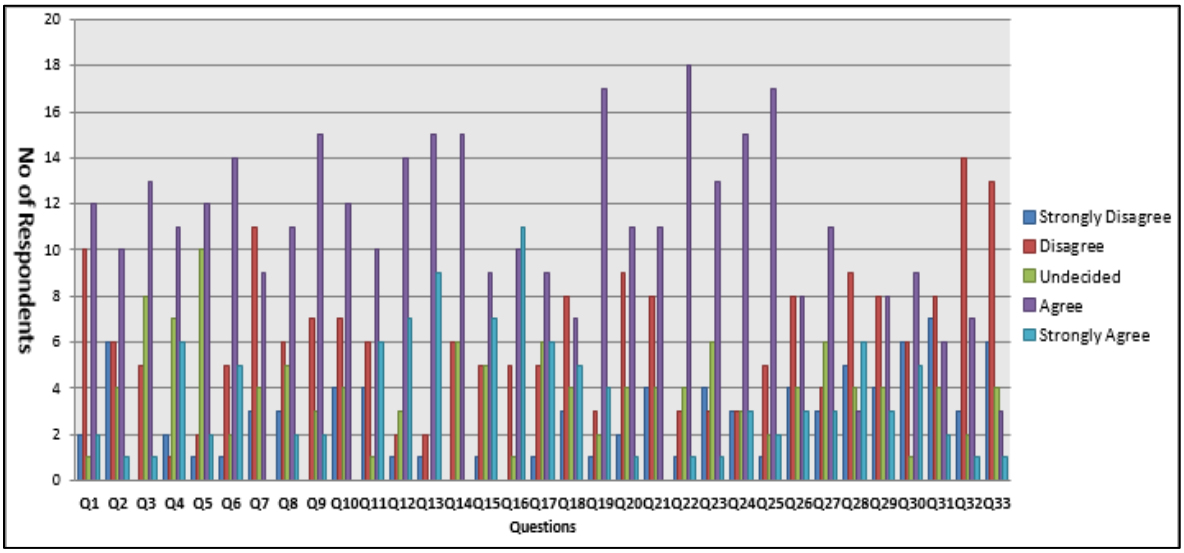


Fig. 2. Students Responses using Likert Scale.

TABLE. I. DATA DESCRIPTION

No of Students	28
No of Questions	33
Likert Scale	1= Strongly Agree 2= Agree 3= Undecided 4= Disagree 5= Strongly Disagree

TABLE. II. RESPONSES OF LIKERT

Questions options	No. of Responses	Percentages
SD	87	9.76%
D	203	22.78%
UD	128	14.36%
A	365	40.96%
SA	108	12.12%

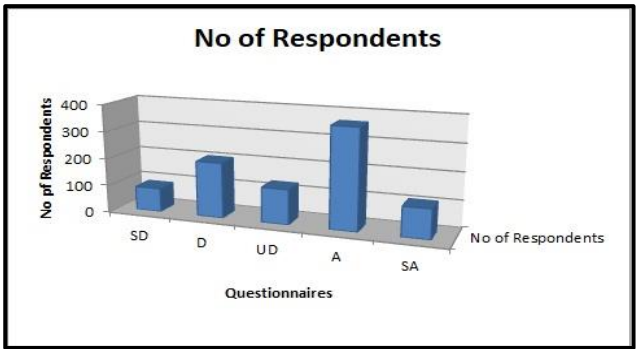


Fig. 3. Response of Questionnaire.

The data in Table II presents the responses and percentages of the students according to the Likert Scale.

This Fig. 3 presents the comparisons of responses which are provided by the respondents.

VI. DATA ANALYSIS AND EVALUATION OF THE RESULTS

For the analysis of the data, NCSS software has been used due to its high accuracy as reported [20-22]. The results of the summarized data that was collected and statistically analyzed accomplish the goal of this research study positively and reports the problems of computing students. The questions included in this study agreement with the results and justify different problems where the expectation of student performance affects and provide the answer to utilize different technologies to resolve these problems.

The data in Table III describes the detailed summary of statistics of 33 questions and all the analysis steps performed individually on every question.

A. Interpretation of the Results

The resulting Table IV summarizes the overall statistical analysis which has applied to the data and shows the result of IQR (Inter Quartile Range) which is a measure of dispersion and illustrates the scattered across the range responses. The IQR is a difference between the 75th and 25th percentile. It is mostly used to quantify scatter points. The percentiles are necessarily useful for normalized ranks. The small IQR (0-1) shows an indication of agreement about the opinion of a particular group while large IQR shows disagreements of the particular group regarding any opinion [24]. Table III shows the median value 4. The median is also called 50th percentile.

According to the statistical interpretations, the median is 4 and IQR from (0-1) means that most respondents show agreement about the particular statement. In our analysis, IQR is 1and the median is 4, therefore, we conclude that most of the students show the agreement regarding the problems highlighted in the questionnaire.

TABLE. III. SUMMARY STATISTICS OF QUESTIONNAIRE

	Variables	Mean	Median	Mode(s) Count	SD	IQR
Summary Statistics	Q1	3.107143	4	4 (13)	1.196887	2
	Q2	2.785714	3	4 (10)	1.25778	2
	Q3	3.392857	4	4 (14)	0.831745	1
	Q4	3.678571	4	4 (12)	1.090483	1
	Q5	3.5	4	4(12)	0.922958	1
	Q6	3.62963	4	4 (14)	1.114525	1
	Q7	2.785714	2.5	2 (11)	1.133893	2
	Q8	3.142857	3.5	4 (12)	1.177388	2
	Q9	3.464286	4	4 (16)	0.961563	1.75
	Q10	2.892857	3	4 (12)	113331	2
	Q11	3.357143	4	4 (10)	1.445665	2.75
	Q12	3.892857	4	4 (15)	0.99403	0.75
	Q13	4.071429	4	4 (16)	0.978607	1
	Q14	3.537143	4	4 (16)	0.82616	1
	Q15	3.642857	4	4 (9)	1.193013	2
	Q16	3.92571	4	5 (11)	1.52407	0.75
	Q17	3.535714	4	4(10)	1.137969	1
	Q18	3.142857	3	2,4 (8)	1.325373	2
	Q19	3.678571	4	4(17)	1.020297	0.75
	Q20	3.035714	3	4(12)	1.104943	2
	Q21	2.857143	3	4(12)	1.1455	2
	Q22	3.535714	4	4(18)	0.881167	1
	Q23	3.178571	4	4(14)	1.156418	1.75
	Q24	3.5	4	4(15)	1.20185	1
	Q25	3.5	4	4(17)	1	1
	Q26	3.5	3	2,4 (8)	1.333333	2
	Q27	3.321429	4	4(11)	1.21879	1.75
	Q28	2.821429	2	2,4(8)	1.441542	2
	Q29	3.107143	3	2,4 (8)	1.333333	2
	Q30	2.607143	4	4(9)	1.523624	2
	Q31	2.678571	2	2(8)	1.314852	2.75
	Q32	2.357143	2	2(14)	1.188013	2
	Q33	2.357143	2	2(13)	1.161553	1

TABLE. IV. SUMMARY STATISTICAL ANALYSIS

Overall Summary of Statistical Analysis	
Mean	3.2571749238416
Median	4
Mode	4
Standard Deviation	0.1677472183443
IRQ	1

B. Reliability Analysis

For the reliability analyses, Cronbach's Alpha method has been used to test and measure the reliability, or internal consistency, of a combined score. Cronbach's Alpha values can be obtained from the range 0 to 1, but you can get negative numbers as well. Sometimes one can get a negative number also which shows that there is something went wrong with the collected samples. The general rule of thumb regarding the reliability is that a Cronbach's Alpha of 0.70 and greater is good reliability, 0.80 and greater is better reliability, and excellent above 0.90 [23,24].

Table V interprets the comprehensive reliability report of this research study and showed the complete depiction of statistical functions applied to 33 questions.

TABLE. V. RELIABILITY REPORT

Variable	Mean	Standard Deviation	Total Mean	Total Std. Dev.	Coef Alpha
Q1	3.074074	1.206582	103.4815	13.89624	0.8057
Q2	2.777778	1.281025	103.7778	14.42843	0.8230
Q3	3.37037	0.8388705	103.1852	14.03851	0.8063
Q4	3.666667	1.1094	102.8889	14.00915	0.8081
Q5	3.444444	0.8915558	103.1111	14.1131	0.8092
Q6	3.62963	1.114525	102.9259	13.69145	0.7977
Q7	2.703704	1.067521	103.8519	14.18749	0.8133
Q8	3.111111	1.187542	103.4444	13.49454	0.7917
Q9	3.444444	0.9740216	103.1111	13.86843	0.8021
Q10	2.888889	1.154701	103.6667	13.52491	0.7924
Q11	3.296296	1.436203	103.2593	13.67646	0.8017
Q12	3.888889	1.012739	102.6667	13.88968	0.8032
Q13	4.074074	0.9971469	102.4815	14.0312	0.8076
Q14	3.333333	0.8320503	103.2222	14.11855	0.8088
Q15	3.592593	1.184151	102.963	14.14616	0.8134
Q16	4	1.1094	102.5556	13.94587	0.8061
Q17	3.518518	1.155933	103.037	13.70705	0.7987
Q18	3.111111	1.339728	103.4444	13.60807	0.7979
Q19	3.740741	0.9842058	102.4818	13.8787	0.8025
Q20	3	1.1094	103.5556	13.76543	0.8001
Q21	2.814815	1.144789	103.7407	13.84694	0.8033
Q22	3.555555	0.8915558	103	13.69447	0.7953
Q23	3.148148	1.166972	103.4074	13.46294	0.7903
Q24	3.444444	1.187542	103.1111	13.46029	0.7905
Q25	3.5185518	1.014145	103.037	13.68458	0.7963
Q26	3.925926	1.298592	103.6296	13.31901	0.7870
Q27	3.259259	1.19591	103.2963	13.903	0.8058
Q28	2.851852	1.459813	103.7037	13.94168	0.8107
Q29	2.925926	1.298695	103.6296	14.02237	0.8110
Q30	3.037037	1.505924	103.5185	14.00651	0.8135
Q31	2.555556	1.310705	104	13.641	0.7986
Q32	2.592592	1.118352	103.963	13.84386	0.8029
Q33	2.259259	1.059484	104.2963	13.6292	0.7949
Total			106.5556	14.22439	0.8078

TABLE. VI. RELIABILITY ANALYSIS

Reliability Analysis	
Cronbach's Alpha	**No of Items**
0.8078	33

Table VI shows the reliability analysis of Cronbach's Alpha that is 0.80 which means that it is better reliability about the opinion given by the particular groups regarding the problems reported in the questionnaire.

VII. CONCLUSIONS

The study was designed with an aim to predict possible problems of students while interacting with computing disciplines. For this purpose, the students of IT department of our university participated in data collection, 28 respondents' provide their responses. The provided responses were analyzed through NCSS software and achieve reliability by Cronbach's alpha and get encouraging results of 80% reliability coefficient and standard deviation of (σ) 0.16774. For showing the agreement regarding the opinion of the subjects is also validated using IQR and Median values. The overall findings of this research study have been enlightening the problems of traditional learning in computing subjects. An attempt has been made to fill the gap of learning barriers by addressing the problems which are faced by the students. There are numerous problems highlighted which adversely affected the student's level of confidence of learning outcomes and their achievements.

VIII. FUTURE RECOMMENDATIONS

In order to enhance and improve the learning of computer science subjects, with the help of a study conducted which highlighted the academic problems besides other issues. Keeping in view the problems of learners, a defined mechanism to be designed to improve the effective and efficient learning environment for a student with the utilization of dimensional encouragement of technologies and learning taxonomies.

REFERENCES

[1] Jean-Pierre, K., & Mejias, M. (2015). Toward Smart Content in Adaptive Learning Systems: Potential, Challenges, and Solutions. In Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS) (p. 80). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[2] Eckerdal, A. (2009). Novice programming students' learning of concepts and practice (Doctoral dissertation, Acta Universitatis Upsaliensis).

[3] Morrison, M., & Newman, T. S. (2001, February). A study of the impact of student background and preparedness on outcomes in CS I. In ACM SIGCSE Bulletin (Vol. 33, No. 1, pp. 179-183). ACM.

[4] Gamalel-Din, S. A. (2010). Smart e-Learning: A greater perspective; from the fourth to the fifth generation e-learning. Egyptian Informatics Journal, 11(1), 39-48.

[5] Efe, H. A., & Efe, R. (2011). Evaluating the effect of computer simulations on secondary biology instruction: An application of Bloom's taxonomy. Scientific Research and Essays, 6(10), 2137-2146.

[6] Renshaw, C. E., & Taylor, H. A. (2000). The educational effectiveness of computer-based instruction. Computers & Geosciences, 26(6), 677-682.

[7] Holt, N. J., Bremner, A., Sutherland, E., Vliek, M., Passer, M., & Smith, R. (2012). Psychology: The science of mind and behaviour. McGraw-Hill Education.

[8] Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. Journal of Interactive Media in education, 2012(3).

[9] Douce, C., Livingstone, D., & Orwell, J. (2005). Automatic test-based assessment of programming: A review. Journal on Educational Resources in Computing (JERIC), 5(3), 4.

[10] Pieterse, V. (2013, April). Automated assessment of programming assignments. In Proceedings of the 3rd Computer Science Education Research Conference on Computer Science Education Research (pp. 45-56). Open Universiteit, Heerlen.

[11] National Science Board, "Science and Engineering Indicators 2014," National Science Foundation (NSB 14-01), Arlington, VA, 2014.

[12] Haungs, M., Clark, C., Clements, J., & Janzen, D. (2012, February). Improving first-year success and retention through interest-based CS0 courses. In Proceedings of the 43rd ACM technical symposium on Computer Science Education (pp. 589-594). ACM.

[13] Aspray, W., & Bernat, A. (2000, March). Recruitment and retention of underrepresented minority graduate students in computer science. In Report on a Workshop by the Coalition to Diversity Computing.

[14] Papadakis, S., & Orfanakis, V. (2016, November). The combined use of Lego Mindstorms NXT and App Inventor for teaching novice programmers. In International Conference EduRobotics 2016 (pp. 193-204). Springer, Cham.

[15] Connolly, T. M., Stansfield, M., & Hainey, T. (2007). An application of game-based learning within software engineering. British Journal of Educational Technology, 38(3), 416-428.

[16] Macias, J. A. (2012). Enhancing project-based learning in software engineering lab teaching through an e-portfolio approach. IEEE Transactions on Education, 55(4), 502-507.

[17] Elias, M. S., & Ali, A. Z. M. (2014). Survey on the challenges faced by the lecturers in using packet tracer simulation in computer networking courses. Procedia-Social and Behavioral Sciences, 131, 11-15.

[18] Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. British Journal of Applied Science & Technology, 7(4), 396.

[19] Quick, J., & Hall, S. (2015). Part three: The quantitative approach. Journal of perioperative Practice, 25(10), 192-196.

[20] Kulkarni, D. K. (2016). Interpretation and display of research results. Indian journal of anesthesia, 60(9), 657.

[21] Subburaj, R., Gopal, G., & Kapur, P. K. (2012). A software reliability growth model for estimating debugging and the learning indices. International Journal of Performability Engineering, 8(5), 539-549.

[22] Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273-1296.

[23] Sam, A. H., Field, S. M., Collares, C. F., van der Vleuten, C. P., Wass, V. J., Melville, C., ... & Meeran, K. (2018). Very - short - answer questions: reliability, discrimination and acceptability.

[24] Krinsley, J. S., & Preiser, J. C. (2015). Time in blood glucose range 70 to 140 mg/dl> 80% is strongly associated with increased survival in non-diabetic critically ill adults. Critical care, 19(1), 179.

Parallel Platform for Supporting Stream Ciphers Over Multi-core Processors

Sally Almanasra
Faculty of Computer Studies
Arab Open University (AOU)
Riyadh, Saudi Arabia

***Abstract*—Designing secure and fast cryptographic primitives is one of the critical issues in the current era. Several domains, including Internet of Things (IoT), military and banking, require fast and secure data encryption over public channels. Most of the existing stream ciphers are designed to work sequentially and therefore not utilizing available computing power. Also, other stream ciphers are designed based on complex mathematical problems which makes these ciphers slower due to the complex computations. For this purpose, a novel parallel platform for enhancing the performance of stream ciphers is presented. The platform is designed to work efficiently over multi-core processors using multithreading techniques. The architecture of the platform relies on independent components that can operate over multiple cores available on the corresponding communication ends. Two groups of stream ciphers were considered as case studies in our experiments. The first category includes stream ciphers of a sequential design, while the second category includes parallelizable stream ciphers. Performance tests and analysis shows that the parallel platform was able to maximize the encryption throughput of the selected stream ciphers dramatically. The enhancements on the encryption throughput is relative to the constructional design of the stream ciphers. Parallelized stream ciphers (Salsa20, DSP-128, and ECSC-128) was able to achieve higher throughput compared to other sequentially designed stream ciphers.**

Keywords*—Stream ciphers; parallel computing; multithreading; cryptographic primitives; multi-core processors

I. Introduction

High-performance computing is progressively in demand in many day-to-day applications. Current computing resources present a tremendous opportunity for creating higher performance models through parallelism. The main concept of parallelism relies on allowing several tasks to be accomplished simultaneously and completed in a shorter period of time. Concurrent use of multiple processing resources is able to solve complex computational problems. A given problem is broken into smaller portions and solved concurrently using multiple computing units. To obtain the best of parallelism, scientists are focusing on faster hardware devices and processing techniques [1][2].

The multithreading technique is one processing technique that aims to create a virtual multiprocessor environment to execute multiple tasks on single processor [3][4]. The recent hardware revolution plays an important a role in improving system performance through multi-core technology. Multi-core processors are designed as a single physical processor that consists of the logical core of more than one processor.

Such processors' architecture enables the multi-core processor to run multiple tasks concurrently in order to achieve a higher performance compared to single-core processors. However, multi-core processors have a great advantage over multi-processors' architecture as a major proportion of intra-communication latency between communicating cores is minimized in multi-core processors, compared to the inter-communication latency carried out in multi-processor architectures. The architecture of multi-core processors is presented in Fig. 1.

Nowadays, a vast range of critical applications require a design of secure and high performance encryption algorithms to facilitate secure communication over public channels [5]. This is also an urgent issue for IoT solutions. In this research we aim to design a high-performance parallel platform to support stream cipher algorithms that depends on complex mathematical operations. Complex algorithms are known to perform slowly as internal operations require massive complex computations. In later sections we discuss some general concepts on parallel computing, which contribute to speeding up systems and applications.

The rest of the paper is organized as follows. Section 2 provides an overview of parallelism over multi-core processors. Section 3 presents the structure of stream ciphers. The design and structure of the parallel platform is introduced in Section 4. A number of performance tests were performed and results are presented in Section 5. A concluding remark is given in Section 6.

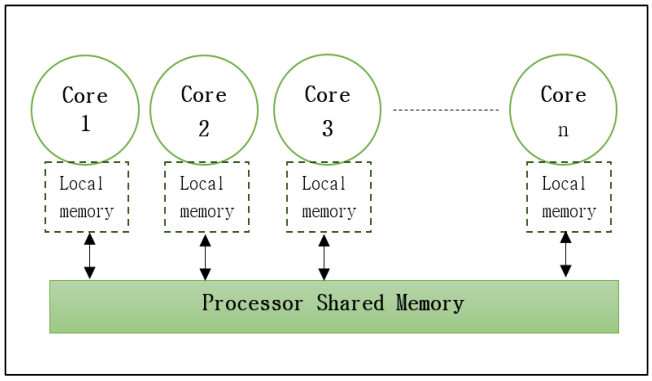


Fig. 1. The Architecture of Multi-Core Processors.

II. Parallelism over Multi-Core Processors

Multi-core computers are commonly used by individuals and enterprises. Software which are designed on a sequential base has become obstructive to performance. In order to make use of the extra cores, new algorithms must be designed in parallel bases. Such parallel designs effectively facilitate the utilization of multi-core processors.

Parallelism is usually presented in the form of threads. This technique maps independent tasks to threads at the lowest level [6]. Multithreading techniques aim to improve the performance of the running processes by allowing proper distribution of tasks among the available cores in a particular computing unit [7]. Fig. 2 illustrates the performance gained by applying multithreading techniques to a multi-core machine.

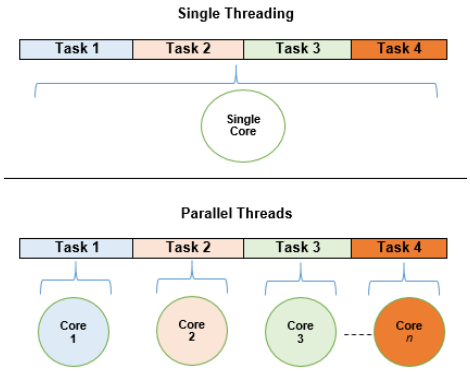


Fig. 2. Sequential Versus Parallel Execution over Multi-Core Processor.

Multithreading allows algorithms to execute several instructions per cycle, resulting in higher processor utilization and significant throughput speedup. Applying parallel techniques in cryptography has become essential for higher throughput and improved performance, especially with the current available resources. Therefore, in this study, we utilize multi-core technology with multithreading techniques to speed up the encryption process in stream ciphers, in order to provide secure and better-performing cryptosystems.

III. Symmetric-Key Encryption: Stream Ciphers

Stream ciphers are one of the cryptographic primitives that are used to secure communication over public and unsecured channels [8]. Stream cipher algorithms generate a pseudorandom keystream to encrypt a stream of plaintext, producing a stream of incomprehensible text known as ciphertext [9], as shown in Fig. 3.

Definition (Encryption): Let $k_1, k_2 \cdots, k_n \in K$ be a set of keystream in the key space K, $m_1, m_2, \cdots m_n \in M$ be a set of plaintext in the plaintext space M, and $c_1, c_2, \cdots c_n \in C$ be a set of ciphertext in the ciphertext space C. The encrypted ciphertext is generated by Equation (1):

$$E_{k_i}(m_i) = c_1, c_2, \cdots c_n \in C \qquad \forall i : 1 \leq i \leq n \qquad (1)$$

From the above definition, the encryption process of a stream cipher $E_k$is bijective for every k_i. The plaintext space and key space are typically represented in bit or byte representations.

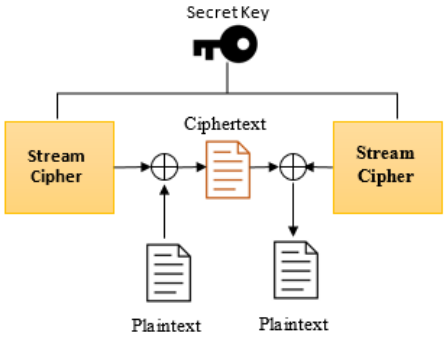


Fig. 3. Stream Cipher Algorithm.

The state of art reveals different designs of stream ciphers. The majority of existing stream ciphers are designed to run over single-core processors (e.g. RC4 [10], Sosemanuk [11]). Very few stream ciphers are designed to support parallelism. Examples of parallelized stream ciphers include Salsa20 [12] and ChaCha [13], DSP-128 [14] and ECSC-128 [15] stream ciphers.

However, the parallelized parts of these algorithms are restricted to some internal sections and do not focus on the general production of keystreams. In the next section, we introduce a parallel platform for supporting efficient parallelization of stream ciphers.

IV. Proposed Parallel Platform for Stream Ciphers

The proposed parallel platform is designed to support stream cipher operations on machines of different numbers of cores. For instance, the sender may encrypt a text on a machine with two cores, while the receiver may decrypt the text on a machine with any number of cores. The platform works flexibly regardless the number of cores on both sides. Our parallel platform is presented in Fig. 4.

The overall design of the parallel platform is divided into multiple parts to ensure maximum parallelism and a balanced workload among the available cores. In addition, the design of the platform also focuses on avoiding synchronization among the running threads for higher performance. This is possible through the use of multiple controllers in the platform.

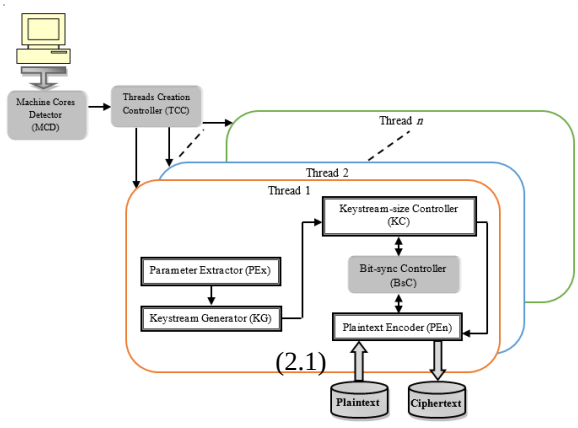


Fig. 4. The Architecture of the Parallel Platform.

The parallelized platform shown in Fig. 4 consists of several components, controllers and detectors. The main components of the platform are: Parameter Extractor (PEx), Keystream Generator (KG), Keystream-size Controller (KC) and Plaintext Encoder (PEn). In addition, the platform uses three supportive controllers and detectors: Machine Core Detector (MCD), Thread Creation Controller (TCC) and Bit-sync Controller (BsC). These controllers and detectors are designed to achieve the optimum level of performance gained from the parallel platform.

The parallel platform is also designed such that there is no direct-dependency among the components. In other words, the design of these components will enable us to easily parallelize the workload between the running threads. From another perspective, we have designed special detectors and controllers to ensure the consistency and accuracy of the keystream generation and plaintext encryption process.

As our platform may support different types of stream ciphers, the parallel platform is able to extract the corresponding parameters of the corresponding stream cipher. Accordingly, the Parameter Extractor component (PEx) extracts the required parameters to be used in other components, as shown in Fig. 5. The extracted parameters vary from one keystream generator to another in terms of the number of parameters, the size of the parameters and the representation of the parameters. For instance, if the DSP-128 stream cipher is selected, two parameters are extracted: parameter C (integer) of 128-bit length and parameter $\mathfrak{B}$ (polynomial) of degree 128.

The functionality of the Keystream Generator (KG) component is the most important part of the platform, whereby it is responsible for generating sequences of keystream bits. In this stage, a counter is used to increment the corresponding parameters, as shown in Fig. 6.

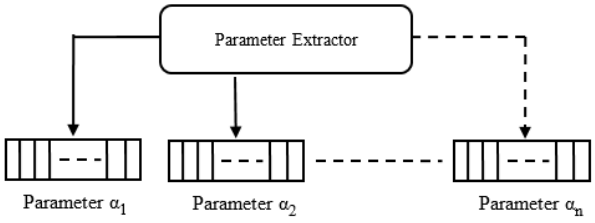


Fig. 5. Parameter Extractor (PEx) Component.

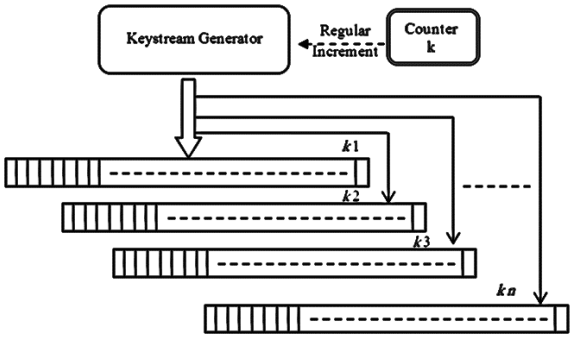


Fig. 6. Parallel Keystream Generated by the Keystream Generator (KG) Component.

The generation process of the keystream bits is mainly dependent on the parameters obtained from PEx to be applied on the mathematical problem being used in the core of the KG component. Fig. 6 illustrates the process of producing *n*-keystreams in *n*-rounds associated with every increment carried out by the counter *k*. Applying parallelism on the KG component to generate multiple keystreams concurrently is possible with the existence of controllers, which will be discussed later in this section. However, the only task that the KG component has to accomplish at this stage is to generate multiple keystreams in parallel, based on the incremented value of the extracted counter parameter, as shown in Fig. 7.

The KG component will use the initial value of the *counter* and the thread number (Thread_ID) to increment the value of the *counter*. At this stage, the KG component creates *n* threads (where *n = no. of_cores*) to handle the generation of new keystreams, concurrently.

Variable keystream lengths might be generated from the KG stage owing to the differences in the deployed stream cipher. Therefore, the Keystream-size Controller (KC) component is designed to standardize and control the size of the generated keystreams and limit their size to a 32-bit length.

As shown in Fig. 8, the keystream controller maps the *n*-bits of the reformatted key to a fixed-size key of *m*–bits length. The mapping process is known as pre-encryption processing. The size of the keystream can vary from one byte to five bytes. If the size of *n* is greater than 32 bits, **KC** will truncate the keystream to 32 bits and uses the rest of the bits in the following round, as presented by Equation (2):

$$K_{new} = Trunc\ (K_S, m) \quad (2)$$

```
FUNCTION KG(no_of_cores)

Pthread Thread[no_of_cores - 1];
Int Flag=1;

FOR i = 0 TO no_of_cores – 1
{
    WHILE (Flag) {

            create_thread (Thread[i], KGS, counter);}

Pthread_join(Thread[i]);
```

Fig. 7. The Code Snippet of KG in the Parallel Platform.

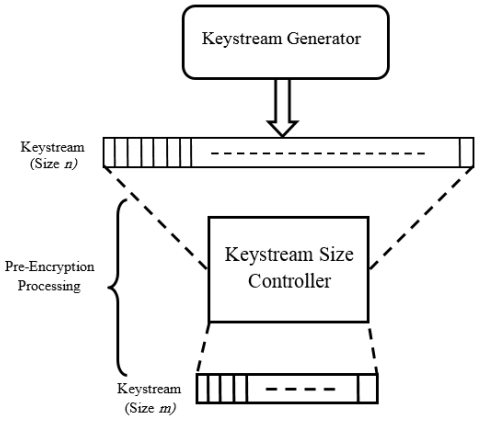


Fig. 8. Keystream-Size Controller (KC) Component.

where *m* is the upper limit of the size of the keystream (e.g. 32-bit). The content of the new keystream is given by Equation (3):

$$K_{new} = K_S[i], \text{ for all } 0 < i < m \quad (3)$$

The process of setting up the length of the final keystream is also known as the pre-encryption process, since the next component will use this key for data encryption. However, the importance of this component comes from the process of standardizing the size of the keystream regardless of the selected generator, which in turn gives the platform additional flexibility.

Consequently, the Plaintext Encoder (PEn) component is carried out. The main task of this component is to encrypt a sequence of plaintext and produce a corresponding ciphertext. The input to this component is the keystream generated by the KG component, as shown in Fig. 9. The encryption process is performed as follows: One word (32-bit) of plaintext is XORed with one word of the keystream. When PEn employs all the bits of the keystream, the KG component will be invoked to generate a new round of keystreams. Equation (4) forms the condition which needs to be satisfied when calling for the KG in a new round:

$$\text{Call (KG): } \sum \eta_{K_s} < \sum \eta_{P_t} \quad (4)$$

where η refers to the unused (available) bits.

Unlike other components, the encryption component (PEn) has a direct dependency on the keystream generator in which PEn must keep checking the number of available bits of the keystream in order to trigger the KG if the number of bits in the keystream is insufficient to perform the encryption.

The five components described above form the basic design of the proposed platform. The next step will parallelize these components and add other controllers and detectors, while applying multithreading techniques on multi-core processors for a fully parallelized platform.

Parallelizing the platform requires a detector called a Machine Cores Detector (MCD) to detect the number of cores on the corresponding machine. The MCD works at low hardware level in which it detects the total number of logical processors (cores) in a particular machine. We refer to the total number of cores as the NOC. The MCD is associated with a Thread Creation Controller (TCC) to create a specific number of threads, as many as the number of the available cores detected by the MCD. The other task of the TCC is to bind each job assigned to a thread with its correspondent core to isolate any potential concurrency issues (e.g. synchronization, system bottleneck, etc.) at the thread level. Accordingly, the TCC allows all jobs to run in parallel, in order to maximize the performance gains from the multi-core processors.

Binding each thread with one specific core is achieved by changing the scheduling policy by calling the processor's affinity routine. Processor affinity is designed to force threads to work on a specific core during the run. This is possible by using the POSIX threading library that provides developers with the one routine known by *pthread_setaffinity_np*. The default scheduling policy usually switches threads from one core to another during the run of multiple threads. Therefore, setting the processor affinity is sufficient to avoid thread switching since thread switching requires copying the thread instructions from its current L1 cache to the L2 cache of the new (switched) core.

For instance, in dual-core processors, the thread creation controller will create two threads and associate them with the two cores on that machine. The two cores share several resources and peripherals on a high-speed on-chip bus except the L1-cache, which is designated for each core. An example of an MCD and TCC implemented on a dual core processor is shown in Fig. 10.

For consistency, the keystream generation component and the plaintext encoder component are associated with an additional controller to ensure correct data encryption of each plaintext byte encrypted by its corresponding keystream. We refer to this controller as a Bit-sync Controller (BsC). This controller will ensure the synchronization between each bit in the plaintext with the corresponding keystream bit, for a correct decryption process. The encryption process ***E*** controlled by BsC is described in Equation (5):

$$C_t = E_{BsC(K_s)}(BsC(P_t)) \quad (5)$$

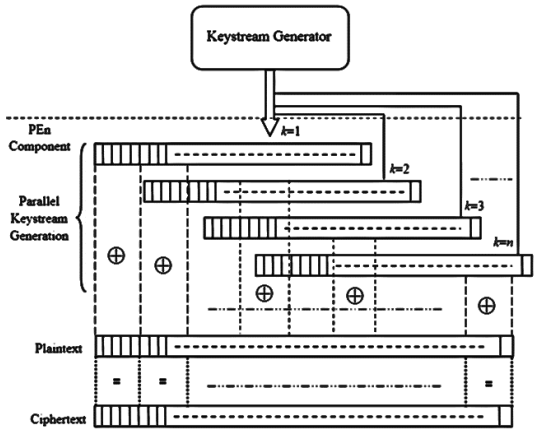


Fig. 9. Plaintext Encoder (PEn) Component.

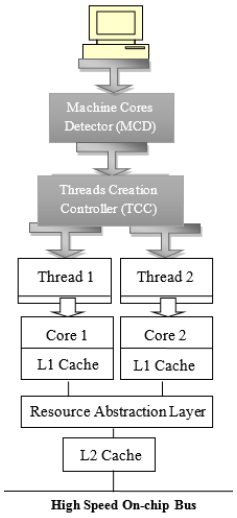


Fig. 10. MCD and TCC in Dual-Core Processor.

where C_t and P_t are the ciphertext and plaintext bytes respectively.

The BsC controller associates each 32-bit of plaintext with each thread so that the parallel threads can run concurrently. The pseudocode of BsC is shown in Fig. 11, and visualized by Fig. 12.

This algorithm is applicable to work on *n*-core processors, which provides a platform with higher scalability with regards to the rapid growth of the processor architecture.

Associating the keystream with the running threads is conceptually controlled in the same way as associating threads with their corresponding plaintext segment. There are two important equations used to increment the counter value associated with each round in the KG component. Let $\alpha_{counter}$ be the initialized counter extracted in PEx, the new initial value of $\alpha_{counter}$ is calculated and stored in a new counter denoted by *Ctr* as shown by Equation (6):

$$Ctr_0 = \alpha_{counter} + Thread\_ID \quad (6)$$

```
//Initial Parameters
NOB = 32 //number of bits
Flag = 1
bit_counter = NOB * (Thread_ID -1) +1
WHILE (more plaintext) //read plaintext for encryption
{
   IF (Flag)
      FOR bit_counter TO NOB*Thread_ID
         P[bit_counter] = plaintext[bit_counter]
   Flag = 0
   bit_counter = bit_counter + (NOB * (NOC-1)) + 1
   limit = bit_counter + (NOB – 1)
   FOR bit_counter TO limit
      P[bit_counter] = plaintext[bit_counter]
}
```

Fig. 11. The Code Snippet of a BsC Controller.

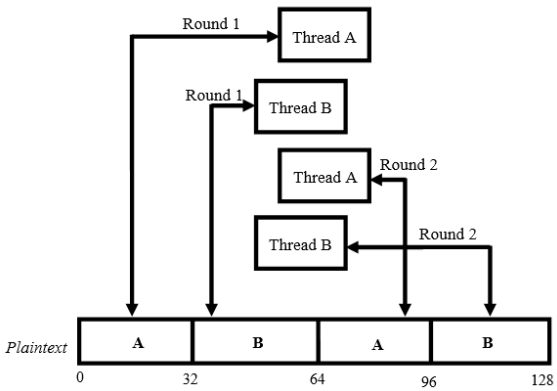


Fig. 12. Bit-Sync Controller (BsC).

The subsequent rounds will increment the value of *Ctr* as shown by Equation (7):

$$Ctr_{j+1} = Ctr_j + NOC \quad (7)$$

When the keystream generation component is requested to perform a new round, calculating *Ctr* by Equation (7) will be carried out. Table I shows an illustration of the associated counter values and plaintext segments with corresponding threads for three rounds of generating new keystreams in an 8-core processor. Let $\alpha_{counter}$ = 0, the incremented *Ctr* with its associated plaintext segments will be as follows:

Based on Table I, the thread with ID 1 will call KG three times and it will accordingly use the values 1, 9 and 17 as its counter values to generate three new 32-bit keystreams. The three generated keystreams by thread 1 will be associated with bits 1 to 32 (1-32), 257-288 and 513-544 of the plaintext, respectively. The functionality of the BsC controller is described in Fig. 13 where each KG component is associated to each of the existing cores.

Note that the plaintext segments are associated with a specific keystream generated by its corresponding counter.

However, the design of the platform and the plaintext encoder will ensure the correct sequence of the ciphertext segments, as shown in Fig. 14.

TABLE. I. ASSOCIATION TABLE BETWEEN THE COUNTER, PLAINTEXT SEGMENT AND THREAD ID

Thread_ID	1	2	3	4	5	6	7	8
Ctr_0	1	2	3	4	5	6	7	8
Plaintext Bits	1-32	33-64	65-96	97-128	129-160	161-192	193-224	225-256
Ctr_1	9	10	11	12	13	14	15	16
Plaintext Bits	257-288	289-320	321-352	353-384	385-416	417-448	449-480	481-512
Ctr_2	17	18	19	20	21	22	23	24
Plaintext Bits	513-544	545-576	577-608	609-640	641-672	673-704	705-736	737-768

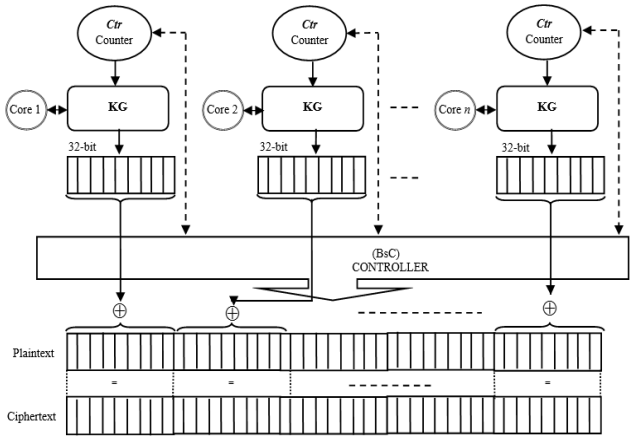


Fig. 13. Parallel Keystream Generation Controlled by BsC.

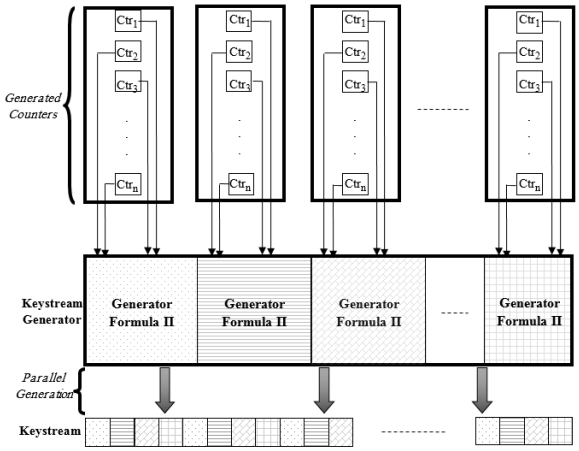


Fig. 14. Concurrency and Consistency in Parallel Keystream.

Fig. 14 illustrates the process of generating multiple independent and parallel counters by multiple threads. Technically, a copy of selected generator KG is associated to each core of the multi-core processor. Accordingly, each copy of the generator will use these independent counters to generate sequences of keystreams.

In this research, we present a high scalable platform that is capable of working on different numbers of cores on multi-core processor. The uniqueness of this parallel platform is that one can encrypt a stream of plaintext on *n*-core processors and decrypt the ciphertext on *m*-core processors (where $n \neq m$), as illustrated in Fig. 15 and 16 (for n=2 and m=4 respectively). This is due to the flexibility and the design of the platform's structure, which ensure the correctness of the encryption and decryption processes on any number of cores.

The process of matching a specific keystream sequence for a specific plaintext segment is a critical task that depends on the appropriate use of the counters. This task must be designed properly owing to its importance in allowing users to encrypt and decrypt their data on different numbers of cores. The following is an example (Example 1) of encrypting plaintext bits using keystreams generated by the keystream generator. The encryption is performed on a dual-core processor, while the decryption is performed on a quad-core processor (Example 2).

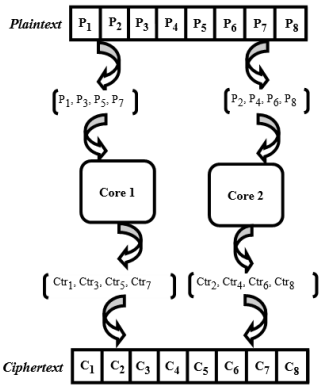


Fig. 15. Encryption Performed on Dual-Core Processor.

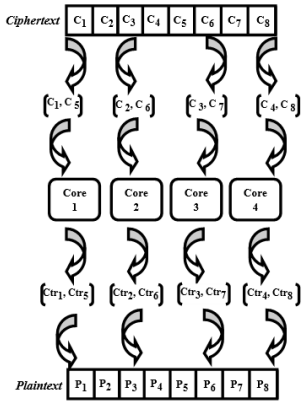


Fig. 16. Decryption Performed on Quad-Core Processor.

Example 1: Let P_t be the plaintext with length of five bytes (40 bits) such that P_t = 1101110111011011111000000111111110011111. The encryption of P_t on a dual-core processor with an 8-bit segment is performed as follows:

First, we divide the plaintext into segments of 8-bit length as follows:

	Segment (1)	*Segment (2)*	*Segment (3)*	*Segment (4)*	*Segment (5)*
$P_t =$	11011101	11011011	11100000	01111111	10011111

Subsequently, each thread associated with its corresponding core will generate a unique value of the counter *C* to be later used to encrypt a specific segment of the plaintext, as shown in Table II.

The resulting ciphertext is formed such that the position of the ciphered segment in the final ciphertext is based on the order of its corresponding counter. The ciphertext C_t will be in the following form:

	Segment (1)	*Segment (2)*	*Segment (3)*	*Segment (4)*	*Segment (5)*
$C_t =$	01110000	11001000	11000111	10111010	1101000

TABLE. II. GENERATING NEW COUNTER VALUE FOR ENCRYPTION ON EACH THREAD

	Core Number	**Counter C**	**KG $k = \Omega^C$**	**Ciphertext Segment C_t**	**Plaintext $P_t = k \oplus C_t$**
***Thread* (1)**	1	1	10101101	Segment (1) 11011101	01110000
***Thread* (1)**	1	3	00100111	Segment (5) 11100000	11000111
***Thread* (2)**	1	5	11110111	Segment (2) 10011111	01101000
***Thread* (3)**	2	2	00010011	Segment (3) 11011011	11001000
***Thread* (4)**	2	4	11000101	Segment (4) 01111111	10111010

One can see the association between the threads and the cores, such that thread 1 with counters 1, 3 and 5 will be executed on core 1, while thread 2 with counters 2 and 4 will be executed on core 2. Another important issue is related to the association between counter *C*, keystream and plaintext segments. This association will assure the consistency and coherence between the interchangeable encryption/decryption processes on a different number of cores.

However, decrypting the ciphertext above on quad-core processor is possible, and the following example (Example 2) shows the relation between the running threads and the ciphertext segments.

Example 2: Let C_t be the ciphertext of five bytes (1 byte = 8bits) such that C_t = 0111000011001000110001111011101 01101000. The decryption of C_t on a quad-core processor with an 8-bit/segment is performed as in Table III:

Similar to the encryption process in Example 1, the plaintext is formed such that the position of the plaintext segment in the final plaintext is based on the order of its corresponding counter. The plaintext P_t will be in the following form:

	Segment (1)	*Segment* (2)	*Segment* (3)	*Segment* (4)	*Segment* (5)
P_t =	11011101	11011011	11100000	01111111	10011111

Fig. 17 visualizes the relationship between multiple threads associated with multiple cores performing data encryption and decryption. However, the previous sub-sections have discussed the functionality of each component in order to understand the connection between those components running on multiple threads, forming a parallelized platform.

The proposed platform is mainly targeted for stream ciphers based on complex mathematical problems due to their high security attributes. The platform is designed to provide the opportunity for a more secure stream cipher to be designed regardless of the speed since the platform is able to provide those stream ciphers with higher efficiency and throughput. The platform is practical and has a great impact on the field of information security systems and cryptography.

TABLE. III. GENERATING NEW COUNTER VALUE FOR DECRYPTION ON EACH THREAD

	Core Number	**Counter C**	**KG $k = \Omega^C$**	**Ciphertext Segment C_t**	**Plaintext $P_t = k \oplus C_t$**
***Thread* (1)**	1	1	10101101	Segment (1) 01110000	11011101
***Thread* (1)**	1	5	11110111	Segment (5) 01101000	10011111
***Thread* (2)**	2	2	00010011	Segment (2) 11001000	11011011
***Thread* (3)**	3	3	00100111	Segment (3) 11000111	01111111
***Thread* (4)**	4	4	11000101	Segment (4) 10111010	10011111

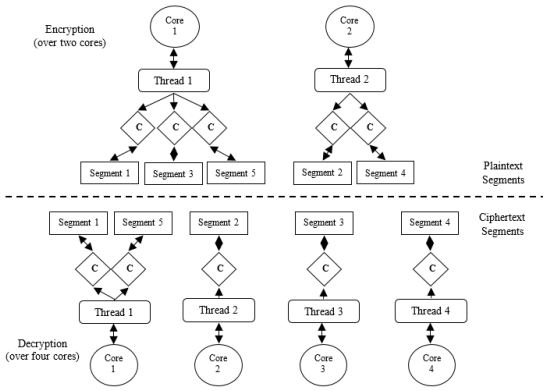


Fig. 17. Performing Encryption and Decryption on a different Number of Cores.

V. SECURITY AND PERFORMANCE ANALYSIS

Evaluating the efficiency of our parallel platform is measured against a set of stream ciphers. These stream ciphers are divided into two categories: parallelized and sequential ciphers. The chosen ciphers in our experiments are summarized in Table IV. The main reason of choosing these two categories of stream cipher is to examine the efficiency of these algorithms in utilizing the capabilities of the parallel platform.

The parallel platform is also measured from the security perspective. We analyze the impact of the parallel platform on the security attributes of the stream ciphers running over the platform. The platform is designed such that it does not affect the security attributes of the plugged-in stream ciphers, since each core is responsible for executing its own workload independently. Hence, there will be no interaction or dependency between any two or more keystream generators running over multiple cores, due to the avoidance of using global shared variables between the running threads.

The security level of the parallelized keystream generation on independent cores is similar to the security level of the sequential version of the stream ciphers. Technically, there are no shared parameters among the cores, which prevents any attempt to criticize the parallel platform on the security of the stream ciphers.

For testing purposes, we ran our model on two workstations. The first workstation (denoted by DualC) used an Intel Core 2 Duo ® E6400 processor of CPU speed 2.13 GHz, L2 cache memory of size 2MB, RAM of size 2GB. The second machine (denoted by QuadC) used an Intel Core 2 Quad ® Q6600 processor of CPU speed 2.40 GHz, L2 cache memory of size 8MB, RAM of size 2GB. The parallel platform was coded in C++ using MinGW-2.05 and tested on Microsoft Windows XP® operating system. POSIX-2.8.0 (*Pthread*) library was used to handle thread-related functions of the model.

Our testing started by examining the performance of the selected stream ciphers over DualC and QuadC machines. The results presented in Fig. 18 and 19 illustrate the performance of the stream cipher running over DualC and QuadC, respectively. Four sets of plaintext have been considered, of the sizes: 100, 500, 1000 and 2000 Mbits.

TABLE. IV. LIST OF STREAM CIPHERS CONSIDERED IN PERFORMANCE ANALYSIS

Stream ciphers	Category
Salsa20	*Parallelized*
DSP-128	*Parallelized*
ECSC-128	*Parallelized*
RC4	*Sequential*
Sosemanuk	*Sequential*

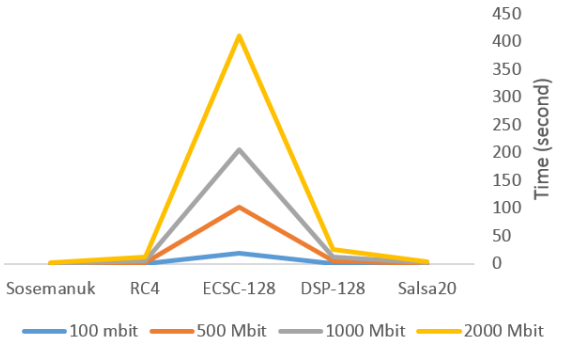


Fig. 18. Performance of Stream Ciphers over DualC Machine.

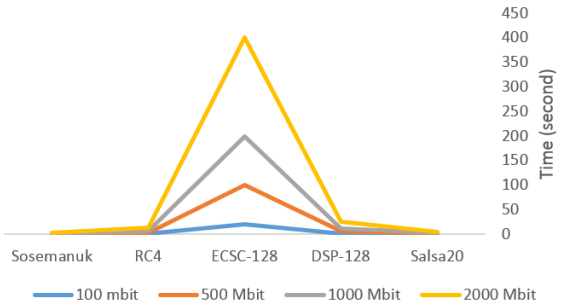


Fig. 19. Performance of Stream Ciphers over QuadC Machine.

Obviously, ECSC-128 is the slowest algorithm among the other algorithms. The results also show that little difference was found on the performance of the stream ciphers running on the quad-core machine compared to the encryption rates obtained on dual-core machines. The utilization of the two extra cores is not well identified by the selected stream ciphers. Table V presents the performance enhancements gained on the QuadC machine compared to the DualC machine.

On the next step, we plugged in the stream ciphers to our parallel platform (denoted by P(*stream cipher*)) to examine the impact of the platform on enhancing the encryption rates of these ciphers. Fig. 20 and 21 presents the results of running the five stream ciphers over the DualC and QuadC machines, respectively.

According to the performance analysis, we found that the parallel platform managed to support the parallelizable stream ciphers to utilize the two cores available on the DualC machine. Table VI shows that the parallel platform was able to enhance the encryption of the three parallelizable stream ciphers dramatically. The encryption rates of Salsa20, DSP-128 and ECSC-128 are enhanced by approximately 31%, 28% and 34%, respectively. However, the sequential stream ciphers are not capable of utilizing the support of the parallel platform.

TABLE. V. ENHANCEMENT RATIO GAINED FOR THE ORIGINAL STREAM CIPHERS RUNNING OVER QUADC MACHINE COMPARED TO DUALC MACHINE

Stream cipher	Enhancement Ratio
Salsa20	11%
DSP-128	6%
ECSC-128	2%
RC4	8%
Sosemanuk	9%

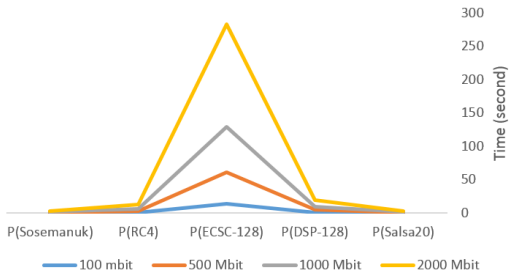


Fig. 20. Performance of Plugged Stream Ciphers on the Parallelized Platform over the DualC Machine.

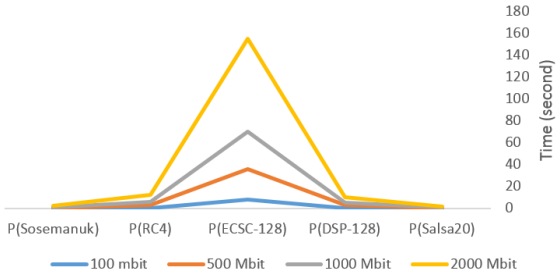


Fig. 21. Performance of Plugged Stream Ciphers on the Parallelized Platform over the QuadC Machine.

TABLE. VI. ENHANCEMENT RATIO GAINED FOR THE STREAM CIPHERS PLUGGED INTO THE PARALLEL PLATFORM AND RUNNING OVER DUALC MACHINE

Stream cipher	Enhancement Ratio
Salsa20	31%
DSP-128	28%
ECSC-128	34%
RC4	6%
Sosemanuk	10%

Similarly, the performance analysis shows that the parallel platform managed to support the parallelizable stream ciphers to utilize the four cores available on the QuadC machine. Table VII shows that the parallel platform was able to significantly enhance the encryption of the three parallelizable stream ciphers. The encryption rates of Salsa20, DSP-128 and ECSC-128 were enhanced by approximately 64%, 55% and 62%, respectively. However, the sequential stream ciphers are not capable of utilizing the support of the parallel platform.

To examine the efficiency of the parallel platform in utilizing the extra cores of QuadC compared to DualC, we compare the efficiency of the parallel platform over DualC and QuadC machines. Unlike the sequential stream ciphers (RC4 and Sosemanuk), results in Table VIII shows that an extra two cores doubled the encryption speed of the other parallelizable stream ciphers.

TABLE. VII. ENHANCEMENT RATIO GAINED FOR THE STREAM CIPHERS PLUGGED INTO THE PARALLEL PLATFORM AND RUNNING OVER QUADC MACHINE

Stream cipher	Enhancement Ratio
Salsa20	64%
DSP-128	55%
ECSC-128	62%
RC4	7%
Sosemanuk	17%

TABLE. VIII. ENHANCEMENT RATIO GAINED BY THE PARALLEL PLATFORM RUNNING OVER QUADC COMPARED TO DUALC MACHINES

Stream cipher	Enhancement Ratio
Salsa20	33%
DSP-128	27%
ECSC-128	28%
RC4	1%
Sosemanuk	7%

We conclude that the design of the stream ciphers plays an important role in utilizing multi-core processors. The parallel platform is able to enhance the encryption rate significantly on the QuadC machine with four cores, while the sequential stream ciphers failed to utilize such computing resources. Fig. 22-26 illustrates the efficiency of the stream ciphers over different environments, where Seq-DualC and Seq-QuadC refer to running the original stream ciphers on DualC and QuadC machines, and Parallel(DualC) and Parallel(QuadC) refers to running the stream ciphers with the support of the parallel platform on DualC and QuadC machines.

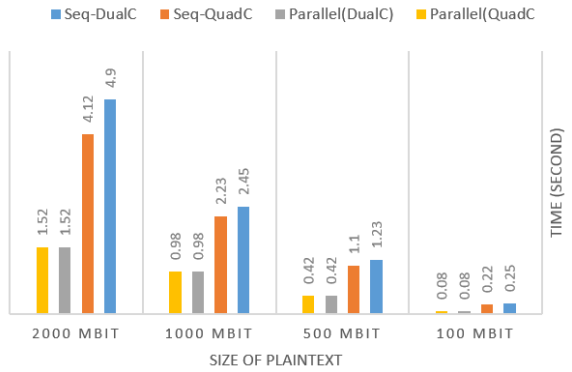


Fig. 22. Performance Efficiency of Salsa20 for different Environments.

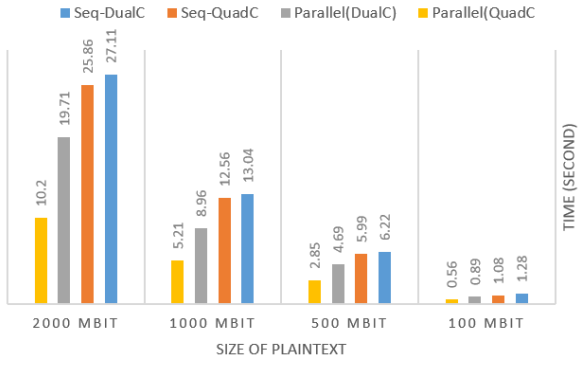


Fig. 23. Performance Efficiency of DSP-128 for different Environments.

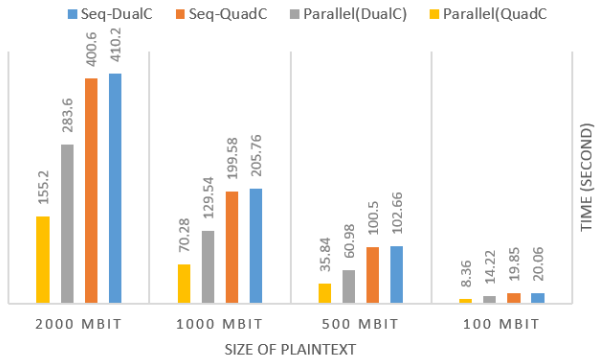


Fig. 24. Performance Efficiency of ECSC-128 for different Environments.

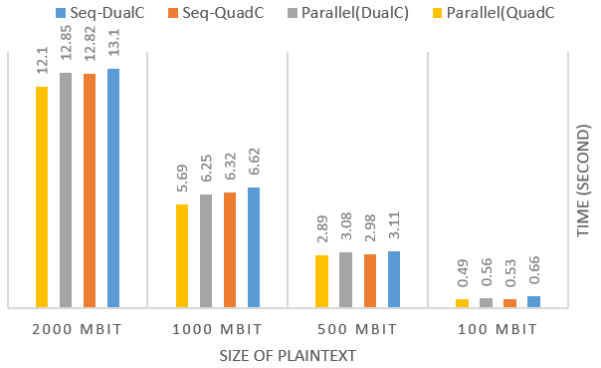


Fig. 25. Performance Efficiency of RC4 for different Environments.

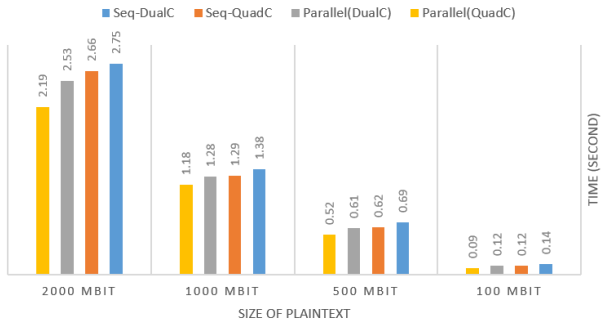


Fig. 26. Performance Efficiency of Sosemanuk for different Environments.

VI. CONCLUSION

In this paper we present a novel parallel platform to enhance the performance of stream ciphers. The underlying architecture of the platform relies on the use of multithreading technology. The platform is designed to be scalable and adaptable to the increasing number of cores in the future. Parallelism on our platform is implemented at two levels: task and data parallelism. Task parallelism is achieved by dividing the workload among the available cores in the corresponding machine, where each core will have its own components and parameters set. On the other hand, data parallelism is achieved by encrypting smaller sets of plaintext in multiple cores, concurrently.

The experiments' results show that parallel stream ciphers (Salsa20, DSP-128, ECSC-128) are capable of achieving higher performance on the parallel platform. The results also

show that increasing the number of cores from two to four cores has doubled the performance of these three algorithms. This is due to the parallelizable design of these ciphers. However, sequential stream ciphers (RC4, Sosemanuk) are not able to utilize the support of the parallel platform running over the quad-core machine.

From the security perspective, the underlying architecture of the parallel platform is constructed to avoid the existence of shared global of local attributes between the running keystream generators. Each core is associated with one independent set of data and operates over separate input keys and counters. Accordingly, the parallel platform does not affect the security of the plugged-in stream ciphers.

ACKNOWLEDGMENT

The author would like to express her thanks to Arab Open University, Saudi Arabia for supporting this study.

REFERENCES

[1] Nutaro, J., and B. Zeigler. "How to apply Amdahl's law to multithreaded multicore processors." Journal of Parallel and Distributed Computing 107: 1-2, 2017.

[2] Dang, H., M. Snir, and W. Gropp. "Eliminating contention bottlenecks in multithreaded MPI." Parallel Computing 69: 1-23. 2017.

[3] Thébault, L., and E. Petit. "Asynchronous and multithreaded communications on irregular applications using vectorized divide and conquer approach." Journal of Parallel and Distributed Computing 16-27: 16-27. 2018.

[4] Soni, V., A. Hadjadj, O. Roussel, and G. Moebs. "Parallel multi-core and multi-processor methods on point-value multiresolution algorithms for hyperbolic conservation laws." Journal of Parallel and Distributed Computing 123: 192-203. 2019.

[5] Ogiela, M. "Cognitive solutions for security and cryptography." Cognitive Systems Research 55: 258-261. 2019.

[6] Hiscock, T., O. Savry, and L. Goubin. "Lightweight instruction-level encryption for embedded processors using stream ciphers." Microprocessors and Microsystems 64: 43-52. 2019.

[7] CharlesLai, B., K. Li, and C. Chiang. "Self adaptable multithreaded object detection on embedded multicore systems." Journal of Parallel and Distributed Computing 78: 25-38. 2015.

[8] Crainicu, B. "Unified Formal Model for Synchronous and Self-Synchronizing Stream Ciphers." Procedia Engineering 181: 620-625. 2017.

[9] Khelifi, F. "On the security of a stream cipher in reversible data hiding schemes operating in the encrypted domain." Signal Processing 143: 336-345. 2018.

[10] Rivest, R. The RC4 Encryption Algorithm. RSA. Document No. 003-013005-100-000000, USA: Data Security Inc. 1992.

[11] Berbain, C., O. Billet, A. Canteaut, N. Courtois, H. Gilbert, L. Goubin, A. Gouget, et al. "Sosemanuk, A Fast Software-oriented Stream Cipher." Accessed June 22, 2019. http://www.ecrypt.eu.org/stream/p3ciphers/sosemanuk/sosemanuk_p3.pdf. 2005.

[12] Bernstein, D. "The Salsa20 Family of Stream Ciphers." In New Stream Cipher Designs, 84-97. Berlin: Springer-Verlag. 2008.

[13] Bernstein, D. "The ChaCha family of stream ciphers." D. J. Bernstein's webpage. Accessed June 20, 2019. http://cr.yp.to/chacha.html. 2005.

[14] Suwais, K., and A. Samsudin. "DSP-128: Stream Cipher Based On Discrete Log Problem And Polynomial Arithmetic." American Journal of Applied Sciences 5 (7): 896-904. 2008.

[15] Suwais, K., and A. Samsudin. "ECSC-128: New Stream Cipher Based on Elliptic Curve Discrete Logarithm Problem." First International Conference on Security of Information and Networks. Famagusta. 13-23. 2007.

Segmentation of Crescent Sand Dunes in High Resolution Satellite Images using a Support Vector Machine for Allometry

M. A. Azzaoui[1], L. Masmoudi[2]
Laboratoire Conception et Systèmes Électronique, signaux et Informatique Université Mohammed V de Rabat, Morocco

H. El Belrhiti[3]
Département des Sciences Fondamentales et Appliquées Institut Agronomique et Vétérinaire Hassan II Rabat, Morocco

I. E. Chaouki[4]
Département Informatique Ecole Nationale des Sciences Appliquées d'Agadir Agadir, Morocco

***Abstract*—The study of sand dunes movement is essential to understand and prevent the desertification phenomenon, and collecting data from the field is a labor intensive task, as deserts contain usually a large number of sand dunes. We propose to use computer vision and machine learning algorithms, combined with remote sensing and specifically high resolution satellite images for collecting data about the position and characteristics of moving sand dunes. We focused on the fastest moving sand dunes called barchans, which are threatening the settlements in the region of Laayoune, Morocco. We developed a process with three stages: In the first stage, we used an image processing approach with cascading Haar features for the detection of dunes location. In the second stage, we used a support vector machine for the segmentation of contours, and in the third stage we used an algorithm to measure the allometric features of barchans dunes. We explored the collected data, and found relevant correlations between dunes length, and width, and horns sizes, which could be used as key indicators for dunes growth and progression. This study is therefore of high interest for urban planners and geologists who study sand dunes and require technical methods, based on machine learning and computer vision to allow them to collect large amount of data from satellite images to understand sand dunes progression and counter desertification problems. The use of cascading Haar feature provided a good accuracy, and the use of Support Vector Machines, along with the high resolution satellite images provided a good precision for the segmentation of barchan dunes contours, allowing the collection of morphological features which provide significant information on barchans sand dunes dynamics.**

Keywords*—Image segmentation; support vector machines; high resolution satellite images; remote sensing; sand dunes; desertification

I. Introduction

A. The Study of Barchan Sand Dunes

Barchan sand dunes are the fastest moving sand dunes, thus, they pose a problem to human settlements in desert arid and semi-arid regions around the world, as they could cover the farmlands, and cause the degradation of cultivated crops, damage houses, and trigger immigration. Also, moving sand dunes can cover intercity roads, which may create accidents, and block the road traffic for many hours, thus limiting the exchange of goods and restrict the economic development of arid and semi-arid regions. In fact, desertification affects around 250 million worldwide and 110 countries could be facing its consequences due to climate change [1]. Therefore, the study of sand dunes, and especially the fastest moving ones called barchans is important, and their large scale implies the use of advanced computer science and machine learning methods, in order to provide more relevant quantitative data, which can be used to assist specialists to understand better this phenomenon, and help decision makers to get an meaningful insight, and thus, contribute to find more efficient solutions to limit the impact of barchans dunes progression on human activities.

B. Detection of Sand Dunes in Satellite Images

The use of satellite imagery along with computer vision and digital image processing algorithms is essential to allow collecting data about sand dunes which are usually scattered on the surface of large deserts. Indeed, researchers have been using remote sensing imagery to study sand dunes and collect data more efficiently in such harsh environments: In the early attempts [2] used remote sensing to manually map sand dunes and develop their taxonomy, later, [3] started the detection of sand dune using their surfaces reflectance. In [4], pairs of ASTER satellite images were combined to extract barchans and collect their morphology measurements and study their dynamics. In [5], convex hulls were used to combine SRTM Digital Elevation Model and orthorectified Landsat TM, then R-statistic index to estimate the degree of self-organization of dunes in Ar Rub' al Khali sand sea on the Arabian Peninsula. In [6], LiDAR was combined with spectral and spatial resolution of QuickBird satellite images, and optical stimulation luminescence was used to create an elevation model for Canadian barchans dunes. In [7], the mineral composition in Gran Desierto dune field in New Mexico, USA was mapped using ratio indexes on images from ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer). In general, remote areas for which the access is sometimes impossible have benefited from using remote sensing along with computer vision and machine learning algorithms advancements. Indeed, availability of images from Landsat, MODIS, IKONOS, HIRISE, and MRO among others allowed not only to study dunes in difficult access areas such as the

Namib Desert and Sahara Desert but also to discover the existence of barchans sand dunes in other planetary systems such as Mars: In [8] images from MRO (Mars Reconnaissance Orbiter) were used to study the seasonal modifications of Mars Northern Polar dunes thanks to high signal-to-noise ratio (SNR) in the dark dunes images. Dunes were also found in Venus: In [9], Cassini Radar images were used and contrasted black dunes to estimate their properties. We used in our previous work [10], the HIRISE (High-Resolution Imaging Science Experiment) satellite images to detect barchan dunes using MSER blobs. Barchans dunes tend to form group patterns [11], which complicates the detection of their position, as they usually overlap, which makes the use of more advanced segmentation methods necessary in order to distinguish dunes from their background.

C. Image Segmentation and Machine Learning

The spatial and spectral properties of satellite images can lead to good results for the detection and segmentation of sand dunes when the configuration of dunes is simple and the images are well contrasted. However, in many cases, the dunes are not in a stable system, and interact with vegetation and crust, therefore, the classical image detection approaches we introduced in the previous chapter face new challenges, especially in more complex dune fields, where the dunes are colliding. Hence, more advanced segmentation techniques were introduced in order to differentiate sand dunes from their surroundings. Researchers have tried to overcome these issues using different image segmentation approaches: In [12] pixel based segmentation was used to study Salinas-El Espartal dunes in the coast of Asturias, Spain. In [13], Maximum Likelihood and Minimum Distance Parallelepiped were demonstrated for classifying barchans. Sub-pixel analysis was extensively used to analyze dunes in Sinai Peninsula, Egypt [14]. In remote sensing, the use of methods based on object-oriented segmentation increased as the resolution and availability of satellite image improved, and the research switched progressively from the use of pixel-based methods to GEOBIA (Geospatial Object Based Image Analysis), allowing the use of more advanced image segmentation methods [15]. Image segmentation could be split into four mains classes: The region-based segmentation, edge-based segmentation and point-based segmentation, along with their combination [16]. Regions can be defined using homogeneity criteria based on a feature space, which is derived from the multi spectral satellite images, and thus have regions can encompass more information that segments, which in turn take advantage of spectral information to convey the mean, variance, median, minimum and maximum values per image bands, and thus, hold more information than single pixels. The Geospatial applications have been integrating segmentation methods developed for other domains such as pedestrian recognition for road safety [17], or for medical image analysis [18]. Hence, the models based on color and texture segmentation, and also relying on shape or motion [19] were used in satellite image processing. In fact, distances, neighborhood and topologies are essential for the object-based methods. Image segmentation has evolved with the availability of benchmark databases and the development of machine learning approaches, which required to be fed with features. Therefore segmentation methods were based on feature engineering, and many descriptors and detectors were developed such as Haralick features [19] which were invariant to quantization gray-levels number, and were calculated using GLCM (Gray-Level Co-occurrence Matrix), thus taking into account image texture. They were used for the discrimination of sand dunes in the Valdes peninsula [20]. The Histogram of Oriented Gradients (HOG) which became popular [21] and Local Binary Patterns (LBP) introduced in [22] served as descriptors and were also used in for the detection of sand dunes [23]. Later, other detectors and descriptors such as Scale-Invariant Feature Transform (SIFT) was used extensively in different computer vision applications [24] such as pedestrian detection, fingerprint recognition, however, at the extent of our knowledge, they were not used in sand dunes detection in satellite images. It is worth noting other important descriptors such as Speed-Up Robust Features (SURF) [25], played an important role. In fact, SIFT performs better than SURF for scale changes, but SURF has lower execution time than SIFT when it used integral images [26]. Fast Retina Keypoint (FREAK) which was introduced by [27] consists of series of Difference of Gaussians (DoG) which are composed over a retinal pattern, was also used in many applications in digital image processing and object recognition. and Binary Robust Independent Elementary Features (BRIEF) which is highly discriminative in the case of few bits [28], and There is also Binary Robust Invariant Scalable Keypoints (BRISK) [29] which has a lower computational cost, Oriented FAST and Rotated BRIEF (ORB) [30] which is rotation invariant and resistant to noise. Also, MSER (Maximally Stable Extremal Regions) detectors were efficient as their time-steps were chosen so as to achieve more stability against changing intensity in different scales, in addition to blur [31]. These descriptors were vectors which represented information, and thus, simple threshold were not very effective to classify complex objects and images. Indeed, the descriptor vectors were later followed by the use of machine learning algorithms, which had to be used generally in the paradigm of supervised learning, where the data was split into a learning set, a validation set and a test set [32]. The use of classification or regression techniques is generally based on LR (Logistic regression), SVM (Support Vector Machines), ANN Artificial Neural Networks, and their multiple variations such as CNN (Convolutional Neural Networks) which includes DL (Deep Learning). The classification of remote sensing images also made use of Markov Random Field (MRV) [33].

D. Our Contribution

As we showed in the previous paragraphs, there were many studies for quantifying sand dunes dynamics, but very few used computer vision and machine learning approaches, and no study in our knowledge went through the entire process from the detection of barchans dunes in satellite image, then image segmentation and finally the automatic measurement and comparison of features with allometry algorithms. Therefore, our contribution is an end to end process which starts from a raw satellite image, used supervised learning to detect dunes, segment their contours using SVM, and then collect they allometry with a dedicated algorithm. In the following chapters, we will provide the detailed steps of our approach, then, we will expose the experimental results and finally, we will give a glance at the upcoming direction of future works, and finally we will conclude.

II. Proposed Work

In the following flowchart Fig. 1, we depict the steps we used to achieve the final goal which is the automatic extraction of dunes measurements. Each of the three lanes represents one major step:

In the first lane, we detect the dunes locations in a satellite image: We start by splitting our high resolution satellite image randomly in two parts: The first part, which is about 70% of the total area, is used for training, and the remaining area, around 30%, was put aside for testing and evaluating the learning. We then started by a simple image enhancement using Weiner filter, then, we manually annotated dunes in order to have a ground truth. Then, we used a cascade classifier based on Haar wavelets to learn the dune model parameters. The result was a trained model, which we evaluated on the test region, in order to get predicted detections of barchans dunes from the cascade classifier model. As we noticed that few dunes were detected more than once by the classifier, we added a processing step which is candidate's fusion, in order to combine overlapping bounding boxes. The result of this first step is thus a set of images containing the detected barchans sand dunes.

In the second lane, we started by using a machine learning approach and specifically supervised learning. We used the dunes annotation of bounding boxes from the first step, and annotated the contours in addition, to get a labelled learning set, and again, we left a test set for evaluation. We proceeded in the learning set and normalized the data, as dunes come in different sizes and shapes. We used shapes alignment, and later on dimensionality reduction in order to compress the information coming from the 2D shapes of dunes, thus extracting the most significant ones. The result was a shape model which describes a mean dune, and the variance which covered most of dune shapes. This result allowed us to have a check on the geometric information. From another hand, to recover the geometric position from the test images, we collected SIFT image descriptors in specific points. More accurately, we extracted SIFT vector descriptor around each annotated contour, and onto the contour itself, then labelled the SIFT points on the contour as positive, and those around the contour as negative.

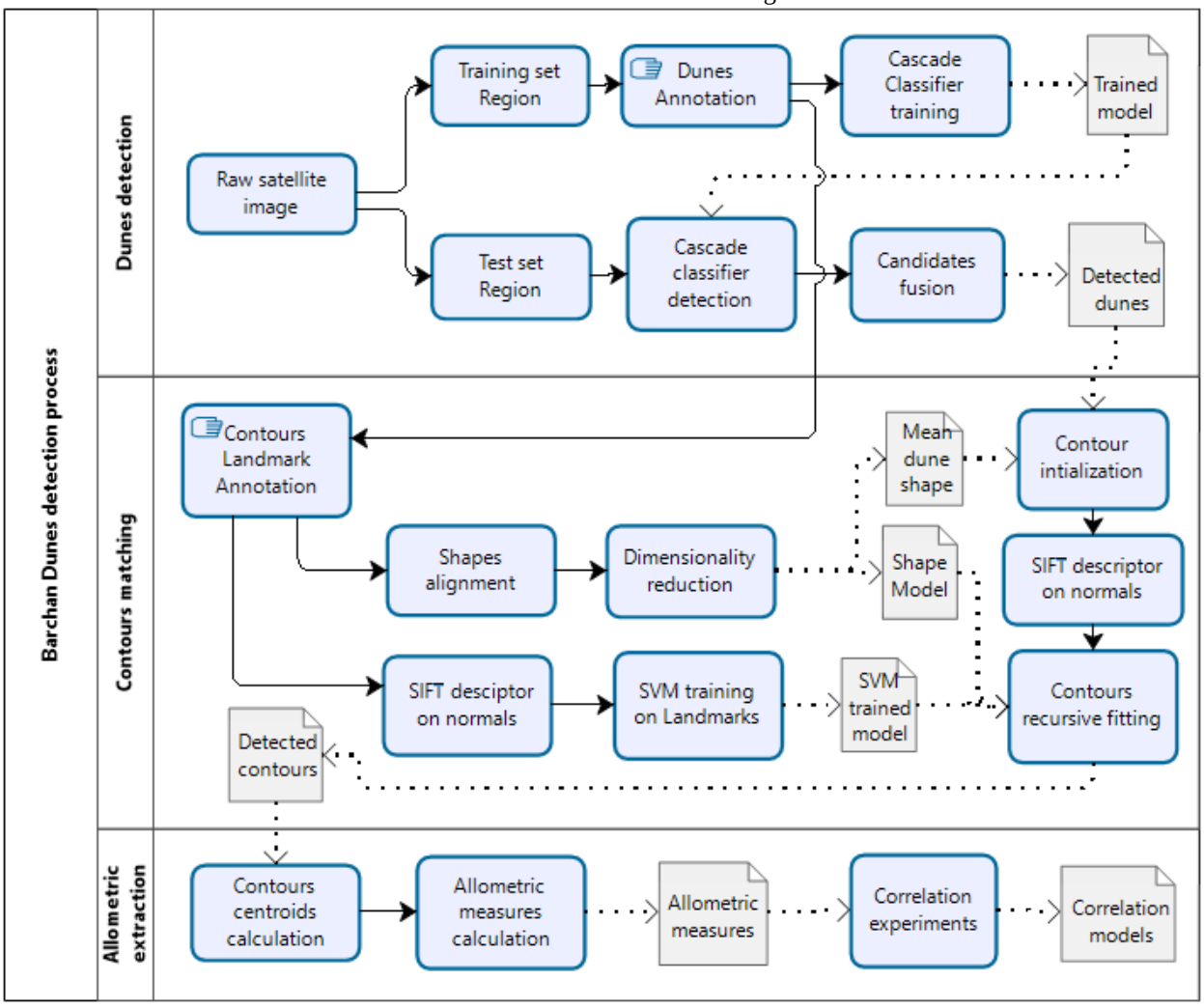


Fig. 1. Flowchart of the Process of Detection, Segmentation and Algorithmic Allometry of Barchans Sand Dunes.

We then trained a Support Vector Machine to classify the SIFT descriptor points as belonging to a contour of a dune, or not if it belongs to the surroundings. The produced model was then used for predicting if the new SIFT points were belonging or not to the contour. Now that the SVM model, shape model, and the mean dune are ready, we proceed used the test images produced by the first step. For each of the test images, we started by plotting the mean dune that is our starting point, from which we searched to match the contours of the dune which was detected in the test image. For this purpose, we started by extracting SIFT features around the curve of the mean dune, along its normal, then we classified them using SVM, as belonging to the dune contour or not. Those who were positive were used to deform the mean dune, into a new shape, which was closer to the shape of the dune in the test image. Then, to check that the new shape was acceptable, we used the shape model produced with PCA to correct the dune shape, and keep it within the acceptable standard deviation of the dune model. Thus, if the shape deviated more than 3σ, it was set 3σ. This operation was done recursively in order to match the contours of the barchans dune, and we stopped after 10 iterations for each image as it was sufficient to converge toward the dune contour in the test image. The result for each dune image was its 3 contours called (Brink, Leeward and Windward.) contours.

In the third lane, we used the contours produced by the former step then calculated for each one, its centroid. These centroids along with the contours will be used by our algorithm which calculates the allometric measures of each dune. First, the line traversing the 3 centroids formed an axis, which we call the dune direction. The other metric automatically extracted are the width, found by drawing an orthogonal line to the direction passing by the crest, and intersecting with the outward contours of the dune. We calculated horns centroids, then horns width, as the sum of both horns distances. This measure was calculated as the distance between the parallel lines to the direction, passing respectively through a horn centroid and the outermost contour. Finally, we explored the collected data and extracted correlation models.

III. Experimental Results

A. Satellite Image Dataset

We used an IKONOS high resolution satellite image from the satellite IKONOS, which is an Earth observation satellite used to collect imagery on multispectral (MS) and panchromatic (PAN) wavelengths. Its band resolutions consist of one panchromatic band with 0.8 m spatial resolution for 450-900 nm wavelengths, and 4 multispectral bands with 4 m spatial resolution with Near Infrared, Red, Green and Blue corresponding respectively to 757-853 nm, 632-698 nm, 506-595 nm and 445-516 nm wavelengths. The location of the image was in the Sahara Desert, more precisely in the South of Morocco, and the limiting coordinates of the image correspond to 27°26'8.6621"N, 13°08'5.2628"W and 27°41'1.0350"N, 13°22'0720"W. The date the image was taken on by the satellite was the 23/07/2003. The total area of the image corresponds approximately to 12 km2. The area contains hundreds of sand dunes, and we focused on a sub area where dunes were not colliding much. Our dataset consisted of 240 images with 120 sand dunes split into 80 sand dunes used as a learning set and 40 used as a test set, plus 120 images without sand dunes, containing mainly the flat desert surrounding the dunes.

B. Metrics and Formula

Barchan sand dunes are three dimensional objects, which appeared as 2D crescent like objects in satellite images. We therefore used an approach where we first scaled the multispectral images to the size of the panchromatic image using bicubic interpolation for 4x4 pixels using equation (1):

$$p(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \quad (1)$$

After enhancing the image using the median filter, we proceed and used a Haar classifier for the first step, which was based on selecting the best weak classifier with respect to the weighted error using equation (2):

$$\varepsilon_t = min_{f,p,\theta}\sum_i w_i |h(x_i,f,p,\theta) - y_i| \quad (2)$$

The results were candidate dunes, surrounded by bounding boxes. They were combined when their overlapping was over 40%. Then each selected bounding boxed was enlarged by 25% then all ROI (Regions of Interest) were normalized by resizing them to 128x128 pixels. In the second step, SIFT descriptors which are based on the difference of Gaussians in (3) were extracted in 10 equidistant points from the contours named landmarks, and 4 pixels along the normal of each landmark, which is a total of 50 SIFT point per dune image.

$$G(x,y,k\sigma) - G(x,y,\sigma) \approx (k-1)\sigma^2 \nabla^2 G \quad (3)$$

After the SIFT features were calculated, the dune shapes represented by their contours were aligned and SIFT descriptors were fed to an SVM, so as the descriptors falling on the contours were defined as positive examples, and the SURF descriptors falling on the normal surrounding the contours landmarks were defined as negative examples. The SVM maximized the separating hyper plan described in (4), where we used a linear Kernel.

$$h(x) = \sum_{k=1}^{p} \alpha_k^* l_k K(x_k, x) + w_0 \quad (4)$$

Subsequently, for testing and evaluation, we used the dunes test examples, and calculated with Mahalanobis distance. Finally, in the third step, we used a set of normalized allometry features defined in Fig. 2, which were geometrical measures calculated with our algorithm. We produced for each dune, the vector V(d) in (5), with w,L,θ,h1,h2,ih,S corresponding respectively to Width, Length, Angle, Horn widths, inter-horns distance, and Surface area.

$$V(d) = (w, L, \theta, h_1, h_2, ih, S) \quad (5)$$

These measures are illustrated in Fig. 2 which shows, the centroids corresponding to different contours: Cw, Cb, Cl to windward, brink and leeward contours respectively, and Clb1, Cw1, to the extremities of leeward and brink, and the extremity of windward curved respectively. NS represents a hypothetical North-South axis.

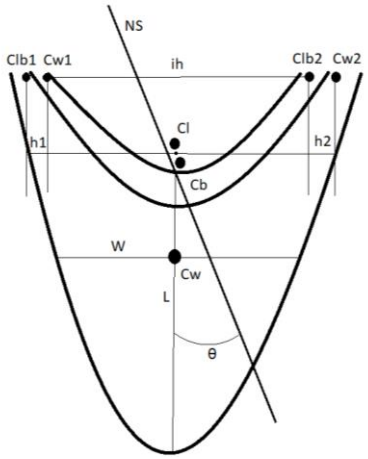


Fig. 2. Allometry Measures of Barchan sand Dunes Contours.

C. Graphics and Comments

We show in Table I the confusion matrix from which we calculated the accuracy of the dunes detection. When we used the fusion of dunes detection, the results returned for the dune detection increased from 65% to 77.5%. This increase is explained by the fact that overlapping detection tends to create false positives, which were generally corresponding to parts of the dunes such as horns. Fusion was reducing the number of false negatives, thus improving the accuracy.

TABLE. I. CONFUSION MATRIX OF DUNES DETECTIONS

Confusion matrix of dune detection		***Truth***	
		Positive	***Negative***
Detection results before fusion	**Positive**	18	23
	Negative	5	34
Detection results after fusion	**Positive**	29	12
	Negative	6	33

The goal of this study is to use machine learning and computer vision approaches in order to collect dunes measures (width, length, horns distances, …) which may be used by other domain specialists. We therefore demonstrate in our own dataset a set of correlations found between these dunes features in Fig. 3.

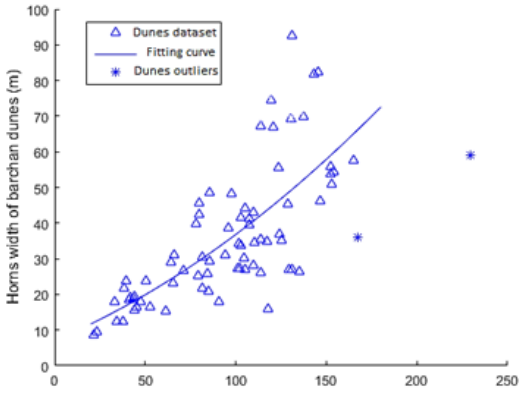
The relationships found in Fig. 3 from top to bottom correspond respectively to the equations (6), (7) and (8).

$$w = 8.10^{-3}(h_1 + h_2)^2 + 0.2\,(h_1 + h_2) + 7.1 \quad (6)$$

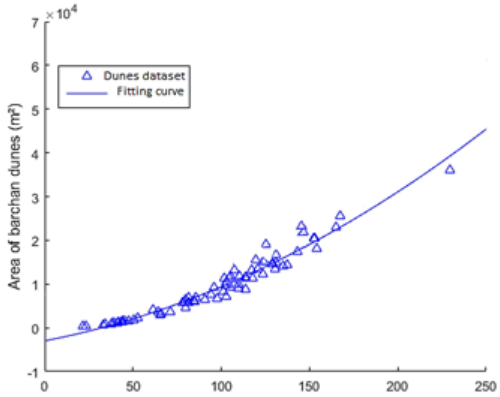
$$S = 10^{-3}\,(5.10^{-4}\,w^2 + 77.10^{-3}\,w - 2.97) \quad (7)$$

$$w = 19.89 + 0.85\,ih \quad (8)$$

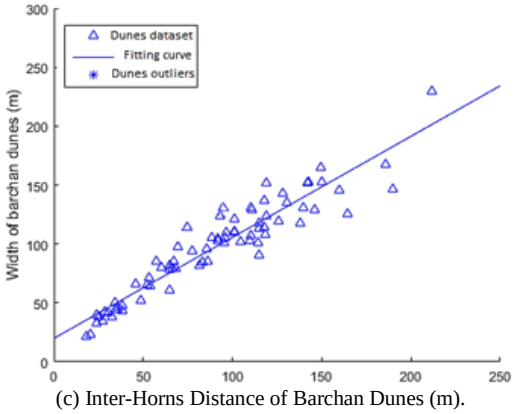
These equations are the results which can be useful to domain experts and help decision makers to estimate the dunes risk and danger. Our contribution consists of the machine learning and computer vision approach organized in a process we illustrated in Fig. 4, describing each step. The original satellite image is enhanced in Fig. 4(a), then the use of a Haar classifier resulted in the detection of dunes indicated by the bounding boxes, the Fig. 4(c) is a zoomed in part of the result. For each bounding box, we normalized the image to 128x128 pixels in Fig. 4(d). Then, we extracted the SIFT descriptor features around the mean dunes, and along the normal, which we provided to a SVM (Support Vector Machine) in order to classify positive SIFT landmarks corresponding to the contours, and negative SIFT, not corresponding to the contour. Thus, by linking the SIFT positive points; we obtained a set of 3 curves represented in Fig. 4(f). Finally, we used an algorithm for measuring allometry features represented in Fig. 2, in order to normalize the data measures. These measures depicted by yellow segments in Fig. 4(f) are then stored in a vector (5). We then used the data which we found using our dunes detection and segmentation approach, to create a set of features vectors (5), and explored correlations.



(a) Width of Barchan Dunes (m).



(b) Width of Barchan Dunes (m).



(c) Inter-Horns Distance of Barchan Dunes (m).

Fig. 3. Dunes data feature correlations examples: (a): Width as a function of horns width: f(w) = h1+h2 (b): Area as a function of width: f(w) = S (c): Width as a function of inter-horns: f(ih) = w

In Fig. 5, we exposed the results we obtained after using our computer vision approach, which allowed us to determine dunes measures, after detecting and segmenting dunes from high resolution satellite images using SVM. We showed in Fig. 5(a), one of the most interesting results we obtained: After collecting data about the surface size of dunes as a function of their angle, we found a normal like distribution where the largest dunes were centered on $\theta = 22.5$ degrees (or NNE). Moreover, the histogram in Fig. 5(c) produced by data collected using our method also showed that most of the dunes are centered on the same angle θ. As it was a specific numbers, we decided to further search for an external data source on to confirm or infirm the results obtained by our method of data collection. Thus, we found that the wind rose in Fig. 5(b) which provided us with the wind direction in the region from which our dataset originated in the satellite image, showed the wind blowing mainly from 22.5 degrees (or NNE). Therefore, it confirmed that our approach based on machine learning and computer vision process to automated process of data collection from satellite images could very accurately measure barchans dunes allometry features, and obtain results, which were coherent with the field.

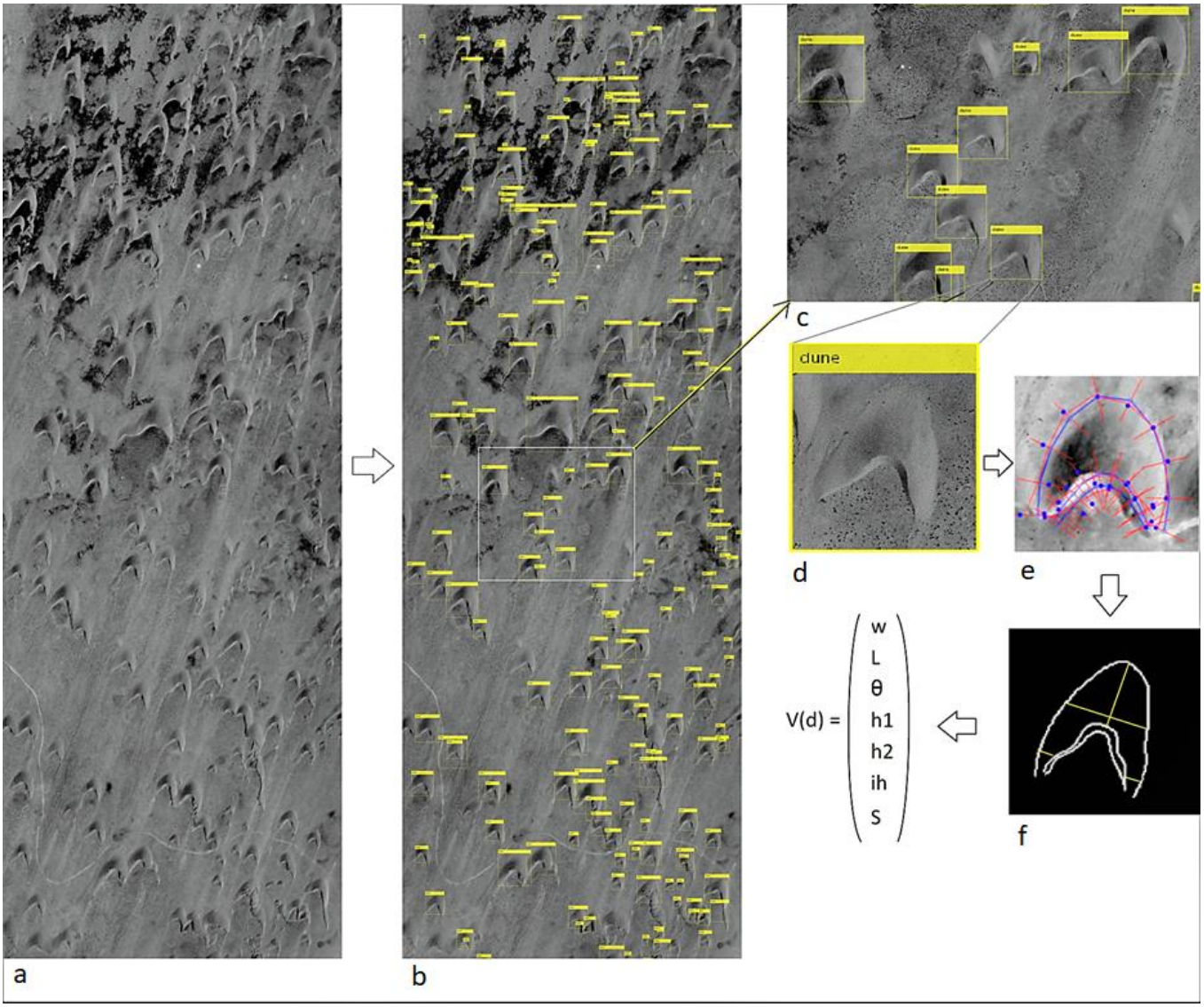
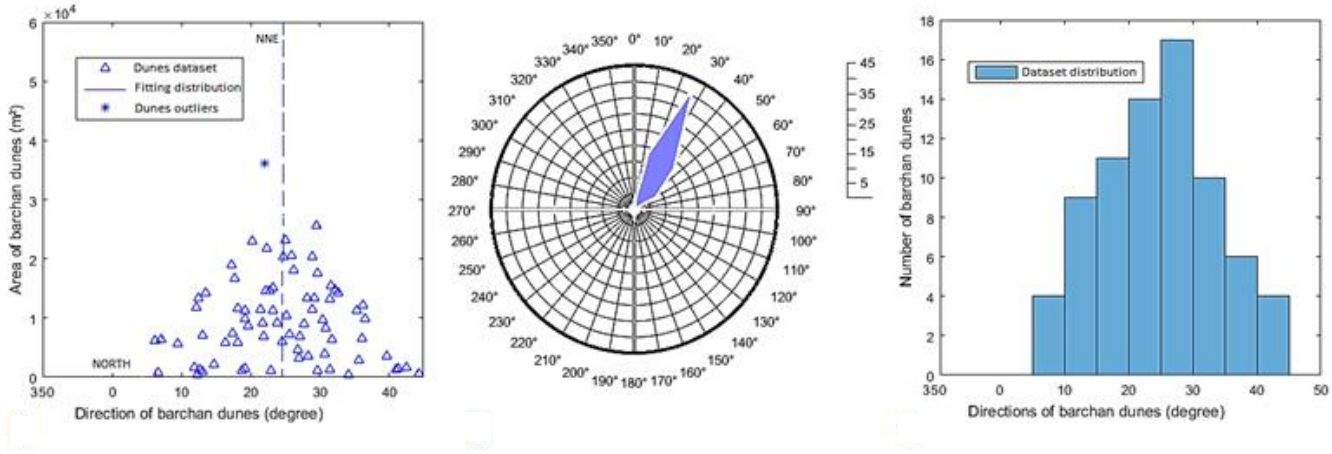


Fig. 4. Image Detection and Contours Segmentation Process for Collecting Dunes Measurements (a): The Satellite Image after Image Enhancement (b): The Satellite Image after Dunes Detection using Haar Wavelets (c): A Zoomed in Portion of the Satellite Image in 2 (d): The Dunes after ROI Normalization to 128x128 Pixels (e): Red Lines are Normals Along which SIFT Descriptors were Calculated. Blue Dots Represent the SIFT Classified as Positive using SVM, Forming Blue Contours (f): Resulting Dune Contours are in White, Dunes Features Measured by our Algorithm are Yellow Segments, Final Result is a Vector (Width , Length, Widths of Horns).



(a) (b) (c)

Fig. 5. Measures of Barchan Sand Dunes Contours (a): Surface Area of Dune as a Function of its Angle (b): The Wind Rose of the Region of our Dataset (c):The Histogram of Dunes Angles.

IV. Conclusion and Future Work

In this paper, we used a computer vision and machine learning approach for the detection, segmentation of sand dunes in high resolution satellite images. Our contribution is an end to end process which starts from the raw high resolution satellite image, which we enhanced, then using a Haar classfier we found dunes candidates, then for each validated candidate, we segmented its contours using SIFT features and an SVM. Moreover, we used an algorithm to measure the allometry of dunes, and finally explored the correlations. Our approach obtained a good accuracy of 77.5%, which we could improve by taking into account dunes which are colliding in future works. Our approach allowed automating the collection of an important amount of data from a high resolution satellite image. The data results were confirmed by field measurement of wind rose. Finally, we found data relationships that we formalized with equations, which desertification experts can rely on to assess the risk of barchans sand dunes and take anticipated actions to protect the populations. Indeed, as smaller dunes tend to move faster, the collected data could identify the dunes having smaller width which may present higher risk.

The overall approach returned satisfactory accuracy results. However, we observed some limitations to our method due to cases of dunes having odd shapes. These deformed dunes are often the result of collision between dunes advancing at different speeds. In future work, we would use more advanced machine learning methods such as Deep Learning for the detection and segmentation of sand dunes in order to improve recall. Also, the use of several high resolution satellite images would allow using a multi temporal analysis of sand dunes.

References

[1] UNCCD: United Nations Convention to Combat Desertification in those countries experiencing serious drought and/or desertification, particularly in Africa. UNCCD explanatory leaflet, UNCCD Secretariat, Bonn (2008).

[2] McKee, E.D., 1979. Introduction to a study of global sand seas. In: McKee, E.D. (Ed.), A Study of Global Sand Seas: United States Geological Survey, Professional Paper, 1052, pp. 3–19.

J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Blumberg, D.G., 1998. Remote sensing of desert dune forms by polarimetric Synthetic Aperture Radar (SAR). Remote Sensing of Environment 65, 204–216.

[4] Vermeesch, P., Drake, N., 2008. Remotely sensed dune celerity and sand flux measurements of the world's fastest barchans (Bodélé, Chad). Geophysical Research Letters 35, L24404.

[5] Bishop, M.A., 2010. Comparative nearest neighbor analysis of megabarchanoid dunes, Ar Rub al Khali sand sea: the application of geographical indices to the understanding of dune field selforganization, maturity and environmental change. Geomorphology 120, 186–194.

[6] Wolfe, S.A., Hugenholtz, C.H., 2009. Barchan dunes stabilized under recent climate warming on the northern Great Plains. Geology 37, 1039–1042.

[7] Scheidt, S., Ramsey, M., Lancaster, N., 2010. Determining soil moisture and sediment availability at White Sands Dune field, New Mexico, from apparent thermal inertia data. Journal of Geophysical Research 115, F02019.

[8] Hansen, C.J., Bourke,M.C., Bridges, N.T., Byrne, S., Colon, C., Diniega, S., Dundas, C., Herkenhoff,K., McEwen, A., Mellon, M., Portyankina, G., Thomas, N., 2011. Seasonal erosion and restoration of Mars' Northern Polar Dunes. Science 331, 575– 578.

[9] Bourke, M.C., Edgett, K.S., Cantor, B.A., 2008. Recent aeolian dune change on Mars. Geomorphology 94, 247–255.

[10] Azzaoui, M. A., Adnani, M., El Belrhiti, H., Chaouki, I. E., and Masmoudi, L.: Detection Of Crescent Sand Dunes Contours In Satellite Images Using An Active Shape Model With A Cascade Classifier, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-4/W12, 17-24, https://doi.org/10.5194/isprs-archives-XLII-4-W12-17-2019, 2019.

[11] Hugenholtz CH, Levin N, Barchyn TE, Baddock M, 2012. Remote sensing and spatial analysis of aeolian sand dunes: a review and outlook. Earth-Science Reviews 111: 319-334.

[12] Flor-Blanco, Germán & Flor, Germán & Pando, Luis. 2012. Evolution of the Salinas-El Espartal and Xagó beach/dune systems in north-western Spain over recent decades: Evidence for responses to natural processes and anthropogenic interventions. Geo-Marine Letters. 33. 10.1007/s00367-012-0301-3.

[13] John A. Richards, Xiuping Jia. 2006. Remote Sensing Digital Image Analysis. 10.1007/3-540-29711-1.

[14] ElSayed Hermas, Sebastien Leprince, Islam Abou El-Magd, Retrieving sand dune movements using sub-pixel correlation of multi-temporal optical remote sensing imagery, northwest Sinai Peninsula, Egypt, Remote Sensing of Environment,Volume 121, 2012, Pages 51-60, ISSN 0034-4257. DOI: 10.1016/j.rse.2012.01.002.

[15] T. Blaschke, Object based image analysis for remote sensing, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 65, Issue 1, 2010, Pages 2-16, ISSN 0924-2716, DOI: 10.1016/j.isprsjprs .2009.06.004.

[16] Schiewe, Jochen. (2012). Segmentation of high-resolution remotely sensed data - Concepts, applications and problems. International Archives of Photogrammetry and Remote Sensing. 34.

[17] Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8926, pp. 613–627. Springer, Heidelberg (2015).

[18] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele and P. Jannin, "Detecting Surgical Tools by Modelling Local Appearance and Global Shape," in IEEE Transactions on Medical Imaging, vol. 34, no. 12, pp. 2603-2617, Dec. 2015. doi: 10.1109/TMI.2015.2450831.

[19] Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. IEEE Trans. Syst.1973, SMC-3, 610–621.

[20] P. D. Blanco , G. I. Metternicht, H. F. del Valle, W. Sione. 2007. Assessment of TERRA-ASTER and RADARSAT imagery for discrimination of dunes in the Valdes peninsula: an object oriented approach Revista de Teledetección. ISSN: 1133-0953. 2007. 28: 87-96.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[22] T. Ojala, M. Pietikäinen, and D. Harwood (1994), "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions", Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR 1994), vol. 1, pp. 582 - 585.

[23] M. A. Azzaoui, M. Adnani, H. El Belrhiti, I. E. Chaouki, L. Masmoudi. 2018. Detection Of Crescent Sand Dunes Contours In Satellite Images Using An Active Shape Model With A Cascade Classifier. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4/W12, 2019.

[24] Lowe, David G. (1999). "Object recognition from local scale-invariant features" (PDF). Proceedings of the International Conference on Computer Vision. 2. pp. 1150–1157. doi:10.1109/ICCV.1999.790410.

[25] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst. 110, 3 (June 2008), 346-359. DOI=http://dx.doi.org/10.1016/j.cviu.2007.09.014.

[26] Darshana Mistry, Asim Banerjee. Comparison of Feature Detection and Matching Approaches: SIFT and SURF. GRD Journals- Global Research and Development Journal for Engineering | Volume 2 | Issue 4 | March 2017 ISSN: 2455-5703.

[27] Alahi, Alexandre & Ortiz, Raphael & Vandergheynst, Pierre. (2012). FREAK: Fast retina keypoint. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 510-517. 10.1109/CVPR.2012.6247715.

[28] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "BRIEF: Binary Robust Independent Elementary Features", 11th European Conference on Computer Vision (ECCV), Heraklion, Crete. LNCS Springer, September 2010.

[29] Stefan Leutenegger, Margarita Chli and Roland Siegwart: BRISK: Binary Robust Invariant Scalable Keypoints. ICCV 2011: 2548-2555.

[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary R. Bradski: ORB: An efficient alternative to SIFT or SURF. ICCV 2011: 2564-2571.

[31] Per-Erik Forssén. Maximally stable colour regions for recognition and matching. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[32] Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. (2012). Learning from data: a short course. [United States]: AMLBook.com.

[33] Li, C.; Wang, J.; Wang, L.; Hu, L.; Gong, P. Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery. Remote Sens. 2014, 6, 964-983.

Dynamic Hand Gesture to Text using Leap Motion

Nur Aliah Nadzirah Jamaludin[1]
Faculty of Computing, University Malaysia of Computer Science and Engineering, Cyberjaya, Selangor, Malaysia

Ong Huey Fang[*2]
School of Information Technology, Monash University Malaysia, Bandar Sunway, Selangor, Malaysia

***Abstract*—This paper presents a prototype for converting dynamic hand gestures to text by using a device called Leap Motion. It is one of the motion tracking technologies, which could be used for recognising hand gestures without the need of wearing any external devices or capturing any images and videos. In this study, five custom dynamic hand gestures of American Sign Language were created with Leap Motion to measure the recognition accuracy of the proposed prototype using the Geometric Template Matching, Artificial Neural Network, and Cross-Correlation algorithms. The experimental results showed that the prototype achieved recognition accuracy of more than 90% in the training phase and about 60% in the testing phase.**

Keywords*—Dynamic hand gesture; leap motion; American sign language; artificial neural network; cross-correlation; geometric template matching

I. Introduction

According to the Cambridge Dictionary, a sign language is a system of hand and body movements representing words, used by and to people who cannot hear or talk. However, not everyone knows and understand sign languages, and they are normally limited to the deaf and mute people and their families. Therefore, assistive devices or systems are required to aid their communication with others or even with machines. Furthermore, sign language is considered a powerful way for human-computer interaction as it is natural, ubiquitous, and meaningful [1, 2].

In general, hand gestures fall into two categories: 1) static; and 2) dynamic. A static hand gesture refers to a still and stable shape of the hand, while a dynamic hand gesture is movements of the hand to express a specific interaction [3]. Although dynamic hand gestures are more practical for real-time application, they require more computational complexity in processing input signals and in building recognition algorithms [4, 5]. The main problems in developing a dynamic gesture recognition system are in detecting and tracking hand shapes and hand movements, which may be complicated with other challenges such as occlusion between fingers, various trajectory, speed and amplitude of hand movements, background environment, and different styles of sign languages [5, 6].

Three technologies can be used to assist in the acquisition of hand gestures: 1) gloved-based; 2) vision-based; and 3) depth-based [3, 7]. Gloved-based technology or data glove uses wearable sensors that typically mounted on fingers and wrist to capture the position and movement of the hand. The PowerGlove is the first commercially available data glove used by the Nintendo Entertainment System as a gaming controller [8, 9]. Other examples of data gloves are such as the Sayre Glove, CyberGloves, and Polhemus. Nonetheless, the use of data gloves is costly and inconvenience for users to wear, as they are typically wired and connected to computers [10]. In contrast, vision-based technology is more prevalent in hand gesture recognition with lower cost, noncontact, natural and intuitive [10, 11, 17]. Vision-based approaches record hand gestures by using a camera, and the captured image or video is processed and analysed with computer vision techniques. However, the remaining challenges of vision-based approaches are mainly in background, lighting, noise, and camera [11]. Besides, it usually captured a large amount of data, thus depends heavily on features preprocessing to provide better accuracy and computation rate [5].

The introduction of low-cost depth-based technology, such as Microsoft's Kinect and Leap Motion, has opened the possibility for informative and accurate dynamic gesture recognition. This approach can provide 3D information of the real world with depth information. A depth image contains not only the human body region but also the background. The user is supposed to stay close to the camera and occupies a significant portion of the camera's field of view [12]. Microsoft's Kinect is a depth camera that acquires gestures with rich depth information and has demonstrated a notable performance of light sensitivity in gesture recognition [18]. However, there are still unresolved issues regarding gesture finding and segmenting with Kinect sensors [10]. Different from the Kinect, the Leap Motion controller is a device explicitly targeted to hand gesture recognition and can directly compute the position of the fingertips and the hand orientation with better accuracy [13]. The device is smaller in size and comes with built-in recognition algorithms.

Although Leap Motion has been adopted in previous works on sign language recognition, most of them focused on static gestures instead of dynamic gestures [14, 15, 16]. The study in [16] showed that the results in static alphabet recognition were low using the built-in algorithms, namely the Geometric Template Matching (GTM), Artificial Neural Network (ANN), and Cross-Correlation (CC). Where the highest average accuracy achieved was only 52.56% with the GTM algorithm. Nevertheless, this study attempted to test further the ability of Leap Motion in recognising and converting dynamic hand gestures to text with by proposing a dynamic hand gesture recognition prototype for American Sign Language (ASL). The proposed prototype created five custom dynamic hand gestures of ASL and tested their accuracy with Leap Motion's built-in recognition algorithms.

The remaining content of this paper is organised as follows: Section II introduces the Leap Motion Controller and Leap Trainer; Section III discusses the design for the proposed

*Corresponding Author.

prototype; Section IV discusses on the experimental results, and finally, the conclusion was given in Section V.

II. Leap Motion

This section explains the overall process in building a dynamic hand gestures recogniser by using Leap Motion. It includes the illustration of workflow for Leap Motion Controller and Leap Trainer.

A. Leap Motion Controller

Leap Motion is a small device connected to USB and can recognise hand motions or gesture accurately when no occlusion happens without the need of wearing any gloves or instruments. The device consists of three Infrared (IR) Light emitters and two cameras. The cameras received the IR lights and track hand gesture up to 200 frames per second and 150° field of view with approximately 8 cubic feet of interactive 3D space [19].

The Leap Motion Controller has a wide range of application. For example, this device had been used for stoking rehabilitation by people from The Intelligent Computer Tutoring Group at the University of Canterbury [20]. The tracking data of the Leap Motion Controller, which contains the palm and fingers' position, direction, velocity can be accessed using its Software Development Kit.

B. Leap Trainer

Leap trainer is one of the main application used to create a custom gesture for Leap Motion Controller. As for now, Leap Motion only supports four gestures, which are a swipe, circle, screen tap and key tap. Circle and swipe gestures are continuous, and these objects can have a state of start, update, and stop. While the screen tap gesture is a discrete gesture [21], and it always has a stop state. However, these four gestures are not enough to cover all the hand gestures available in the ASL. Each sign language has a different sign to convey the same meaning in different languages. There are two ways of creating custom gestures using Leap Motion device, which are: 1) writing code to recognise a custom gesture, and 2) allowing the device to learn a gesture which can be recognised.

Fig. 1 below shows the workflow of how a Leap Trainer trained the custom gestures by using the Leap Motion Controller. The description for each component in the workflow is as below:

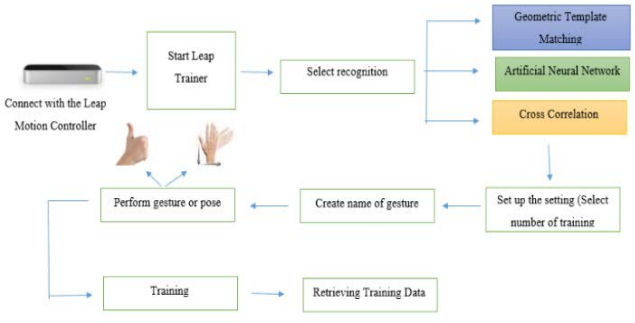


Fig. 1. The Workflow of a Leap Motion Controller.

- Connect with the Leap Motion Controller: To run the Leap Trainer, the user must connect to the Leap Motion Controller.
- Start Leap Trainer: If the device is connected, it detects user hand movements on top of the Leap Motion device.
- Select recognition: Before start creating custom gesture, the user may need to select an algorithm they want to use. The choices are Geometric Template Matching, Artificial Neural Network and Cross-Correlation.
- Set up the setting: When the user clicks the Options button, they can modify settings that have been provided. User is advised to change the number of training gesture since it helps to increase the recognition of the custom hand gestures.
- Create name of the gesture: After done with the setting part, the user can put a name at the textbox available at the top left to start creating custom hand gestures.
- Perform gesture or pose: User can perform any motions, either gesture or pose. A gesture means the hand gesture is intended to change over a period of time, whereas pose is observed at the spurt of time.
- Training: After a few times of training gesture (according to the number of training gesture that has been set up), the application now try to learn the gesture and user can start repeating the gesture for training the algorithm to learn the hand gesture.
- Retrieving training data: Lastly, after the algorithm had learned the hand gesture, the user can retrieve the data and save it into the database.

III. Proposed Prototype

This section presents designs of the proposed prototype, including the hand gesture recognition process, the user interface design and the database design.

A. Design Workflow

Fig. 2 shows the design workflow used in this paper with two main phases. The first phase is the training phase. The training phase is part of the process to train the data, extract data and save it into the database. The data need to be trained to get the best result during the translation of the motion of hand gesture to text. The second phase is the real-time phase, where the testing of hand gesture recognition is done real-time with the techniques of GTM, ANN, and CC. This phase will try to match a gesture with the most suitable text and give an output via the prototype.

B. Sign Trainer

A Sign Trainer page allows the user to create a custom gesture, to train hand gestures and store data into the database. The creation of custom gestures can be recognised using three different techniques which are the GTM, ANN, and CC. This study used several specific programming languages and scripting, which are JavaScript, HTML, CSS and image files. Python files also created to store custom gesture data into the

database. Fig. 3 shows the workflow and architecture of a Sign Trainer. The main parts are:

- Connect with Leap Motion Controller
- Start creating hand gestures
- Store custom gesture data
- Load hand gesture in the database.

The user interface for Sign Trainer page is shown in Fig. 4, which allows users to add custom gestures. This custom gesture can be created according to three different classifiers which are the GTM, ANN, and CC. Before doing the customisation of hand gestures, users are allowed to change the settings listed and choose a recognition algorithm when they clicked on the Options button. The custom gestures data will produce in a JSON file.

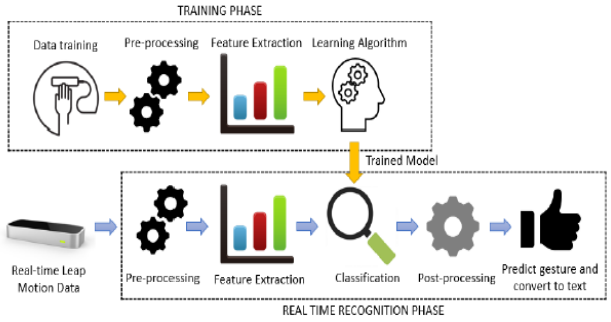


Fig. 2. The Workflow of Hand Gesture Recognition.

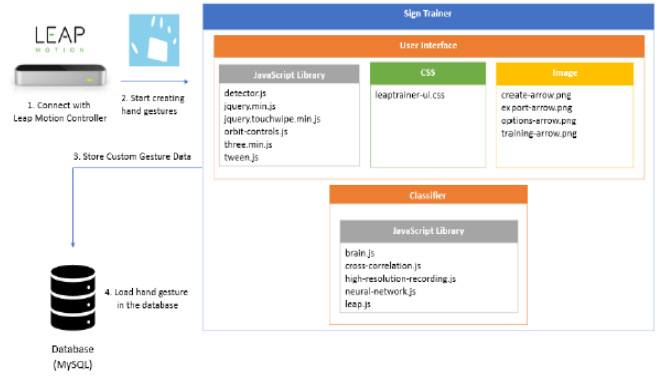


Fig. 3. Architecture and Workflow of Sign Trainer.

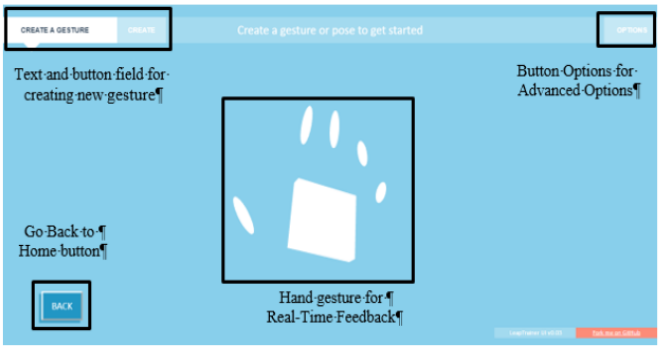


Fig. 4. Sign Trainer Page.

There are some other options provided on this page, as shown in Fig. 5. These options are the default implementations of an algorithm and may be removed or redundant in sub-classes of the Sign Trainer controller. The Option button also let the user choose one of the techniques as the recognition strategy, which are the GM, ANN or CC.

C. Sign Translator

Sign Translator is the page to convert a hand gesture into corresponding text for ASL. This page first connects to the MySQL database all the gesture data available. If the user wants to use this application, they first need to choose the three main algorithms that have been implemented in the Sign Trainer. The workflow and architecture, as shown in Fig. 6, and the main parts of Sign Translator are:

- Connect with Leap Motion Controller
- Retrieve hand gesture data from the database
- Load the stored data
- Output result

Fig. 7 shows the page that intended to convert the sign language into text. The user can perform any sign while connecting to the Leap Motion Controller. Then, after one of the algorithms been chosen, a pop-up alert will appear and shows that the stored data already loaded to the application. Automatically, the application will later look for a match of the current hand gesture with the stored data. If a match occurs, the result with percentage of matching will appear on top of the screen, as shown in Fig. 8.

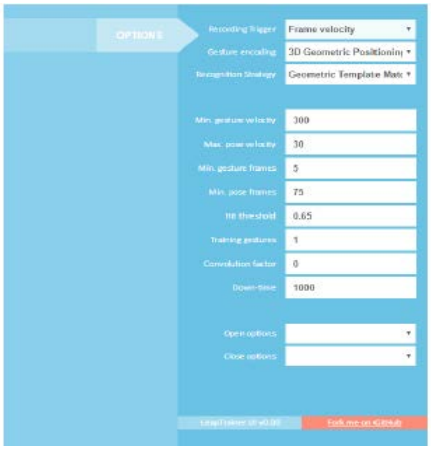


Fig. 5. Options in Sign Trainer Page

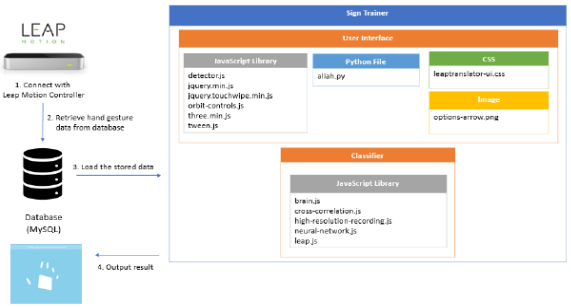


Fig. 6. Architecture and Workflow of Sign Translator Page.

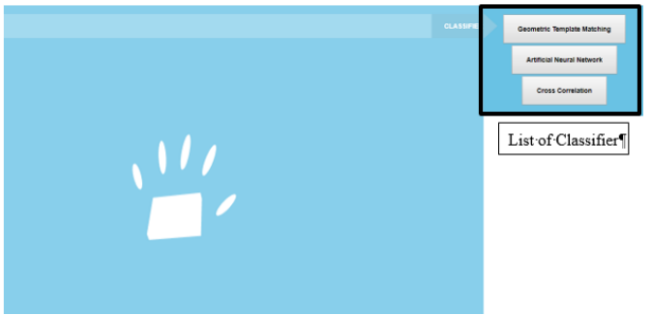


Fig. 7. List of Classifiers in Sign Translator Page.

Fig. 8. List of Hand Gestures in Database.

IV. Experiment

This section presents the experiments conducted using the proposed prototype that runs on the Leap Motion Controller. The results of in recognising dynamic hand gesture into text are also discussed.

A. Experiment Setup

This study performed two experiments to test the accuracy of the training phase and the testing phase. The tests were held to find out the performance of the proposed prototype. The accuracy value is calculated by using the following formula:

$$Accuracy = \frac{(\text{Number of correct signs})}{(\text{Number of signs performed})} \quad (1)$$

Five datasets were used in this study. For each gesture, the number of tests conducted was two times, and the average accuracy was calculated. Fig. 9 shows the dynamic hand gestures that were used in this study, namely the sign language of ‘YES’, ‘NO’, ‘CIRCLE’, ‘BOOK’, and ‘HOUSE’.

B. Training Results

Table I to Table III shows the results of the training phase that had been conducted on the Sign Training page. The results show that the ANN technique achieved the highest accuracy in recognising the five dynamic sign language.

The bar chart shown in Fig. 10 compared the results obtained from three different algorithms, which are GTM, ANN, and CC. The comparison shows that ANN always gets the highest accuracy in every type of gesture. Besides, the results suggest all the three techniques struggle in recognising particular signs such as 'HOUSE', 'NO', and 'BOOK'. The CC technique was not able to detect the hand gesture of 'HOUSE', although given several trials.

Fig. 9. Dynamic Hand Gestures Tested in this Study.

TABLE. I. Results of Dynamic Hand Gestures Recognition using Geometric Template Matching

Gesture	No. of training gesture	Threshold	Accuracy (%)
YES	2	0.65	96.33
NO	2	0.65	67.33
CIRCLE	2	0.65	48.33
BOOK	2	0.65	72.00
HOUSE	2	0.65	23.33

TABLE. II. Results of Dynamic Hand Gestures Recognition using Artificial Neural Network

Gesture	No. of training gesture	Threshold	Accuracy (%)
YES	2	0.65	98.00
NO	2	0.65	96.00
CIRCLE	2	0.65	91.33
BOOK	2	0.65	95.00
HOUSE	2	0.65	95.33

TABLE. III. Results of Dynamic Hand Gestures Recognition using Cross-Correlation

Gesture	No. of training gesture	Threshold	Accuracy (%)
YES	2	0.65	83.67
NO	2	0.65	23.33
CIRCLE	2	0.65	70.67
BOOK	2	0.65	43.33
HOUSE	2	0.65	0.00

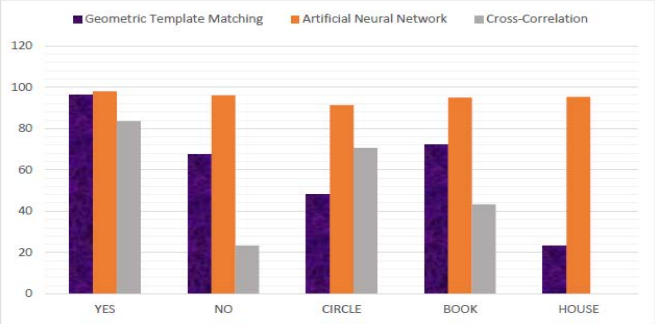


Fig. 10. Comparison of Recognition Training Results for Geometric Template Matching, Artificial Neural Network and Cross-Correlation.

C. Testing Results

For the testing results, only GMM was implemented in the experiment. Table IV shows the results of GMM in converting hand gesture to text. The results of the cumulative average accuracy of translating hand gesture to a text are 60.00%, as shown in Fig. 11. This accuracy can be a good start for more improvement in future work.

TABLE. IV. RESULTS OF TESTING USING GEOMETRIC TEMPLATE MATCHING

Gesture	Test #1	Test #2	Test #3	Average
YES	1	1	1	100.00%
NO	1	1	0	66.67%
CIRCLE	1	0	1	66.67%
BOOK	0	1	0	33.33%
HOUSE	1	0	0	33.33%
Cumulative				60.00%

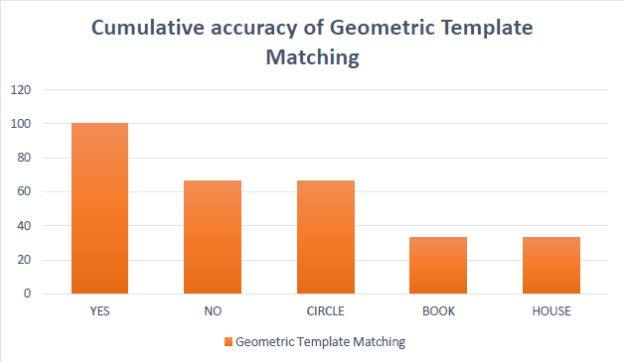


Fig. 11. Cumulative Accuracy of Geometric Template Matching.

V. CONCLUSION

This paper proposed a prototype that can convert dynamic hand gestures into text by using one of the depth-based technologies known as Leap Motion. The objective of the prototype is to be a medium of communication between deaf and mute by using a lightweight device that is capable of detecting and tracking hand motions. However, the experimental results suggest that Leap Motion was unable to track finger position precisely, especially when using both hands to do the 'HOUSE' gesture, and when fingers are in the first position. Most of the time, the hand gesture of 'YES' is the fastest and easiest to be detected. This gives hints on the tracking capability of Leap Motion and the recognition algorithms required for improvement in tracking similar dynamic hand gestures more precisely and accurately.

Nonetheless, with Leap Motion technology, it has become possible for converting hand gesture to text to recognise various actions that can be interpreted into words and understandable for the deaf and mute people. For future work, improvement can be made in term of performance and accuracy, where better gesture recognition algorithms are required not only to recognise dynamic gestures that can be translated into a word, but also to recognise continuous dynamic gestures that can be translated into sentences. Moreover, better ways of capturing tracking data, creating custom gestures and recognising gestures are needed for Leap Motion in building an effective dynamic hand gesture to text system.

REFERENCES

[1] P. Premaratne, S. Yang, Z. Zhou, and N. Bandara, "Dynamic Hand Gesture Recognition Framework," in Intelligent Computing Methodologies, Cham, 2014, vol. 8589, pp. 834-845: Springer International Publishing.

[2] H. Cheng, L. Yang, and Z. Liu, "Survey on 3D Hand Gesture Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 9, pp. 1659-1673, 2016.

[3] J. S. Sonkusare, N. B. Chopade, R. Sor, and S. L. Tade, "A Review on Hand Gesture Recognition System," in 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 790-794.

[4] J. Jia, G. Tu, X. Deng, C. Zhao, and W. Yi, "Real-time hand gestures system based on leap motion," Concurrency and Computation: Practice and Experience, vol. 31, no. 10, pp. e4898, 2019.

[5] Z.-j. Liang, S.-b. Liao, and B.-z. Hu, "3D Convolutional Neural Networks for Dynamic Sign Language Recognition," The Computer Journal, vol. 61, no. 11, pp. 1724-1736, 2018.

[6] X. Zhang, J. Wang, X. Wang, and X. Ma, "Improvement of Dynamic Hand Gesture Recognition Based on HMM Algorithm," in 2016 International Conference on Information System and Artificial Intelligence (ISAI), 2016, pp. 401-406.

[7] R. R. Itkarkar and A. V. Nandi, "A survey of 2D and 3D imaging used in hand gesture recognition for human-computer interaction (HCI)," in 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2016, pp. 188-193.

[8] M. Kadous, "Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language," 1970.

[9] C. Preetham, G. Ramakrishnan, S. Kumar, A. Tamse, and N. Krishnapura, "Hand Talk-Implementation of a Gesture Recognising Glove," in 2013 Texas Instruments India Educators' Conference, 2013, pp. 328-331.

[10] X. Ma and J. Peng, "Kinect Sensor-Based Long-Distance Hand Gesture Recognition and Fingertip Detection with Depth Information," Journal of Sensors, vol. 2018, pp. 9, 2018.

[11] Y. Zhu, Z. Yang, and B. Yuan, "Vision Based Hand Gesture Recognition," in 2013 International Conference on Service Sciences (ICSS), 2013, pp. 260-265.

[12] C. Wu and C. H. Lin, "Depth-based hand gesture recognition for home appliance control," in 2013 IEEE International Symposium on Consumer Electronics (ISCE), 2013, pp. 279-280.

[13] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 1565-1569.

[14] A. Vaitkevičius, M. Taroza, T. Blazauskas, R. Damasevicius, R. Maskeliunas, and M. Woźniak, "Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion," Applied Sciences, vol. 9, p. 445, 2019.

[15] T. W. Chong and B.-G. Lee, "American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach," (in eng), Sensors (Basel, Switzerland), vol. 18, no. 10, p. 3554, 2018.

[16] F. Khan, H. Ong, and N. Bahar, "A Sign Language to Text Converter Using Leap Motion," International Journal on Advanced Science, Engineering and Information Technology, vol. 6, p. 1089, 2016.

[17] B. Garcia, S. A. Viesca (2014) Real-time American Sign Language Recognition with Convolutional Neural Networks.

[18] D. H. Vo, H. H. Huynh, P. M. Doan and J. Meunier, “Dynamic Gesture Classification for Vietnamese Sign Language Recognition” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 3, p 2017.

[19] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," (in eng), Sensors (Basel, Switzerland), vol. 13, no. 5, pp. 6380-6393, 2013.

[20] A. Bracegirdle, T. Mitrovic, and M. Mathews (2014). Investigating the usability of the leap motion controller. Gesture-based interaction with a 3d virtual environment.

[21] R. O'Leary (2013). Leaptrainer.js. Retrieved from Github: https://github.com/roboleary/LeapTrainer.js.

Investigating Factors Affecting Knowledge Management Practices in Public Sectors

Subashini Ganapathy[1], Zulkefli Mansor[2], Kamsuriah Ahmad[3]
Faculty of Information Science and Technology, National University of Malaysia Bangi, Selangor, Malaysia

***Abstract*—Knowledge Management (KM) is a systematic approach in creating, sharing, using and managing information effectively sustain knowledge regardless public or private organizations. It helps organizations to make better decision making in order to achieve the goals and increase the productivity. However, many public organizations are still facing challenges to adopt knowledge management practices compared to private organization due to lack of awareness. They are not aware of the influenced factors such as people, process, and technology. Therefore, this paper identifies influencing factors that contributed to the successful KM practices in public sectors. This study employs quantitative approaches by distributing a set of questionnaires to 83 IT practitioners in public organizations. 63 returned responses were analyzed using the Rasch Measurement Model. The findings indicated that there is a lack of participation amongst the staff in practicing efficient knowledge management due to they are still not ready to accept changes to the new system, lack of exposure and behavior. In addition, looking at critical success factor such as on the human resources (HR), there is a lack of encouragement such as rewards and recognition given to employees who practices KM in the organization. As a result, this paper highlights the most influential factors for effective knowledge management practices in terms of people, process and technology. We hope that the results can be used as a guideline to rectify the challenges in KM practices especially in the public organizations.**

***Keywords*—Critical success factors; influencing factors; knowledge management; public sector**

I. Introduction

Knowledge Management (KM) is a systematic effort to encourage and facilitate the flow of knowledge within the organization to speed up organizational learning [1]. Knowledge-based information plays a very important role in the public sector as well as in the private sector. Public administration especially marked as an important element in any nation because of its contribution to economic growth and development of a country as compared. Basically, public organizations showing its tendencies to culture and challenged when it comes to introduction of new management initiatives. Implementation of KM in the public sector is still not clear even longstanding and there is resistance to change, particularly from the perspective of the individual [2].

To be more precise, there is resistance to change, particularly from the perspective of the individual to any related KM practices or system. As concerned, there is evidence showed that more studies done in private sectors with a small contribution towards the public sector [3]. It is also found that [1] research on public sector is relatively less compared to the private sector.

Successful companies are those who are consistently practicing, creating, sharing, storing and disseminating source of knowledge properly. These success rates become more precise with the adaptation of new technologies and products [4]. Considering in Malaysia, various studies have been conducted to determine the factors that influence the practice of KM. However, understanding KM in Malaysian context is still difficult due to its limited studies [5]. Among the factors that were studied in terms of human resources, confidence of individuals and organizational culture, technology, individual performance, awareness and understanding of people, leadership and strategy, the involvement of top management and external factors related to external influences. However, most of the studies have related to factors affecting the KM in the private sector compared to the public sectors.

A study by [2] highlighted that, the practice of KM in the public sector is still weak. Secondly, it is found that there is less appropriate criteria are examined to determine the factors that influence KM practices the most. Taking this into consideration, this current study is to further explore the KM practices in Malaysia especially in the public sector. And a thorough study will be done to get to know the most influenced factors that contributing to the success of KM in the public sector today so that it becomes a guide to all parties.

II. Related Works

The concept of knowledge management (KM) is an important activity to be carried out effectively by the organization. We reside in a world of rapid change driven by globalization, the knowledge-based economy combined with the development of information, communication and technology (ICT). This transformation, however, not solely poses some challenges, however conjointly offers opportunities for each private and public sector. KM focuses on the effective ways to manage knowledge to acquire competitive advantage [6] and marked as an important aspect for organizations to have efficient knowledge resources managing round the world. Most of the huge companies in the private sector have been actively taking initiatives to adopt and practice KM along with tool and techniques. Determination in the public sector in terms of KM initiatives is less compared to the private sector [7].

The importance of managing knowledge is classified to whom to share, what is to be shared, how it is to be shared, and ultimately sharing and using it. Managing information or

Sponsored by Faculty of Information Science and Technology, National University of Malaysia (UKM)

knowledge fabricates benefits when the knowledge is shared, used and reused. The benefits of KM practice can be seen effectively when there is an atmosphere of motivation and trust for people to share and use knowledge, when there are systematic processes to find and create knowledge, and, when needed. This process is supported by adequate technology to store and make knowledge relatively simple to find and share. With regards to this, understanding the KM in the Malaysian context is difficult because the study was limited [5]. Knowledge management (KM) was born in accordance with the initial development of the "Multimedia Super Corridor (MSC)" in the 1990s in Malaysia.

Most of the public and private sector organizations have already adapted the concept of KM in Malaysia. Draft regulation regarding KM has been in practice in most of the private sector than in the public sector, such as MIMOS, Siemens, Bank Negara Malaysia, Nokia Malaysia, Telekom, Tenaga Nasional and Petronas [8]. In fact, there are several KM systems were used as a support for the creation, sharing, storage and dissemination of knowledge. Malaysia, like other countries that practice KM faced various obstacles and challenge. And those challenges need to be addressed and further research to solve effectively. In that case, various studies have been conducted in both public and private sector organizations. Consequently, researchers have identified several obstacles that are common in both sectors. Organizational changes at the beginning is difficult, the willingness of the organization should be assessed prior to proceeding with any initiative [9].

One of the major obstacles in practicing KM is the behavior of the workers themselves, and one of the sub-factor that influence behavior is that, there's a lack of sharing culture with colleagues. On the other hand, there are some employees who are reluctant to ask their colleagues for information and knowledge. The reason is that, they might be thinking that the colleagues will look down if they ask so because of low knowledge [10]. Cultural exchange is also a challenge faced by most organizations. KM successful practice requires an open culture that encourages knowledge sharing. Individual factors play as an important role [11] in the process of cultural change [9].

In another point of view, evolution of Information Communication Technology (ICT) turns into new challenge in KM implementation. Speedy growth especially in ICT produces large amounts of information around the world requires organizations to align their knowledge and technology in this direction [12]. Other than that, another critical success factor such as lack of top management support also found to be challenged in the use and practice of KM. With the full support of the management, organization can even provide a favorable environment [13] that will encourage workers to have more commitment and will promote knowledge creating practice to form a better organizational performance. Conversely, without a proper involvement and supports from top management would also cause inefficient practices [14].

Based on the current literature review, limitations of technology are found out as an challenge by some researchers because of many processes require information technology (IT) to KM practices. Infrastructure information systems rely on the use of network technology such as the internet, lotus notes and communications systems for transferring knowledge. This is a major contributor to the application and dissemination of knowledge [8]. It can be concluded that the technology is important for the efficient of overall KM regulation system.

Apart from that, critical success factor such as organizational structure was also said to be one of the factors which influence KM practices [15]. Organizational structure plays important role in the knowledge conversion. Organization need to have favorable structure [16] in order to implement KM successfully. Team size most likely affect the performance of KM practices in the overall. Bigger team size may contribute to efficient knowledge sharing performance. Roles and responsibility of the leader were also playing roles in promoting good leadership styles in giving guidance for employees to practice KM effectively.

III. Research Methodology

This study focuses on middle management and top management, including their processes and activities regarding KM practices. A quantitative approach was conducted through a questionnaire survey. One of the focuses of this paper is on the outcomes of the questionnaire survey after analysis has completed. Therefore, the questions were categorized into three main aspects which are people, process, and technology. The questionnaires were distributed to the public organization as per targeted sample. In this case, five ministries have been chosen randomly as a target to study the factors influence the public sector. In addition, an online version of the questionnaires was distributed. Respondents who responded were the quality managers, information technology personnel and personnel involved indirectly in the KM in the ministry.

The questionnaires will be distributed to check the KM practices in the organization through survey. The questionnaires were developed based on the following key process area as per shown in Fig. 1.

The questionnaires were developed based on the following key process area as per shown in Fig. 1. The questionnaires will be distributed to check the KM practices in the organization through survey. There is a total of 24 items developed based on the people, technology and process key area. People criteria were assessed from organizational culture in terms on the employee's (trust, chances, confident, willingness to accept changes, commitment, participation, and teamwork skills), organizational structure (team size, community of practice, leadership roles) and human resource (resources, time management, promotion, and rewards) perspectives. While, technology criteria, were evaluated based on the technology infrastructure that supports the practices and Information Technology (IT) tools that being used to perform KM practices. Process criteria scoped to its knowledge creation, storing, sharing and knowledge application (dissemination).

The research question (RQ) for the study is:

RQ1: What are the most factors influencing KM practices in the organization?

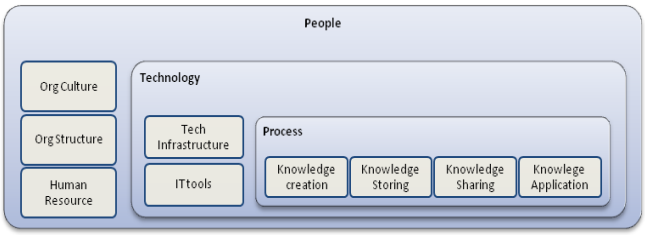


Fig. 1. Questionnaires Criteria.

The data retrieved will be analyzed using Rasch model. Rasch statistical analysis used because of its ideal statistical techniques that allow the scales to be modified [17] by removing unreliable items. In that case, WINSTEPS software used to analyze the data. INFIT and outfit statistics were used to measure the fit of the items in the Rasch model. Possible attributes defined the probability of response in an item [18]. The Rasch model is expected to measure the KM practices based on the influencing factors. The similar approach will be used to evaluate the KM practices in the organizations in this study.

A. Normality Test

This normality test is used to see the distribution of the sample size. This is important to see the sample collected falls within the range and its Skewness as well as Kurtosis. According to [19], values that fall within the range of -3 to +3 for the Kurtosis test, and -2 to +2 for the Skewness test is considered within the normal range. Therefore, the normality of this study is normal. Table II shows the mean, standard deviation, Skewness and Kurtosis of the sample gathered.

B. Reliability Test

Data were analyzed using Winstep Rasch analysis software. To test the internal consistency of the questionnaires, the results of the analysis showed that the value of Cronbach's Alpha (α) is 0.68 as shown in Table I below. The value of Cronbach's alpha for all constructs must be above 0.6. Ideally the Cronbach's alpha should be above 0.7. Since the current value is within 0.65 and 0.70, it is minimally acceptable [20].

For summary statistic, INFIT ZSTD for Person as well as Item is close to 0 and MNSQ are close to 1.0, so the data gathered to consider fitting the model. An item MNSQ fit statistic of 1.0 meaning that the item was perfectly fit the expected model. Table II shows the fits statistic for data fits. INFIT ZSTD for the values for all items are lesser than 1.5, so the all items are considered acceptable in this study. Parameter level ranges between 0.5-1.5 as shown in Table II considered productive for measurement hence it is reliable.

TABLE. I. SUMMARY STATISTICS

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E	INFIT		OUTFIT	
					MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
MEAN	87.0	25.0	0.48	0.23	1.02	-.27	1.01	-.24
SEM	1.0	.0	0.05	.00	.07	.27	.07	.26
P.SD	7.8	.0	0.43	.01	.58	2.12	.55	2.03
S.SD	7.8	.0	0.43	.01	.58	2.14	.55	2.05
MAX	103.0	25.0	1.45	.27	2.38	3.96	2.26	3.65
MIN	71.0	25.0	-.32	.21	.20	-4.41	.20	-4.45
REAL RMSE .26 TRUE SD .33 SEPARATION 1.27 PERSON RELIABILITY .62 MODEL RMSE .23 TRUE SD .36 SEPARATION 1.52 PERSON RELIABILITY .70 S.E. OF PERSON MEAN = .05								
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00 CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .68 SEM = 4.39								

TABLE. II. FIT STATISTICS

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E	INFIT		OUTFIT	
					MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
MEAN	221.0	63.0	0.00	0.15	1.01	.08	1.02	.13
SEM	5.1	.0	0.11	.00	.03	.14	.03	.15
P.SD	24.7	.0	0.55	.01	.12	.67	.13	.70
S.SD	25.2	.0	0.56	.01	.12	.69	.13	.72
MAX	269.0	63.0	0.92	.18	1.32	1.70	1.34	1.82
MIN.	177.0	63.0	-1.19	.14	.85	-.90	.81	-1.08
REAL RMSE .15 TRUE SD .52 SEPARATION 3.40 ITEM RELIABILITY .92 MODEL RMSE .15 TRUE SD .53 SEPARATION 3.50 ITEM RELIABILITY .92 S.E. OF ITEM MEAN = .11								
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00 UMEAN = .0000 USCALE = 1.0000								

IV. Results and Findings

This study used five point Likert scale (strongly disagree, disagree, neutral, agree and strongly agree) type of questions and the respondent have to rate the statement from strongly disagree to strongly agree scaling. A survey was distributed to 83 respondents and returned by 63 with the yielding rate of 76.8%. Five ministries responded. The entire ministry is under the category of 100-500 employees. Table III below shows the demographic profile of the respondents.

In this study, there are a total of 28 respondents which consist of (44.4%) female respondents and 35 (55.6%) male respondents. The percentage shows that the male respondents are much higher than the female respondents. Basically, most of the respondents are from the Information Technology department, which compromise 44.4% and 33.3% of human resource department. The rest of the unit were also indirectly play roles in knowledge management practices since they are also in charge of managing knowledge and information on the organization in particular division or department. There are respondents from six departments such as finance, IT, accounts, human resource and the other two from the top management department. All the participants were from middle level as well as manager's post. These profiles of participants provide valuable information about the context in which, the research findings are applicable. From the response, 28 employees are from the Information Technology (IT) department. Most of the employees who responded the survey are from the IT post (44.4%) such as secretary, secretary, assistant, IT officer and IT assistants. These four positions basically play roles in information management and indirectly involved in terms of KM practices in their organization. Table III above represents the actual information about the positions held by the participants in the five ministries from eight departments.

TABLE. III. Demographic Profile of Respondents (N=63)

Variable	Category	Frequency	Percentage
Gender	Male	35	55.6%
	Female	28	44.4%
	Total	**63**	**100.0%**
Age group	25 to 29	12	19.0%
	30 to 34	19	30.2%
	35 to 39	11	17.5%
	40 to 44	5	7.9%
	45 to 49	4	6.3%
	50 & above	12	19.0%
	Total	**63**	**100.0%**
Ministry	Ministry of Works	29	46.0%
	Housing & Local Government	8	12.7%
	Federal & Regional	7	11.1%
	Science, Technology & Innovation	11	17.5%
	Rural & Regional Development	8	12.7%
	Total	**63**	**100.0%**
Year of Experience in Ministry	1 – 3 years	18	28.6%
	3 – 5 years	17	27.0%
	5 – 7 years	3	4.8%
	7 – 10 years	13	20.6%
	10 – 14 years	8	12.7%
	More than 15 years	4	6.3%
	Total	**63**	**100.0%**
Position	Secretary	3	4.8%
	Secretary Assistant	12	19.0%
	Chief IT Officer	4	6.3%
	IT Officer	15	23.8%
	IT Assistant	7	11.1%
	Administrative Officer	11	17.5%
	Administrative Assistant	11	17.5%
	Total	**63**	**100.0%**
Department	Secretary Chief Office	2	3.2%
	Deputy Secretary Chief Office	2	3.2%
	Information Technology	28	44.4%
	Account	5	7.9%
	Finance	5	7.9%
	Human Resource	21	33.3%
	Total	**63**	**100.0%**

Most of the staff with the percentage of (30.2%) are from 30 to 34 years of age category and 19.0% is in 25 to 29 years of age category. It implies that the young generation is in the managerial positions. Therefore, for them, it is easy to manage and capture the knowledge management with the world trend. Besides that, a number of 44.4% of the respondent has more than 5 years of experience in their organization. In that figure, almost 20.6% of them have 7 to 10 years of experience. Therefore, the results indicate that the respondents have enough experience and knowledge about their position and practices. As a result, respondent is aware of the management structure of their organization. In fact, implementing the KM initiatives in these organizations supposed to be easier because of the employee's long experience in the ministry. The reason why KM remains weak is to be further explored.

In order to know the opinion towards KM, the respondents were asked whether they know about KM in general at the first place. This question was to ensure that the respondents have some sort of awareness or grasp of KM concepts. Almost 70.6% have answered that they have heard of this concept and remaining 29.4% have not heard about it. When asked about whether they are aware of KM practices in their organization, half of the respondents are aware of the practices, while 29.0% are not aware, and the remaining respondents are not sure whether the practices are there in their organization. Similarly, the perception on the level of KM practices in their organization is still puzzled from the respondent's perspective. Almost 12 respondents (35.3%) said that they are not sure where their KM practices are up to. Almost 20.6% of the respondents are not sure about the level of KM practices in their organizations. This is due to lack of individual commitment to the organization. The other reason could be that they are not given any awareness on the importance of KM.

A. People

In this section, the respondent was asked to agree with the statement in terms of people or individual's perspective in the organization. The responses are quite moderate and fall into neutral to agree level. The respondents were asked to rate their level of agreement based on the organizational culture, organizational structure, and human resource (HR) perspective and how these factors influence their KM practices.

1) Organizational culture: Organizational factors were evaluated based on the opportunity, trust and encouragement given to participate in KM practices. Almost 47.1% are agreeing with the statement and this showed that they are given some sort of motivation as well as encouragement to participate in KM. Fig. 2 below shows the statistic on the level of agreement in terms of trust, encouragement and opportunity given in practicing KM.

In another view, when asked about the radius of a participation sub-factor in KM practices, most of the respondent agrees (35.3%) that there is less participation inside the organization, 5.9% strongly agree with the statement. Almost 20.6% are not agreeing with the statement. It is found that, mostly the employees are not willing to accept changes of the system. Additionally, in terms of team cooperation, most of them (28 respondent) are agreed that there is teamwork in solving certain issues in the organization related to KM practices.

2) Organizational structure: On the other hand, the respondent was asked to rate the agreement on the size of the group, community of practice and roles and responsibility related to KM practices. The respondent was asked whether the bigger size group plays an important role in knowledge sharing practices in the organization, and most of them (44.1%) said that group size does give impact to knowledge sharing. The bigger group promotes positive knowledge sharing. Only 3 people (8.8%) don't agree with the statement. Apart from this, in terms of practicing community of practice (COP), most respondents answered in a moderate manner which is 41.2%. Almost 29.4% have agreed with the statement that the roles and responsibilities of knowledge officer don't practice the KM practices.

3) Human Resource (HR): Looking at the human resource management (HRM) perspectives, management actually provides resources and time management to enroll in the learning and sharing exercises for motivation. Promotion or appraisal is given to their work commitment. However, when asked about providing rewards and recognition of KM related activities. The response is just moderate (50%) in terms of giving recognition. From this, it can be concluded that the management is less giving the reward to the employee who actively practice KM. Without proper recognition, the level of commitment from the employee towards KM practices indirectly effects the overall participations in the organization.

B. Technology

In the perception of technology perspective, almost 50% of the respondents were agreed that the organization has cutting-edge Information Communication Technology (ICT) infrastructures that support KM initiative at first place. E-mail was used as a primary resource for knowledge sharing. Almost 41.2% respondent agreed that they are using E-mail in practicing KM. The employees are mainly using E-mail option for creating, sharing and distribute knowledge and information among colleagues.

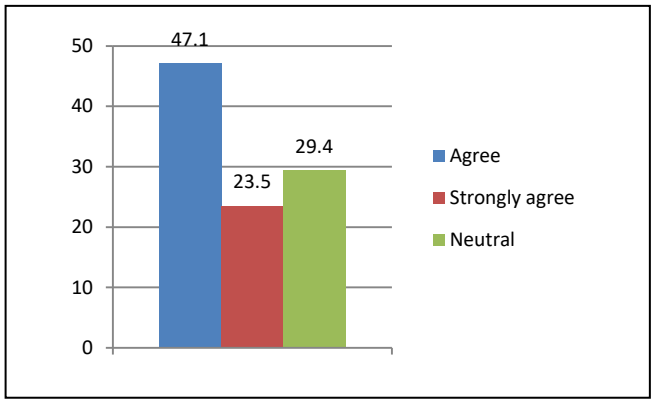


Fig. 2. Level of Trust, Encouragement and Opportunity by Respondents.

Fig. 3 above shows the level of agreement in using email and groupware as one of the KM practices option. It is found that the organizations are less using the groupware for knowledge sharing.

The internet was mostly used with (70.6%) to support KM practices and intranet (64.7%) for knowledge storing and sharing activity. Knowledge Management System (KMS) was particularly used in some of the studied organization for storing information and knowledge. The other tools used are business intelligence tools and WhatsApp group. Fig. 4 below shows the IT tools in practicing KM.

C. Process

This section with regards to KM processes which is knowledge creation, storing, sharing and knowledge dissemination. Basically, there is knowledge creation in the organization with the percentage of 79.4%. Knowledge created in a formal way by having frequent meeting and brainstorming session among the employees in the organization. Knowledge storing activities are formal and following the standard where most of the employees are using some basic tools such as data management system as well as cloud to store the information or knowledge. Knowledge sharing happens in an informal way. It is found that there is selective knowledge sharing activity with certain individual with a range of 17.6% in the organization. To be more precise, information, skills or expertise is shared with some employees only.

Fig. 5 above shows the full rate of knowledge sharing criteria. Additionally, for knowledge dissemination activities are quite interesting where the knowledge dissemination happens in an informal way within the group (44.1%).

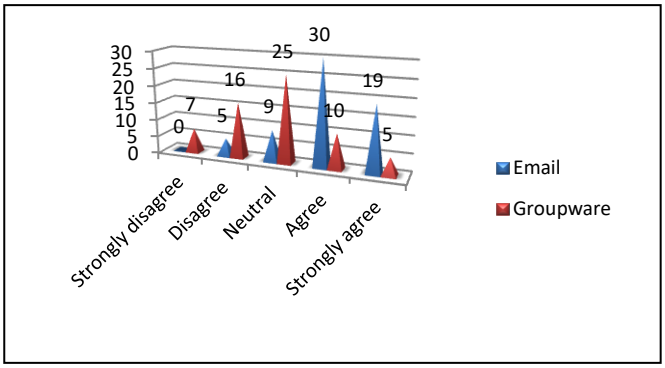


Fig. 3. Email and Groupware usages.

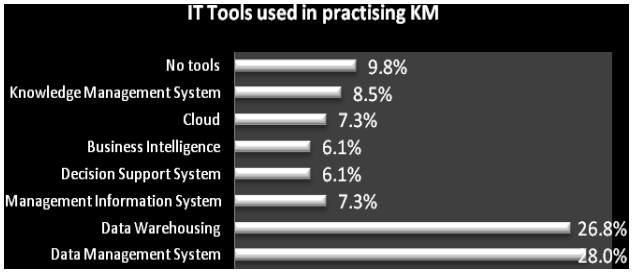


Fig. 4. ICT Tools.

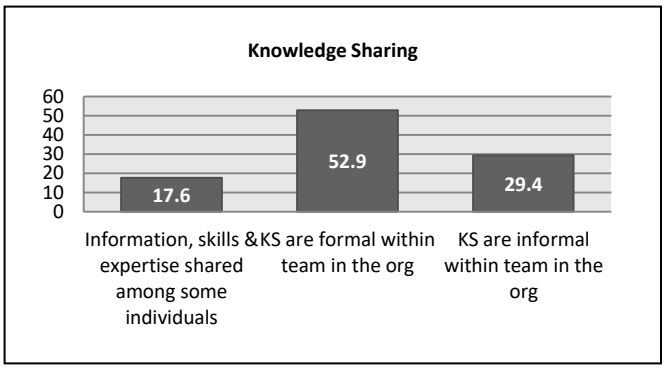


Fig. 5. Knowledge Sharing Activity.

Additionally, for Rasch analysis, an item-person map (Fig. 6) used 24 items during the initial analysis. The map stipulates that the sample gathered overall was satisfactory. The most difficult item was "accept changes" with item difficulty 0.92, while the easiest item was "teamwork" with -1.19.

Fig. 7 shows the overall category of scaling rate of this study. It is observed that category between rating 2 (disagree) and rating 3 (Neutral) distance is 0.31 (Table IV). This stipulated that these two categories are less probable to be discovered. This is due to fact that the respondents are given many categories as an option [10].

There are five factors (Fig. 8) that influenced most of the employees in terms of KM practices in their organization which is the readiness to accept the new system, less reward and recognition toward practicing KM from HR perspectives.

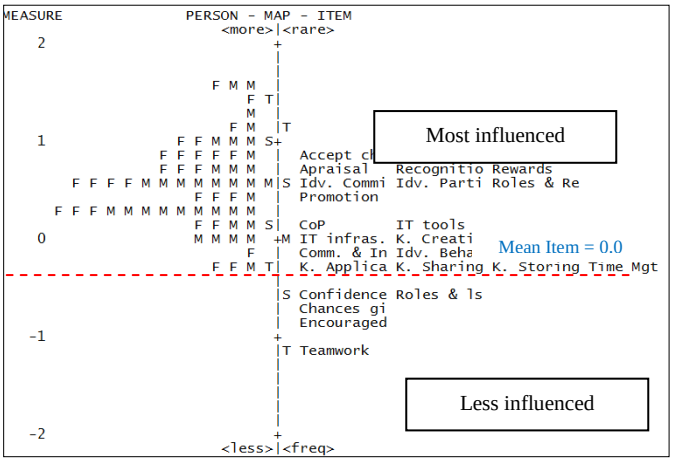


Fig. 6. Person-Item Map.

TABLE. IV. CATEGORY STRUCTURE SUMMARY

Category	Score	Observed Count (%)	Outfit MNSQ	Threshold Andrich
Strongly Disagree	1	-0.11	1.16	NONE
Disagree	2	0.22	1.29	-1.26
Neutral	3	0.22	0.85	-0.95
Agree	4	0.74	0.87	0.38
Strongly Agree	5	1.21	0.94	1.83

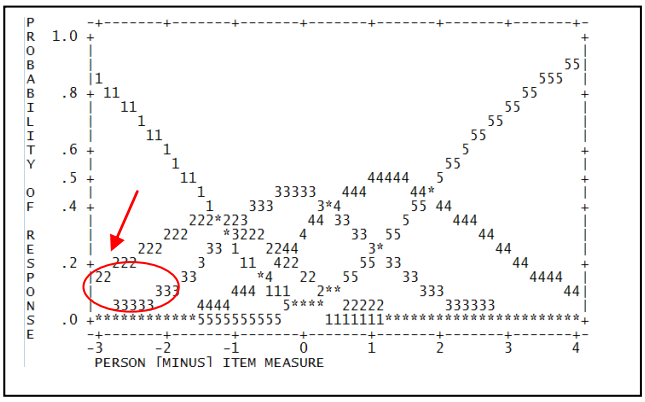


Fig. 7. Category Probability Curves for 5 Scale Rating.

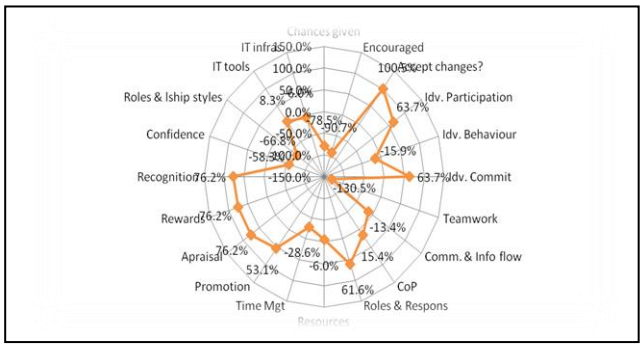


Fig. 8. Factor Influenced KM Practices in the Public Sector.

Lastly, individual participation and commitment was also seen as a difficult sub factor in the organizations. From technological perspectives, there are adequate supporting tools for practicing knowledge. The process of KM seems to be practiced at an earlier stage by some of the employees; however, without collective individual commitment and participation the whole KM initiatives won't be successful in overall. Apart from that knowledge sharing in team seems to be very approachable. Responsibility and roles of Chief Knowledge Officer (CKO), Knowledge consultant (KC), knowledge worker (KW) and knowledge manager (KM) from the study are quite less to see. Without the proper supports and encouragement from top level may influence levels of motivation from the worker's perspective.

V. Conclusions

Based on the findings, almost half of the respondents are aware of KM practices in their organization. However, it is claimed that the respondent faced some challenges while practicing KM in their daily work. In fact, the level of KM practices in their organization was also unsure. This showed that the respondent is lack of knowledge in up-to-date information about their organization practices of KM. The data revealed there several problems at people's perspective. Basically, these organizations don't have any challenges on the technology where the organization is furnished with supporting infrastructure to practice KM. The process of capturing, creating, storing and sharing in these organizations seems to be properly practiced. From this study, it is found that there are main challenges comes from people's perspective. The problems are briefly described as follows:

- Individual are not ready to accept changes when the new system come.
- Less of individual participation toward KM practices.
- Less individual commitment from individual perspective to KM practices.
- Lack of reward from HR.
- Lack of recognition from HR.

Apart from identifying the critical success factor of KM practices and implementation only, the research on KM practices would be better if the researcher emphasize on some sorts of measurement and identify the level of implementation to overcome challenges that being faced in the organization.

Acknowledgment

This research was sponsored by Faculty of Information Science and Technology, National University of Malaysia (UKM) under Dana Pecutan Penerbitan PP-FTSM-2019.

References

[1] M. Massaro, J. Dumay, and A. Garlatti, "Public sector knowledge management: A structured literature review," Journal of Knowledge Management, vol. 19(3), pp. 530-558, 2015.

[2] H. Laihonen, and S. Mantyla, "Strategic knowledge management and evolving local government," Journal of Knowledge Management, Vol. 22, no. 1, pp. 219-234, 2018.

[3] O. Syed, and F. Rowland, "Benchmarking knowledge management in a public organisation in Malaysia," An International Journal, Vol. 11, no. 3, pp. 238 – 266, 2004.

[4] I. Nonaka, and N. Konno, "The concept of "Ba": Building a foundation for knowledge creation. California management review, vol. 40(3), pp. 40-54, 1998.

[5] N. A. Kassim, M. F. Baharuddin, and Z. A Samad, "Knowledge Management Practices and Organizational Performance in Malaysia Government Institution," International Journal for Infonomics (IJI), vol. 9, no. 4, 2016.

[6] Z. Ahmed, Z. Mansor, K. Ahmad, " An Analysis of Knowledge Management Challenges in Agile Global Software Development," Journal of Telecommunication, Electronic and Computer Engineering., vol. 9, pp. 63-66, 2017.

[7] A. T. Amayah, "Determinants of knowledge sharing in a public sector organization", Journal of Knowledge Management, Vol. 17 No. 3, pp. 454-471, 2013.

[8] J. Moshari, "Knowledge Management Issues in Malaysian Organizations: the Perceptions of Leaders," Journal of Knowledge Management, Economics and Information Technology, vol. 3(5), pp. 15-27, 2013.

[9] D. T. Holt, S. E. Bartczak, S. W. Clark, M. R. Trent, "The development of an instrument to measure readiness for knowledge management," Knowledge Management Research & Practice, vol. 5(2), pp. 75-92, 2007.

[10] A. M. Mills, and T. A. Smith, "Knowledge management and organizational performance: A decomposed view," Journal of knowledge management, vol. 15(1), pp. 156-171, 2011.

[11] A. Tehraninasr, and M. U. R. A. L. I. Raman, "Knowledge management obstacles in Malaysia: An exploratory study," Public Sector ICT Management Review, vol. 3(1), pp. 15-20, 2009.

[12] Abdullah and H. Date, "Public Sector Knowledge Management: A Generic Framework," Public Sector ICT Management Review, Vol. 3(1), pp. 1- 14, 2009.

[13] Singh, S. Kumar, S. Gupta, B. Donatella, and K. Shampy, "Top management knowledge value, knowledge sharing practices, open innovation and organizational performance," Journal of Business Research, 3 May 2019.

[14] P. Jain, "Knowledge Management In e-Government," Journal of knowledge management practice, vol. 10, no. 4 pp. 1-11, 2009.

[15] Mills, A.M. and T.A. Smith, "Knowledge management and organizational performance: a decomposed view," Journal of knowledge management, vol. 15(1), pp. 156-171, 2011.

[16] M. Z. Islam, S. M. Jasimuddin, and I. Hasan, "The role of technology and socialization in linking organizational context and knowledge conversion: The case of Malaysian Service Organizations," International Journal of Information Management, vol. 37, no. 5, pp. 497-503, 2017.

[17] J. M. Linacre, WINSTEPS Rasch measurement computer program. Chicago: WINSTEPS.com, 2006.

[18] K. Ahmad, M. A. Nazlena, and Z. Suhaila. "An improved course assessment measurement for analyzing learning outcomes performance using Rasch model," *Procedia-Social and Behavioral Sciences* vol. 18, pp. 442-449, 2011.

[19] D'Agostino, and B. Ralph, "Tests for the normal distribution." In Goodness-of-fit-techniques, pp. 367-420. Routledge, 2017.

[20] R. A. L. F. Griethuijsen, M. W. Eijck, H. Haste, P. J. Brok, N. C. Skinner, and N. Mansour, "Global patterns in students' views of science and interest in science," Research in Science Education, vol. 45(4), pp. 581–603, 2014.

Image Steganography using Combined Nearest and Farthest Neighbors Methods

Farhana Sharmin[1], Muhammad Ibrahim Khan[2]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Chattogram, Bangladesh

***Abstract*—Security is invariably a significant concern during communication. With the ease of communication, there is always a pending threat of intrusion. Steganography is one such way to achieve security by concealing confidential information within a more innocent looking media like image, audio, video etc. In this paper, a new technique is proposed that uses the relationship of a pixel with its Nearest Neighbor and Farthest Neighbor to hide secret information into that pixel. The cover image is divided into2x2 non-overlapping blocks. According to the vulnerability of the relationship among the pixels, blocks are labelled as Stable and Unstable. The Stable block hides 'k' secret while unstable block hides 'n' secret bits. 25 types of different set of 'k' and 'n' is examined to evaluate the performance of proposed method. 2k method is applied to improve the quality of stego image. The experimental result shows that the proposed technique hides a significant number of secret bits with high PSNR. While compared with other existing methods, the proposed method achieves a much higher visual quality than that of those methods.**

Keywords*—Steganography; cryptography; Pixel Value Difference (PVD); image processing; cover image; information security; stego image

I. Introduction

With the progress made in digital technology, information in analog form are being rapidly converted into their digital counterparts. Digital technology and the Internet today offer available cover medias for use in steganography. The frequency of digital image on the Internet makes it the most common option among different cover medias. Internet technology has bridged the global distance between people and organizations by allowing for fast and cheap communication. In areas such as online intellectual property rights, criminal, political and commercial communications there is always a fear of third-party intruder who is seeking access to our communication. It is not enough to rely upon the assumption that communications in today's high-tech world are safe and not being intercepted. Ensuring secure communication is essential. Different tools ensuring secure communication include: hiding the content or nature of communication (encryption, water-marking, steganography), hiding the parties to a communication (anonymity) and conceal a communication that has been taken place (obscurity). A perfect Steganographic system is however an ideal concept. Research in this field aims at increasing capacity, security and robustness of the system. Steganography is a mechanism that implants a secret message into unspectacular cover media to prevent arousing the suspicion of an eavesdropper. The cover media thus serves as a container for the secret information. The outcome is a stego media that appears indistinguishable from the original media and becomes undetectable. In the modern world digital steganography is used for secret communication as well as authentication. A steganographic system consists of two major parts: cover media and stego media. The confidential information that needs to be hidden is embedded in seemingly harmless media. This is the cover media. The result is a stego media that appears identical to the original cover media. Steganography hides information by altering certain characteristics of the cover media that cannot be detected by visual inspection. It makes use of both redundancies that exists in the digital cover media as well as the limitations of the human eye. The design goal of proposed technique is: larger payload and higher PSNR without compromising security.

II. Literature Review

Different types of steganographic techniques are proposed to serve different purposes. Original Pixel Value Differencing (PVD) method for gray scale image was proposed by Wu and Tsai in 2003 [1]. Here the cover image is segmented into non-overlapping blocks consisting two consecutive pixels. The difference values between the pixels are classified into ranges. Based on human vision's tolerance to gray value fluctuation, the interval of 'Range' is decided. Width of the ranges decides the number of secret bits to be hidden in a pixel. Thus, it is an adaptive method. To minimize the exposure of hiding effect of PVD method, Modulus Function (MF) is employed in [2]. In [3], turnover policy and novel adjusting process is employed to improve the performance of method proposed in [2]. Capacity promoting technique is proposed in [4] to increase hiding capacity by finding more edge areas. A new quantization range table, in [5], is designed on the basis of the perfect square number to achieve better image quality and higher capacity. By redefining the remainder function to a more general form, in [6], an indeterminate equation is derived to reduce image distortion. In [7], another steganographic technique is offered by using pixel value differencing on colour images which also ensures that no pixel value in stego image exceeds the range 0 to 255. Where pixel value does not cross the range, the original PVD method is used and elsewhere proposed method has been used for embedding data. Instead of focusing on individual color component, three color components are constituted into two overlapping blocks in [8], where PVD technique is used separately on each block to conceal secret data. In [9], two steganographic technique is proposed based on adaptive quantization range. In first technique, image is divided into 2x2 non overlapping pixel blocks and second technique divides the image into 3x3 overlapping pixel blocks. First technique offers

higher capacity where higher PSNR is achieved by using the second technique which in consequence offers higher security. Smart pixel-adjustment technique is proposed in [10] to reduce visual distortion. Simple LSB substitution method [11] also achieves great image quality by applying Optimal Pixel Adjustment Process (OPAP). To provide greater embedding capacity and less manipulation of image, the side match method [12] exploits correlation between neighboring pixels. It hides secret data depending on the degree of smoothness. Pixel Value Modification (PVM) method [13] uses modulus function to embed secret information in a color image. Combined Pixel Value Differencing and Pixel Indicator Technique [14] is proposed to ensure higher security. Here one channel acts as an indicator where other two channels hold the secret information. Just Noticeable Difference (JND) technique and method of Contrast Sensitivity Function (CSF) is used in [15]. It is an edge detection method that utilizes partial information (3 bits from MSB) of each pixel value. The authors proposed a mathematical method, named 2k correction, for better imperceptibility. Author in [16] proposed the modified least significanr bits and modulus function with pixel value differencing techniques. Both steganography and cryptography is used in [17] for improved security. A higher quality form of the cover object is utilized in [18]. Author in [19] employed technique based on LSB manipulation and inclusion of redundant noise.

III. Proposed Method

The proposed method can be applied both on gray image and color image. In case of color image, the Red, Green and Blue channels are separated first. Then embedding and extraction algorithm is repeated on all three channels separately. Gray image does not need any pre-processing.

A. Embedding Process

Step 1: Divide the cover image into 2X2 non-overlapping blocks consisting four pixels: P_1, P_2, P_3 and P_4 as shown in Fig. 1.

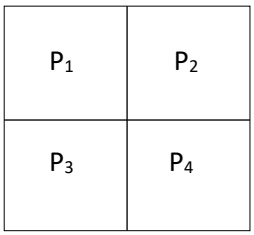


Fig. 1. Cover Image Block.

Step 2: Calculate the distance of P_1 with its other neighbor pixels P_2, P_3 and P_4 by taking the differences as follows:

Difference between P1 and P2 = D1 = |P1 – P2|

Difference between P1 and P3 = D2 = |P1 – P3|

Difference between P1 and P4 = D3 = |P1 – P4|

Step 3: Determine the Nearest Neighbor (NN) and Farthest Neighbor (FN) of P_1. Compare D_1, D_2 and D_3 calculated in Step 2. The pixel that has the lowest distance from P_1 is called its Nearest Neighbor and the pixel that has the highest distance from P_1 is called its Farthest Neighbor. If D_i is the lowest distance then P_{i+1} is the nearest neighbor and If D_i is the highest distance then P_{i+1} is the farthest neighbor, where i = 1, 2, 3.

Step 4: Determine the Range of P_1 according to the following four relationship conditions:

Condition 1: If NN > P_1 < FN then

Range = [0, *min* (*NN* - 1, *FN* - 1)]

Condition 2: If NN < P_1 > FN then

Range = [*max* (*NN* + 1, *FN* + 1), 255]

Condition 3: If NN ≥ P_1 ≥ FN then

Range = [*FN* + 1, *NN* - 1]

Condition 4: If NN ≤ P_1 ≤ FN then

Range = [*NN* + *1*, *FN* - 1]

Step 5: Calculate,

|Range| = upper limit – lower limit + 1

Step 6: Calculate the number of secret bits, n, to be hidden into P_1.

$n = \min(\text{floor}(\log_2(|\text{Range}|)), k)$

where '**k**' may vary from 1 to 5. During experiment it is found that the method reaches the maximum capacity at k = 5.

Step 7: Take 'n' secret bits from the secret bit stream and convert into its decimal value 'b'.

Step 8: Modify the value of pixel P_1:

$P_1' = P_1 - (P_1 \bmod 2^n) + b$

Step 9: To keep the new pixel value P_1' nearest to P_1, 2k Correction method is applied. The steps are as follows:

Step 9.1*:* If $|P_1' - P_1| > 2^{n-1}$ then go to Step 9.2. Otherwise P_1' remains unchanged.

Step 9.2: If $(P_1' - P_1) > 0$, then $P_1'' = P_1' - 2^n$ else $P_1'' = P_1' + 2^n$

Step 10: Check the stability of the relationship among the pixels in the block using Step 2 to Step 4. For correct message extraction, it is required to preserve the same relationship condition it has among the pixels even after hiding secret data. If the relationship among the pixels remains unchanged then the block is labelled as Stable Block and P_1'' remains unchanged. Otherwise the block is labelled as Unstable Block. In Unstable Block, all four pixels hide 'n' secret data. 'n' may vary from 1 to 5. For n>5, though the hiding capacity increases significantly but PSNR also drops drastically, which leads to vulnerability of the method. For this reason, the value of 'n' is limited from 1 to 5. After hiding n bits, P_1, P_2, P_3 and P_4 will be changed to P_1', P_2', P_3' and P_4' as follows:

$P_1' = P_1 - (P_1 \bmod 2^n) + b_1$

$P_2' = P_2 - (P_2 \bmod 2^n) + b_2$

$P_3' = P_3 - (P_3 \bmod 2^n) + b_3$

$P_4' = P_4 - (P_4 \bmod 2^n) + b_4$

Where, n is the secret binary bits and b_1, b_2, b_3 and b_4 are corresponding decimal values of those secret bits.

To increase security, secret information is embedded in random fashion. Secret message is embedded into stable blocks first then into unstable blocks. In that way, message is hidden in random order instead of sequential order.

For example, in Fig. 2, if 'abcdefghijklmnop' is the secret message, then it is embedded as 'jabkcdlemfgnohpi'. After embedding into all blocks of the cover image, we get the stego image.

B. Extraction Process

Step 1: Divide the stego image into 2X2 non overlapping blocks and identify stable blocks (Fig. 3) and unstable blocks (Fig. 4).

In case of stable blocks:

Step 2: Calculate the distance of P_1' with P_2, P_3 and P_4.

Difference between P_1' and P_2 = $D_1' = |P_1' - P_2|$

Difference between P_1' and P_3 = $D_2' = |P_1' - P_3|$

Difference between P_1' and P_4 = $D_3' = |P_1' - P_4|$

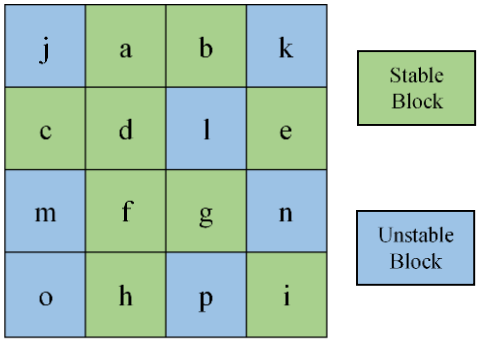


Fig. 2. Embedding Secret Message in Random Order.

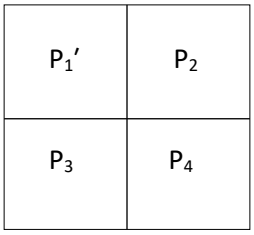


Fig. 3. Stable Block.

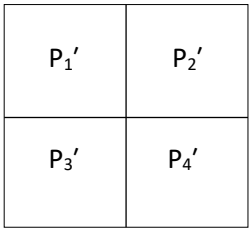


Fig. 4. Unstable Block.

Step 3: Determine the Nearest Neighbor (NN) and Farthest Neighbor (FN) of P_1' same as described in Step 3 of embedding process.

Step 4: Determine the Range of P_1' according to the four relationship conditions specified in Step 4 of embedding process.

Step 5: Calculate, $|Range| = upper\ limit - lower\ limit + 1$

Step 6: Calculate the number of secret bits 'n' to be extracted from P_1'.

$n = \min(\text{floor}(\log_2(|\text{Range}|)), k)$

Step 7: Extract **n** secret bits from the stego pixel P_1'.

$b = P_1' \bmod 2^n$

Step 8: Convert the decimal value b into its corresponding binary value. Obtained binary bits are the secret bits.

In case of Unstable Block, secret message is extracted from all four pixels using following equations:

$$PSNR = 10log_{10}\left(\frac{MAX\ I^2}{MSE}\right)$$

$$MSE = \frac{1}{\text{c m n}}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

$b_1 = P_1' \bmod 2^n$

$b_2 = P_2' \bmod 2^n$

$b_3 = P_3' \bmod 2^n$

$b_4 = P_4' \bmod 2^n$

By converting b_1, b_2, b_3 and b_4 into their corresponding binary value, we get the secret bits hidden into the pixels.

IV. Experimental Results

Most widely used 512X512X3 color images (Lena, Baboon, Pepper, Jet, Sailboat, House, Splash, Tiffany etc.) and 512X512 gray images (Lena, Boat, Baboon, Tank, Pepper, Tiffany, Jet, Elaine, Barbara, etc.) are used to conceal the secret information. In order to assess the efficiency of proposed method, two parameters are employed: Capacity and Peak Signal to Noise Ratio (PSNR). Capacity is the maximum number of secret bits an image can hide. Peak Signal to Noise Ratio (PSNR) quantifies the quality of stego image.

Here, MSE calculates the Mean Squared Error. I and K are cover and stego image respectively where both have m rows and n columns, c is the number of channels of the image. MAX_I is the maximum pixel value of cover image. Higher PSNR indicates better image quality. Usually, when the PSNR is greater than 30 dB, Human Visual System (HVS) is unable to detect the changes made in an image [20].

In steganography, capacity and PSNR walk in reverse way. The more the capacity, usually the lower is the PSNR. Lower PSNR leads to vulnerability of a method. In our proposed method, 'k' and 'n' play a major role to control the capacity.

As for example, Table I (color image) and Table II (gray image) depict how capacity (in bits) as well as PSNR (in db) varies for different values of '**k**' and '**n**' for image Lena. Users can choose any set of (k,n) to meet their priority requirement.

Taking 30 db as the threshold level PSNR, Table III and Table IV depicts the highest capacity with threshold level PSNR and capacity with highest PSNR for color images and gray images respectively. Capacity is measured in bits and PSNR is measured in db. From the experimental results, we can see that the proposed method can hide a large amount of secret information yet keeping a high PSNR.

Fig. 5 shows some color images and their corresponding stego images where some gray images and their respective stego images are shown in Fig. 6. However, it is also clear from the figures that it is evident that Human Visual System (HVS) does not perceive distortion.

TABLE. I. CAPACITY AND PSNR OF COLOR IMAGE LENA FOR DIFFERENT VALUES OF 'K' AND 'N'

Type of (k,n)	Capacity	PSNR
Type 1 (1,1)	742866	56.20
Type 2 (1,2)	1471202	51.45
Type 3 (1,3)	2199538	45.83
Type 4 (1,4)	2927874	39.85
Type 5 (1,5)	3470530	34.13
Type 6 (2,1)	712671	56.40
Type 7 (2,2)	1380255	51.76
Type 8 (2,3)	2047839	46.18
Type 9 (2,4)	2715423	40.22
Type 10 (2,5)	3381967	34.28
Type 11 (3,1)	702121	56.23
Type 12 (3,2)	1326941	51.88
Type 13 (3,3)	1951761	46.42
Type 14 (3,4)	2576581	40.50
Type 15 (3,5)	3201401	34.57
Type 16 (4,1)	702444	50.43
Type 17 (4,2)	1058920	49.44
Type 18 (4,3)	1415396	46.83
Type 19 (4,4)	1771872	42.29
Type 20 (4,5)	2128348	36.83
Type 21 (5,1)	702612	55.46
Type 22 (5,2)	1303484	51.69
Type 23 (5,3)	1904356	46.48
Type 24 (5,4)	2505228	40.63
Type 25 (5,5)	3106100	34.73

TABLE. II. CAPACITY AND PSNR OF GRAY IMAGE LENA FOR DIFFERENT VALUES OF 'K' AND 'N'

Type of (k,n)	Capacity	PSNR
Type 1 (1,1)	242106	51.55
Type 2 (1,2)	477530	46.82
Type 3 (1,3)	712954	41.19
Type 4 (1,4)	948378	35.22
Type 5 (1,5)	1158682	29.39
Type 6 (2,1)	232365	51.74
Type 7 (2,2)	448009	47.11
Type 8 (2,3)	663653	41.55
Type 9 (2,4)	879297	35.60
Type 10 (2,5)	1094941	29.68
Type 11 (3,1)	229963	51.65
Type 12 (3,2)	435735	47.22
Type 13 (3,3)	641507	41.73
Type 14 (3,4)	847279	35.79
Type 15 (3,5)	1053051	29.89
Type 16 (4,1)	230056	45.95
Type 17 (4,2)	351460	44.87
Type 18 (4,3)	472864	42.13
Type 19 (4,4)	594268	37.51
Type 20 (4,5)	715672	32.03
Type 21 (5,1)	229831	51.26
Type 22 (5,2)	430735	47.12
Type 23 (5,3)	631639	41.78
Type 24 (5,4)	832543	35.89
Type 25 (5,5)	1033447	29.98

TABLE. III. CAPACITY AND PSNR FOR PROPOSED METHOD (COLOR IMAGE)

Cover Image (512 x 512 x 3)	Highest Capacity with threshold level PSNR		Capacity with highest PSNR	
	Capacity	*PSNR*	*Capacity*	*PSNR*
Lena	3470530	34.13	712671	56.40
Baboon	3490302	34.14	768667	56.03
Pepper	3474527	34.28	730659	56.25
Sailboat	3481380	34.18	745340	56.17
Car house	3475411	34.19	708470	56.46
Splash	3321788	34.21	669662	56.75
Jet	3424467	34.33	684031	56.64
Tiffany	3471557	34.64	710055	56.40

TABLE. IV. CAPACITY AND PSNR FOR PROPOSED METHOD (GRAY IMAGE)

Cover Image (512 x 512)	Highest Capacity with threshold level PSNR		Capacity with highest PSNR	
	Capacity	*PSNR*	*Capacity*	*PSNR*
Lena	948378	35.22	232365	51.74
Baboon	1013763	34.97	249900	51.36
Pepper	1038439	30.01	237555	51.62
Jet	1038254	30.00	226950	51.91
Airplane	910753	30.74	191549	52.74
Boat	1002498	35.02	244560	51.47
Lake	998313	35.04	244987	51.48
Tiffany	1017169	30.11	228397	51.83
Elaine	1010298	34.96	247009	51.43
Goldhill	994878	35.08	242765	51.52
Cameraman	1018789	30.05	217845	52.16
Carhouse	932328	35.30	233172	51.75
Tank	1047990	30.00	242889	51.49
Truck	980733	35.16	238737	51.60
Couple	985443	35.08	240799	51.53
Zelda	1027092	30.00	230426	51.77
Barbara	979443	35.09	241677	51.54

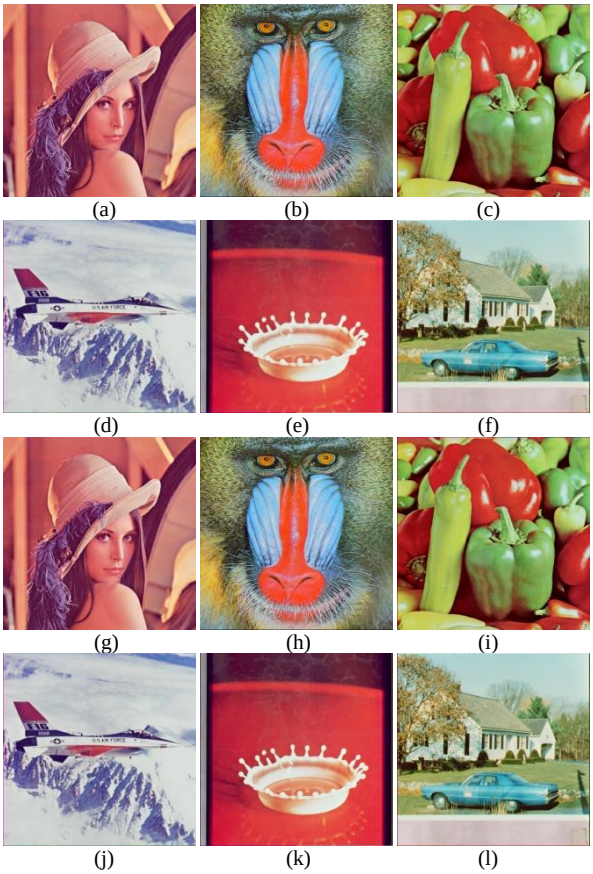


Fig. 5. Color Cover Images: (a) Lena (b) Baboon (c) Peppers (d) Jet (e) Splash (f) Carhouse and (g) – (l) are their Respective Stego Images for Type 18 (k = 4, n = 3).

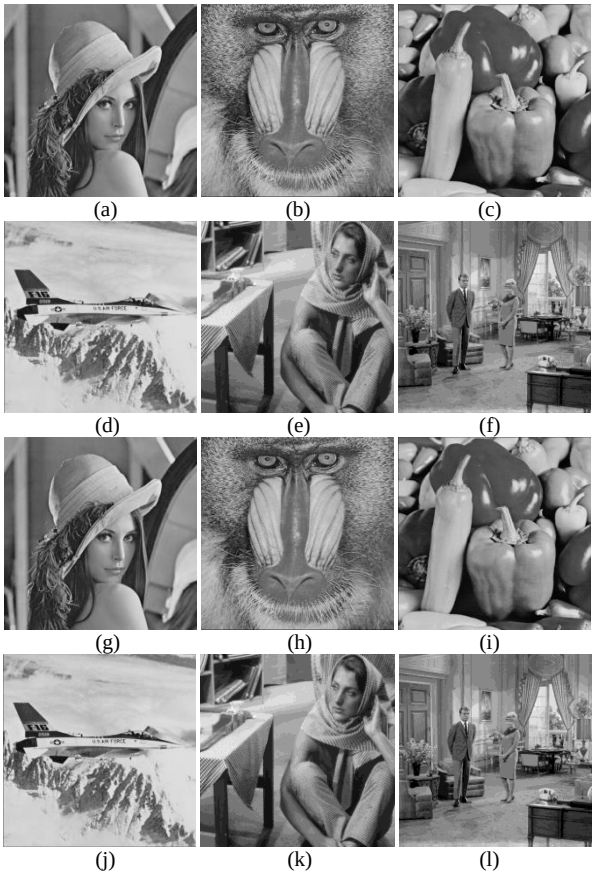


Fig. 6. Gray Cover Images: (a) Lena (b) Baboon (c) Peppers (d) Jet (e) Splash (f) Carhouse and (g) – (l) are their Respective Stego Images for Type 18 (k = 4, n = 3).

V. COMPARATIVE ANALYSIS AND DISCUSSION

To analyse how well the proposed method is working, it is compared to some relevant methods proposed before like Wu and Tsai's method [1], Joo et al.'s method [3], Yang et al.'s method [4], Tseng and Leng's method [5], Mandal and Das's method [7], Prasad and Pal's method [8], Swain's method [9] and Yang and Wang's method [10]. In order to have a comparative analysis of the outcomes of proposed method, Table V and Table VI is constructed to show the comparison between proposed method and other existing methods for color image and gray image respectively. For each existing method, PSNR at the highest capacity is compared with the PSNR of proposed method at the same capacity. From Table V and Table VI, it is evident that proposed method outperformed the existing methods each time. For same capacity, proposed method's PSNR is significantly higher than other methods compared. In case of color image, proposed method's average PSNR is 26.43%, 49.21%, 11.02% and 41.32% higher than that of Mandal and Das's method, Yang and Wang's method, Swain's method and Prasad & Pal's method, respectively. Applying the proposed method on gray images also produces better result than other methods compared. Here also, for same capacity, proposed method's average PSNR is 16.29%, 3.50%, 16.94% and 12.81% higher than the PSNR gained in Wu & Tsai's method, Tseng et al.'s method, Yang et al.'s method and Joo et al.'s methods, respectively.

TABLE. V. COMPARISON BETWEEN PREVIOUS METHODS AND PROPOSED METHOD FOR COLOR IMAGES

Cover Image (512 x 512 x 3)	Mandal and Das's method		Proposed method	Yang and Wang's method		Proposed method	Swain's method		Proposed method	Prasad and Pal's method		Proposed method
	Capacity	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*
Lena	1166296	42.26	52.45	196608	41.58	62.14	1341192	46.17	51.89	1976671	31.01	46.34
Baboon	1159328	38.44	52.43	196608	33.29	61.96	1489945	48.49	51.38	2219715	32.29	45.79
Peppers	1167960	42.28	52.42	196608	39.43	62.01	1350251	47.06	51.81	1783210	30.10	46.81
Jet	1165184	42.60	52.51	196608	43.73	62.35	1267690	46.18	52.14	1753707	35.66	46.88
Sailboat	1146224	40.66	52.48	196608	47.41	62.04	1424967	47.29	51.59	2130772	33.11	46.16
Car-House	1162992	41.41	52.49	196608	41.34	62.26	1339985	44.73	51.90	2079088	34.59	46.08
Splash	1173856	42.86	52.52	196608	44.86	62.39	-	-	-	-	-	-

TABLE. VI. COMPARISON BETWEEN PREVIOUS METHODS AND PROPOSED METHOD FOR GRAY IMAGES

Cover Image (512 x 512)	Wu and Tsai's method		Proposed Method	Tseng et al. method		Proposed Method	Yang et al. method		Proposed Method	Joo et al. Method		Proposed Method
	Capacity	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*	*Capacity*	*PSNR*	*PSNR*
Lena	407680	41.79	47.51	215740	50.70	52.07	528966	36.75	42.56	407152	43.4	47.52
Baboon	450328	37.90	47.04	241719	48.57	51.50	559222	34.30	42.22	456344	39.2	46.98
Jet	409944	40.97	47.52	204682	50.89	52.37	-	-	-	409768	42.8	47.52
Peppers	405480	41.73	47.50	217290	50.57	52.01	528791	36.83	42.54	406520	42.5	47.49
Elaine	407128	42.1	47.44	-	-	-	-	-	-	407144	43.5	47.44
Tiffany	403764	41.47	47.53	210935	50.86	52.18	528678	36.35	42.49	-	-	-
Boat	418560	39.6	47.29	-	-	-	-	-	-	419920	41.0	47.27
Tank	403990	42.3	47.33	-	-	-	-	-	-	-	-	-
Lake	420912	40.0	47.34	224915	49.86	51.85	535984	36.51	42.41	421296	41.5	47.34
Gold hill	405634	42.20	47.48	-	-	-	529319	37.15	42.51	-	-	-
Zelda	398584	42.66	47.63	-	-	-	526145	37.78	42.60	-	-	-
Barbara	442529	36.24	47.12	-	-	-	548206	34.80	42.33	-	-	-

VI. CONCLUSION

An efficient technique is introduced to hide a significant amount of secret information keeping the distortion unnoticeable. Capacity and PSNR is quite higher than other existing methods discussed. The greatest advantage of our proposed method is, users can easily tailor this method according to their need by taking the desired Type of (k,n). Either to achieve higher capacity or to maintain a higher PSNR, whatever the target may be, just tuning the values of 'k' and 'n' is sufficient. Random embedding order is followed instead of traditional sequential order by hiding secret information into Stable Blocks first than into Unstable Blocks. This step is working as an extra layer of security if any unauthorized attack happens. Since only the sender and recipient know the correct order of extraction, the intruder is supposed to get a meaningless message in that case. Future work may include utilizing the unstable blocks in different way like hiding variable rate secret data instead of fixed 'n' bits secret data.

REFERENCES

[1] D.-C. Wu and W.-H. Tsai, "A steganographic method for images by pixel-value differencing," Pattern Recognition Letters, vol. 24, no. 9-10, pp. 1613–1626, 2003.

[2] C. Wang, N. Wu, C. Tsai and M. Hwang, "A high quality steganographic method with pixel-value differencing and modulus function", Journal of Systems and Software, vol. 81, no. 1, pp. 150-158, 2008.

[3] J. Joo, H. Lee and H. Lee, "Improved Steganographic Method Preserving Pixel-Value Differencing Histogram with Modulus Function", EURASIP Journal on Advances in Signal Processing, vol. 2010, no. 1, 2010.

[4] C. Yang, C. Weng, H. Tso and S. Wang, "A data hiding scheme using the varieties of pixel-value differencing in multimedia images", Journal of Systems and Software, vol. 84, no. 4, pp. 669-678, 2011.

[5] H.-W. Tseng and H.-S. Leng, "A Steganographic Method Based on Pixel-Value Differencing and the Perfect Square Number," Journal of Applied Mathematics, vol. 2013, pp. 1–8, 2013.

[6] W. Zhao, Z. Jie, L. Xin, and W. Qiaoyan, "Data embedding based on pixel value differencing and modulus function using indeterminate equation," The Journal of China Universities of Posts and Telecommunications, vol. 22, no. 1, pp. 95–100, 2015.

[7] J. K. Mandal and D. Das, "Colour Image Steganography based on Pixel Value Differencing in Spatial Domain," IJIST International Journal of Information Sciences and Techniques, vol. 2, no. 4, pp. 83–93, 2012.

[8] S. Prasad and A. Pal, "An RGB colour image steganography scheme using overlapping block-based pixel-value differencing", Royal Society Open Science, vol. 4, no. 4, p. 161066, 2017.G. Swain, "Adaptive pixel value differencing steganography using both vertical and horizontal edges", Multimedia Tools and Applications, vol. 75, no. 21, pp. 13541-13556, 2015.

[9] C.-Y. Yang and W.-F. Wang, "Block-Based Colour Image Steganography Using Smart Pixel-Adjustment," Advances in Intelligent Systems and Computing Genetic and Evolutionary Computing, pp. 145–154, 2015.

[10] C.-K. Chan and L. Cheng, "Hiding data in images by simple LSB substitution," Pattern Recognition, vol. 37, no. 3, pp. 469–474, 2004.

[11] C.-C. Chang and H.-W. Tseng, "A steganographic method for digital images using side match," Pattern Recognition Letters, vol. 25, no. 12, pp. 1431–1437, 2004.

[12] V. Nagaraj, V. Vijayalakshmi, and G. Zayaraz, "Color Image Steganography based on Pixel Value Modification Method Using Modulus Function," IERI Procedia, vol. 4, pp. 17–24, 2013.

[13] M. G. Gouthamanaath and A. Kangaiammal, "Color Image Steganography using Combined Pixel Value Differencing and Pixel Indicator Technique in Spatial Domain," IJCA Proceedings on National Conference on Research Issues in Image Analysis and Mining Intelligence, pp. 20–23, Jun. 2015.

[14] J.-G. Yu, E.-J. Yoon, S.-H. Shin, and K.-Y. Yoo, "A New Image Steganography Based on 2k Correction and Edge-Detection," Fifth International Conference on Information Technology: New Generations (itng 2008), 2008.

[15] T.-S. Chen, C.-C. Chang, and M.-S. Hwang, "A virtual image cryptosystem based upon vector quantization," IEEE Transactions on Image Processing, vol. 7, no. 10, pp. 1485–1488, Oct. 1998.

[16] M. Sabokdast and M. Mohammadi, "A steganographic method for images with modulus function and modified LSB replacement based on PVD," The 5th Conference on Information and Knowledge Technology, 2013.

[17] S. Gupta, A. Goyal, and B. Bhushan, "Information Hiding Using Least Significant Bit Steganography and Cryptography," IJMECS International Journal of Modern Education and Computer Science, vol. 4, no. 6, pp. 27–34, 2012.

[18] T. Denemark and J. Fridrich, "Side-informed steganography with additive distortion," 2015 IEEE International Workshop on Information Forensics and Security (WIFS), 2015.

[19] Chadha A., Satam N., Sood R., and Bade D., "An Efficient Method for Image and Audio Steganography using Least Significant Bit Substitution," International Journal of Computer Applications, vol. 77, no. 13, pp. 37-45, 2013.

[20] M. Hussain and M. Hussain, "A Survey of Image Steganography Techniques," International Journal of Advanced Science and Technology, vol. Vol. 54, pp. 113–124, May 2013.

Decision Support System for Employee Candidate Selection using AHP and PM Methods

Soleman
Department of Computer Science
Faculty of Information Technology
Budi Luhur University, Jakarta, Indonesia

***Abstract*—PT. Prima Grafika is a digital printing company that is looking for the best prospective employees. Based on research and observations that have been made, the company only uses administrative data in selecting prospective employees. Processing a lot of data and documents with the same applicant's name slowing down the data processing and exceeding the time limit has been determined to get the best employees. The solution to this problem is by making a web-based system in recruitment, a decision support system combined with Analytic Hierarchy Process (AHP) method as a weighting to conduct priority criteria analysis by pairwise comparisons between two criterion so that all criteria are covered, and the Profile Matching (PM) method as a ranking. This study uses three methods of testing, namely, Black Box Testing system tested by 60 respondents who can be accepted by the company; User Acceptance Testing (UAT) obtained from 10 respondents with an ideal score of 900 produced an actual score of 779 or 86.1%, in total this whole system is acceptable; and Delone and McLean Model Test Results obtained from 10 respondents with an ideal score of 850 produced an actual score of 726 or 85.7% which is very good. With the results of ranking: A001 - Cantika Dewi = 4.12, A004-Arif Yulkianto = 3.98, A002 Eprriadi = 3,913, A003-Rika Novriani = 3,467**.

Keywords*—*Prospective employees; decision support systems; Analytic Hierarchy Process (AHP); profile matching

I. Introduction

Large companies and the development of a company cannot be separated from the role of employees. PT. Prima Grafika is a full service Digital Printing company, including Repro, Design, Computer to Plate, Computer to Film, Digital Press, Digital Offset, and Photobook. This company grew into a large enough company by opening several branches so that a selection of prospective new employees was made to occupy the required position. New employee selection aims to get the best employees that can work well.

Observations have been made for the employee selection process at PT. Prima Grafika. Data was obtained that the employee selection process was only carried out with administrative processes such as diplomas and certificates of work experience possessed by applicants. In addition, this process was difficult because of the large amount of applicant data that had to be processed; many applicant files had the same name and were vulnerable to being swapped.

The solution that will be applied by researchers in the selection of new employees at the PT Prima Grafika is to create a web-based application in recruitment by implementing the decision support system of the Analytic Hierarchy Process (AHP) combination method in weighting, conducting priority analysis of criterion with paired comparison methods between the two criterion and Profile Matching (PM) as the calculation of the final result (ranking).

This process will save time and money for the company because the process will be automatic and the result can be read based on the rank. The employees with the highest score are the best employee who will get a position that fits the company's criteria. Decision support systems expand the ability of decision makers to process data or information using a computer system that processes data into information for decision-making on semi-structured specific issues like in this company.

A similar study conducted in 2016 by Moedjiono, et al. using AHP and PM methods but had different goals. Their research was conducted to support decisions in order to improve the effectiveness of sending the best employees, increasing customer satisfaction from 3 Aspects of 7 criteria especially Intellectual Aspects and Work Attitude. While in this research a decision support system will be applied using a combination of Analytic Hierarchy Process and Profile matching methods, with three aspects (Aspects of Nature and Behavior, Intellectual Aspect, Administrative Aspect) which are divided into 11 sub aspects or criteria, namely, Work Motivation, Appearance, Accuracy, Speed, Intelligence, Language Mastery, Problem Solving, Domicile, Diploma, Work Experience and Age.

II. Theoretical Background

A. Decision Support System

According to Pratiwy (2016) understanding of decision support systems proposed by McLeod which states that a decision support system is a system of producing information aimed at a problem that must be created by managers, decision support system is an information system intended to help management in solving problems they faced.

B. Study Overview

Research on decision support systems using the Analytical Hierarchy Process (AHP) approach and the Profile Matching (PM) method or those relating to employee selection have been conducted by several people including.

Decision support system used in the selection of candidates or employees with a comparison of two methods in

the calculation process, that the Analytical Hierarchy Process (AHP) is better / complex than the Multi Factor Evaluation Process (MFEP), determined by four criteria: interview (0.558) , writing test (0.122), psychological test (0.263), and health test (0.057), that is, candidates who meet the requirements are accepted by 2 candidates and 4 candidates with values of 8.42 and 8.23 (for MFEP) and 0.277 and 0.342 (for AHP), respectively; the results are consistent standards[1]. Decision support system that is used can help to rank the top down employees who have the best performance and so it can be in the decision making to choose alternative employee mutations 40%, 60% [2]. Decision support system used to determine the best employee can be proven that consistency below <0.100 with IR is 0.5 results that have been obtained are criteria such as the following: Disciplinary criteria: CR = 0.95, Responsibility criteria: CR = 0.58, Skills : CR = O, 95, and Cooperation CR = 0.95 [3]. Decision support system with a combination of research methods can be used as a decision maker in order to increase the effectiveness of sending the best candidates / employees, and increase customer satisfaction, and reduce costs and increase the company's gross operating margin. The final result of software quality according to respondents is Good with the percentage of respondents at 83.87%[4]. decision support system in the comparison between Analytical Hierarchy Process (AHP) and Simple Additive Weighting (SAW) methods, is used to find out the accuracy of each method in determining employee employee performance appraisal, namely the Highest Score {(SAW: 0.871) (AHP) : 0.368), Lowest score (SAW: 0.686) (AHP: 0.110), average final score (SAW: 0.767) (AHP: 0.2), Total from the final score accumulation (SAW: 3.834) (AHP: 1) Accuracy AHP results are more accurate than SAW [5]. The decision support system for the Promethee method. Promethee Method is one method of determining the order or priority in MCDM (Multi Criterion Decisin Making) or decision making [6]. Decision Support Systems can support all who support decision making to help improve and facilitate the decision making process. The results of this study will produce a ranking ranking of employees and this application can help the results of decisions in choosing alternative employees who excel [7]. Decision Support System with the MFEP method in determining high achieving students gives subjective and intuitive consideration in giving weight to each criterion [8]. Decision support system uses the PM method in determining employee achievement with aspects of the criteria: disciplined achievement, integrity, organizational commitment, leadership teamwork, work performance and service oriented [9]. The results of his research the method used for performance assessment it is very difficult to state where the method is better to use than others because it depends on the type and business size; classifies many criteria decision making techniques (MCDM) such as TOPSIS,FTOPSIS, AHP and FAHP, fuzzy multistage, cascaded fuzzy, Neuro-Fuzzy (NF) and Fuzzy evaluation is type-2. So the right technique must be selected based on the existing application [10]. Research with an AHP-based metalearning algorithm is proposed to identify the guided classification algorithm that is most suitable for developing cynical decision support systems. The results of the ten publicly available medical datasets show that the support vector machine has the highest potential to perform well on various medical data sets [11]. This study is on the adoption of social media in employee recruitment and selection by testing the Unified Theory of Acceptance and Use of Technology (UTAUT), the results obtained from modeling the partial least square path, using cross-sectional data collected from 224 recruiters, showing that the hypothesis The core of UTAUT is supported, namely the positive impact of performance expectations [12]. The research that the backfilling algorithm does not provide better scheduling if there is the same type of rental and must be conjugative. By using AHP (Analytic Hierarchy Process) as a decision maker in the backfilling algorithm increases the performance of the refilling algorithm by scheduling more rent amounts and minimizing the rejection of leases using AHP [13]. This study applies the Methods for Order Preference with Ideal Solution Equality (TOPSIS) to determine vocational schools. Criteria for Heart Decision Criteria Requires C1, C2 C3 Facilities, Quality Accreditation, C4 HR Students, C5 Fees [14].

This research applies the system Decision Support performance appraisal employee for employee selection achievers. In profile matching, good or bad identification of groups of employees is done. The employees inside the group was measured using some assessment criteria namely aspects of discipline, collaboration and work performance [15]; research assessment of employee performance. Conduct an assessment in giving awards for outstanding employees including using a support system a decision in helping a solution problem. The method used in giving awards for outstanding employees namely the profile method matching. Criteria or aspects use on the award decision support system. These outstanding civil servants (PNS) are discipline, integrity, organizational commitment, leadership, cooperation, work performance and service oriented [16]. Decision support system used shows that the consistency ratio (CR) of the five criteria is CR = 0.0976 which shows that consistency in all criteria has more weight with the amount of accuracy: 43% when compared to other criteria with teaching methods (30%) teaching quality (16%), initiative (7%), skills and personal (4%) [17]. Employee performance appraisal using the profile matching method that is able to select employees with achievements with determined aspects, namely, aspects of work targets and behavioral aspects and seven sub criteria [18]. Decision support systems are applied in the selection of ERP systems using software applications that are being developed. This software application makes it easy to apply the AHP process to the selection of ERP packages and will allow updates and improvements to the model [19]. Research with decision support systems is using AHP and TOPSIS methods algorithm as an effective tool for supporting machines selection decision. In this study, weights Different criteria are calculated using the AHP method and to choose the most preferred machine one of the well-known MCDM methods namely the TOPSIS method has used and checked back using Microsoft Office Superior [20].

C. Analytic Hierarchy Process (AHP) Method

The AHP method was developed by Thomas L. Saaty, a mathematician. This method is a framework for making decisions effectively, simplify, speeding up complex

decisions. Give numerical value to subjective considerations about the importance of each variable and synthesize these various considerations to determine which variable has the highest priority and act to influence the outcome of the situation.

The stages in using the Analytic Hierarchy Process (AHP) Method in solving problems there are some basic principles that must be understood, namely:

Create a hierarchy. Complex systems can be understood by breaking them down into supporting criteria, constructing criteria in a hierarchical manner, and combining them, such as the hierarchical structure in Fig. 1.

1) Evaluation of criteria and alternatives is done by comparison of various issues, scale 9 is the best scale for expressing opinions. The values and definitions of qualitative opinions from the Saaty comparison scale can be measured using the analysis in Table I.

2) Determine priorities (synthesis of priority). For each criterion and alternative used, pairwise comparison is needed. The values of all criteria and alternative comparisons can be adjusted according to predetermined judgments to produce weights and priorities.

Establish priorities and prepare pairwise comparisons that compare all criteria for each hierarchy if the operating subsystems have n operating criteria namely A_1, A_2,…,A*n* then the results of the comparison of each operating criteria will form a matrix A size n × n in Table II as follows:

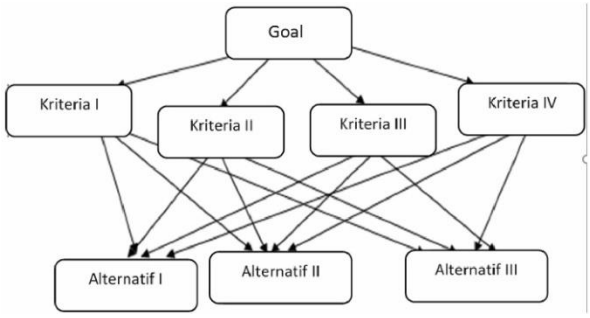


Fig. 1. Hierarchy Structure of AHP

TABLE. I. PAIRWISE COMPARISON SCALE

Scale of Importance	Information
1	Both criteria has the same influence
3	Assessment is in favor of one of the criteria than their pair
5	Assessment is very favoring one of the criteria compared to their pair
7	One criterion is very influential and its dominance is real
9	Evidence that one of the criteria is more important than the partner at the highest level of confidence.
2,4,6,8	The value is given if there is doubt between two adjacent values
Reverse	If criterion i has one of the numbers above compared to criterion j, then j has the inverse value when compared to criterion i

TABLE. II. PAIRWISE COMPARISON MATRIX [5]

	A_1	A_2	⋯	A_n
A_1	a_{11}	a_{12}	⋯	a_{1n}
A_2	a_{21}	a_{12}	⋯	a_{2n}
⋮	⋮	⋮	⋱	⋮
A_n	an_1	a_{n2}	⋯	a_{mn}

With a pairwise comparison matrix, normalization is performed using the following steps:

3) The weight of each column j is summed, the total column values are denoted by S_{ij}

$$S_{ij} = \sum_{i=1}^{n} a_{ij} \quad (1)$$

4) The value of each column is divided by the total value of the column. The results of that division are symbolized by $V_{ij.}$

$$V_{ij} = \frac{a_{ij}}{S_{ij}} \quad (2)$$

Ij = 1,2,3, … , n

Furthermore, by calculating the relative priority vectors of each criterion by averaging the normalized weights with the i-th row. The i-th priority criteria is dsymbolized by Pi.

$$\boldsymbol{P_i} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{Qi}}{\boldsymbol{n}} \quad (3)$$

5) Logical consistency. Consistency has two meanings namely similar objects can be grouped according to uniformity and relevance, and relationships between objects based on certain criteria

Assessment with a matrix often has small changes in the value of *aij* which affect the maximum eigenvalue. Whereas the maximum eigen value deviation determines the size of consistency. Indicators of consistency are measured through the consistency index as follows:

$$CI = \frac{\lambda_{maks} - n}{n-1} \quad (4)$$

AHP measures all consistency assessments using the consistency ratio (CR) formulated:

$$CR = \frac{CI}{R1} \quad (5)$$

A certain level of consistency is needed in determining priorities to get the best results. CR value ≤ 0.100 is consistent if it was not, it needs to be revised. The limit of consistency determined by Saaty, measured using the Consistency Ratio (CR), which is a comparison of the consistency index with a random generator value (RI).

N is the order of matrix n with the number 11 based on the order of the matrix, whereas RI is the Random Index for 0 <n <11. This value depends on the order matrix n. To see the percentage of ratio consistency from pairwise comparison in the AHP process is in Table III (derived from Saaty's book) as follows:

TABLE. III. RI VALUE RANDOM INDEX [6]

N	1	2	3	4	5	6
RI	0.00	0.00	0.58	0.90	1.12	1.24
N		7	8	9	10	11
RI		1.32	1.41	1.45	1.49	1.51

D. Profile Matching Method

According by [21] profile matching is a situation where the client will get the expected results. In the profile matching process, there will be a comparison process between individual competencies into standard competencies, in this case an ideal applicant profile so that competency differences can be known (also called gaps). The smaller the gap produced, the greater the weight value. Applicants who have a large weight value means have a greater chance to be accepted in the company, seen in Table IV as follows:

Gap = Attribute Value - Target Value

Calculation and Grouping of Core Factors and Secondary Factors After determining the weight of the gap values for all aspects in the same way, each aspect is further divided into two groups, namely, the core factor (main factor) and secondary factor (supporting factors). Calculation of core factors and secondary factors can be shown in equation [7].

1) Core Factor Calculation

$$NCF = \frac{\sum NC(i,s,p,x)}{\sum IC} \quad (6)$$

Information:

NCF : The average value of the core factor

NC (i, s, p, x) : Total number of core factor values

(administration, competence, interview, psychological test)

IC : Number of core factor items Calculation of

2) Secondary Factor

$$NSF = \frac{\sum NC(i,s,p,x)}{\sum IS} \quad (7)$$

Information:

NSF : average value of secondary factor

NS (i,s,p,x) : Total number of secondary

(administration, competence,

interview, psychological test)

IS : Number of secondary factor items

3) Calculation of Total Value

N(I,S,P,X)=(X)%.NCF(I,S,P,X)+(X)%.NSF(I,S,P,X) (8)

Information:

NCF (i,s,p,x) : Average value of core factor

(administration, competence,

interview, psychological test)

NSF (i,s,p,x) : Average value of secondary factor

(administration, competence,

interview, psychological test)

N (i, s, p, x): Total value of the aspect

(administration, competence,

interview, psychological test)

(X)% : The percent value inputted

4) Calculation of Ranking Determination

The final results of candidates submitted to fill a certain position in the company.

Ranking =(x)%.Ni+(x)%.Ns=(x)%.Np+(x)%.Nx (9)

Information :

Ni: Administration value

Ns: Competency score

NP: The value of the interview

Nx: Psychological values

(X)%: The percent value inputted

E. Employee Candidate Selection

Employee selection is the first effort a company must make to find employees that are in line with the company's expectations based on ability and in accordance with the position it occupies.

The definition of employee selection according to some experts is:

1) The decision making process chooses someone to fill a work position based on the suitability of individual characteristics with the needs of the job position [8].

2) The process of obtaining and using information about job applicants to determine who should be accepted to occupy short-term and long-term positions [9].

3) The process of identifying and selecting people from a group of applicants who are the most suitable or most qualified for a particular position or position [10].

TABLE. IV. GAP WEIGHT VALUE

Gap	Weight Value	Information
0	5	No gap (Competency matched with need)
1	4,5	Individual competency is 1 level more
-1	4	Individual competency is 1 level less
2	3,5	Individual competency is 2 level more
-2	3	Individual competency is 2 level less
3	2,5	Individual competency is 3 level more
-3	2	Individual competency is 3 level less
4	1,5	Individual competency is 4 level more
-4	1	Individual competency is 4 level less

III. Systems Design and Application

A. Method of Analysis

Analysis of the data of this study uses quantitative data which is a form of research method used to examine a particular population or sample, collecting data using research instruments with the aim to test the hypotheses that have been set. After the data is collected, the data is transcribed by applying the Analytic Hierarchy Process (AHP) as a determinant of the weight of the criteria, where each evaluation criterion is compared with each other and the Profile Matching (PM) method as a method for determining alternative final values. Reports generated by this system are ranking reports of all alternatives (applicants) that have been sorted by the largest alternative value. The next step is to create an application with the PHP and MySQL programming languages as the database. Prospective employees received are Prospective employees who occupy the highest ranking position of the number of applicants available, based on the number of company needs for the selected prospective employees.

B. Technique of Analysis

The analysis technique used in this study uses the Object Oriented Analysis (OOA) approach or object-oriented analysis with UML. The analysis process is carried out on the results of the stages of data collection by interviewing, observing and studying the literature to get a specification of the system requirements to be developed. In the analysis process that will be carried out are:

1) Analysis of Data and Information obtained from interviews, observations, and literature studies.

2) Analysis of Functional, non-functional and user needs. Functional requirements modeling to describe the system functions and users involved as well as what functions will be obtained by each user is modeled using a use case diagram.

3) In this research the decision support system approach technique uses a combination of Analytic Hierarchy Process (AHP) and Profile Matching (PM) methods.

Use Case diagram in Fig. 2 is a system design process that will be made in selecting prospective employees. As an actor or actor the assessment in the system design is carried out by recruitment staff who carry out the selection process. Use Case diagram illustrates that recruitment staff can log in to process alternative data, criteria, criteria weights, Profiles, Calculations, Print and logout from the system As shown in Fig. 2 as follows:

C. Design Technique

This sequence diagram illustrates the dynamic collaboration between objects that shows a series of messages between actors (Staff Recruitment) and the system. Following is the Sequence diagram of a prospective employee selection decision support system as shown in Fig. 3 as follows:

Class diagram of the decision support system for prospective employee selection, in this study describes the structure and description of the class and the relationship between classes, as shown in the following Fig. 4:

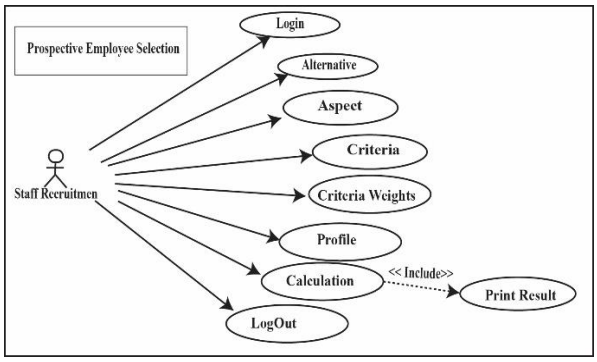


Fig. 2. Use Case Diagram of Decision Support System.

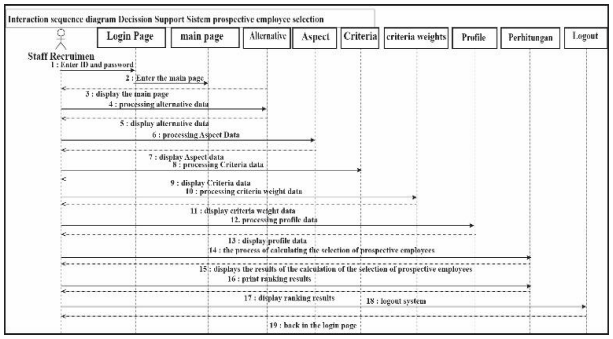


Fig. 3. Sequence Diagram of Prospective Employee Selection.

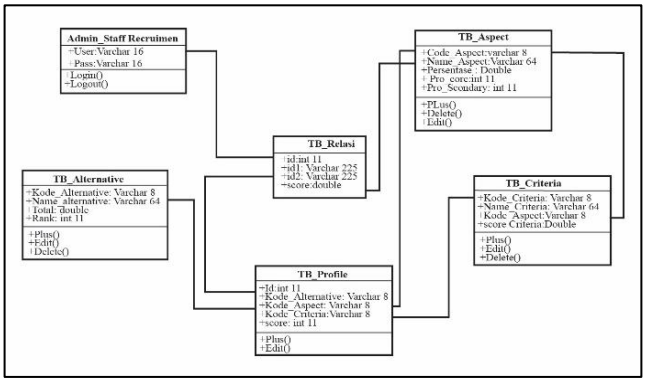


Fig. 4. Class Diagram of Prospective Employee Selection.

IV. Result and Discussion

A. Data Collection

The results of interviews with HRD PT. Prima Grafika establishes 3 Aspects and 11 Sub-Aspects (criteria) and four sample applicants data for 2018 that can be used as validation calculations in the decision support system. The following are the prospective employee criteria determined by the HRD PT. Prima Grafika can be seen in Table V as follows:

There are two types of Sub aspects (criteria), namely:

1) Core Factor, the most important or prominent criteria (competencies) or most needed by an assessment with a value of 60%.

2) Secondary Factor (supporting factors), which are items other than those in the core factor with a value of 40%.

B. AHP Weighting Process

Determining the value of the pairwise matrix comparison is done by the AHP method. Compare input data between criteria in a paired matrix using the AHP importance intensity scale. This process is carried out to determine the value of the ratio of consistency of comparison (CR), where the CR condition <0.1. The first aspect of the nature and behavior aspects based on work, appearance, accuracy, speed, sub-aspects can be seen in Table VI as follows:

The Second Aspect of Intellectual Aspect based on Sub Aspect of Intelligence, Handling Problems, Language Mastery, can be seen in Table VII as follows:

The Third Aspect of Administrative Aspects based on Sub Aspects of work motivation, Appearance, Accuracy, Speed can be seen in Table VIII as follows:

Example Explanation: Aspect Comparison Matrix The properties and behavior in Table VI. are evaluated and added to each column to get the following results: Column AI01 = 1 + 0.5 + 0.333 + 0.25 = 2.083. The next step is to divide the numbers in Table VI. by the number of each column, so that a normalization matrix is formed. AI01 column AI01 line divided by the number of AI01 columns = 1 / 2,083 = 0.48, and the average AI01 line = (0.48 + 0.5217391+ 0.461538+ 0.4 / 4)=0.465819.

After that the Consistency Measure Phase = the same two array matrix of the same number of rows as A101 code and the same number of columns with the weight value = 1,2,3,4 (in Table VI) + average weight (in Table IX) divided weights A101 The results can be seen in Table IX, X and XI as follows:

TABLE. V. ASPECT AND CRITERION (SUB ASPECT)

Aspect	Code	Criterion (sub aspect)	Factor
Nature and Behavior aspect	AI01	Work's motivation (MK)	Core
	AI02	Appearance (PNP)	Secondary
	AI03	Accuracy (KTL)	Secondary
	AI04	Speed (KCP)	Secondary
Intellectual aspect	AII01	Intelligence (KCS)	Core
	AlI02	Problem solving (PMS)	Secondary
	AII03	Language mastery (PBS)	Secondary
Administrative Aspect	AIII01	Education/Diploma (IJZ)	Core
	AIII02	Domicile (DMS)	Secondary
	AIII03	Work experience (PKJ)	Secondary
	AIII04	Age(USA)	Secondary

TABLE. VI. COMPARISON MATRIX OF ASPECTS OF NATURE AND BEHAVIOR

Kode	AI01	AI02	AI03	AI04
AI01	1	2	3	4
AI02	0.5	1	2	3
AI03	0.333	0.5	1	2
AI04	0.25	0.333	0.5	1
Total	2.083	3.833	6.5	10

TABLE. VII. MATRIX INTELLIGENCE COMPARISON

Kode	AII01	AII02	AII03
AII01	1	5	2
AII02	0.2	1	0.2
AII03	0.5	5	1
Total	1.7	11	3.2

TABLE. VIII. MATRIX ADMINISTRATIVE COMPARISON

Kode	AIII01	AIII02	AIII03	AII04
AIII01	1	9	2	7
AIII02	0.11	1	0.2	1
AIII03	0.5	5	1	9
AIII04	0.143	1	0.111	1
Total	1.754	16	3.311	18

TABLE. IX. NORMALIZATION OF ASPECTS OF NATURE AND BEHAVIOR

Code	AI01	AI02	AI03	AI04	weight	CM
AI01	0.48	0.5217391	0.461538	0.4	0.465819	4.051335
AI02	0.24	0.2608696	0.307692	0.3	0.27714	4.041634
AI03	0.16	0.1304348	0.153846	0.2	0.16107	4.015971
AI04	0.12	0.0869565	0.076923	0.1	0.09597	4.015218
						16.12416
						4

TABLE. X. NORMALIZATION OF INTELLECTUAL ASPECTS

Kode	AII01	AII02	AII03	Weight	CM
AII01	0.588	0.45	0.625	0.556	3.085
AII02	0.118	0.09	0.063	0.090	3.014
AII03	0.294	0.45	0.313	0.354	3.063
					9.161726
					3

TABLE. XI. NORMALIZATION OF ADMINISTRATIVE ASPECTS

Kode	AIII01	AIII02	AIII03	AIII04	Weight	CM
AIII01	0.57	0.5625	0.604027	0.389	0.465819	4.051335
AIII02	0.063348416	0.0625	0.060403	0.056	0.27714	4.041634
AIII03	0.285067873	0.3125	0.302013	0.5	0.16107	4.015971
AIII04	0.081447964	0.0625	0.033557	0.056	0.09597	4.015218
						16.12416
						4

Example remarks in Table IX: The total value with 4 results on the Consistency Measure on the index ratio is 0.90 and the Consistancy Index: (16.12416 / 4) –4) / (4-1) = 0.010 i.e. as a limit of the consistency measured using the Consistency Ratio (CR). Consistancy Ratio: 0.01, namely, the consistency ratio shows the extent to which the analyst is consistent in giving a value to the comparison matrix. In general, the results of the analysis are considered consistent if they have a CR of 10%.

The Last Stage The result of the weight value based on the AHP calculation is the value of the criterion that has the greatest weight. The criteria that prioritize the results are consistent with the results of Aspect weights and behavioral order of the criteria that have the highest score to the smallest score that can be seen in Tables XII, XIII and XIV as follows:

C. Determine the Profile Value

Profile values are alternative values for each criterion (range 1-5). This is the value obtained from the alternative, while the Criteria Weight is the expected value to be sought from the alternative. In this stage the value data of prospective employees will be generated in accordance with existing values in knowledge-based. The first value sought is value according to company norms. In psychological values and interviews on aspects of nature and behavior, intellectual aspects can be seen in Table XV as follows:

Administrative value determined by PT. Prima Grafika can be seen in Tables XVI, XVII, XVIII and XIX as follows:

The abbreviation information from 3 aspects and 11 sub aspects are Work Motivation (WM), Appearance (AP), Accuracy (AC), Speed (SP), Intelligence (IN), Language Mastery (LM), Problem Solving (PS), Diploma (DI), Domicile (DM), Work Experience (WE), Age (AG) can be seen in Table XX as follows:

TABLE. XII. SCORE RESULT OF WEIGHT ASPECTS OF NATURE AND BEHAVIOUR

Code	Criterion	Score
AI01	Work's motivation	5
AI02	Appearance	3
AI03	Accuracy	2
AI04	Speed	1

TABLE. XIII. SCORE RESULT OF INTELLECTUAL ASPECT

Code	Criterion	Score
AII01	Intelligence	5
AII02	Problem solving	1
AII03	Mastery of language	3

TABLE. XIV. SCORE RESULT OF WEIGHT ASPECT OF ADMINISTRATIVE

Code	Criterion	Score
AIII01	Education / Diploma	5
AIII02	Domicile	1
AIII03	Work experience	3
AIII04	Age	1
AIII01	Education / Diploma	5

TABLE. XV. PSYCHOLOGICAL TEST AND INTERVIEW SCORE

Category	Score
Very good	5
Good	4
Good enough	3
Enough	2
Lacking	1

TABLE. XVI. WEIGHT OF EDUCATION VALUE

Category	Score
S2	5
S1	4
D3	3
High school (SMA)	2
Junior high (SD-SMP)	1

TABLE. XVII. WEIGHT OF DOMICILE VALUE

Location distance (Km)	Score
≤ 2	5
3-5	4
6-7	3
8-9	2
≥10	1

TABLE. XVIII. WEIGHT OF WORK EXPERIENCE VALUE

Work experience (years)	Score
>10	5
5-10	4
3-4	3
1-2	2
0 th	1

TABLE. XIX. WEIGHT VALUE FOR AGE

Age	Score
18-20 th	5
21-25 th	4
26-30 th	3
31-35 th	2
>35 th	1

TABLE. XX. ALTERNATIVE VALUE

KODE		Sub-criterion value									
	WM	AP	AC	SP	IN	LM	PS	DI	DM	WE	AG
A001	SB	B	B	B	B	CB	B	S1	6	3	21
A002	B	B	B	CB	SB	SB	SB	S1	6	5	21
A003	CB	B	SB	B	B	B	SB	S1	6	3	21
A004	B	B	B	B	SB	B	B	S1	6	5	24

Profile value is the weight of alternative values from the conversion results of sub criteria evaluation can be seen in Table XXI.

D. Calculation Process of Profile Matching

After knowing the results of the Criteria Weighting, the next step is to calculate using Profile Matching. In the calculation process of this study the authors take the example of four prospective employees who will be made simulations for the calculation process, can be seen in Table XXII.

Example Explanation Calculation of the nature and behavior aspects based on the values that have been converted to the weight value of each sub-criterion, the column section is the code Aspect criteria and the row section is an alternative that has been determined by PT. Prima Grafika, can be seen in Tables XXIII, XXIV and XXV.

TABLE. XXI. PROFILE VALUE

CODE	Score of sub-criterion										
	WM	AP	AC	SP	IN	LM	PS	DI	DM	WE	AG
A001	5	4	4	4	4	3	4	4	3	3	4
A002	4	4	4	3	5	5	5	4	3	4	4
A003	3	4	5	4	4	4	5	4	3	3	4
A004	4	4	4	4	5	4	4	4	3	4	4

TABLE. XXII. LIST OF SYSTEM TEST ALTERNATIVE

No	Kode	Alternatif
1	A001	Cantika Dewi
2	A002	Eprriadi
3	A003	Rika Novriani
4	A004	Arif yulkianto

TABLE. XXIII. SCORE OF NATURAL AND BEHAVIOUR ASPECT

Code	AI01	AI02	AI03	AI04
A001	5	4	4	4
A002	4	4	4	3
A003	3	4	5	4
A004	4	4	4	4
Criterion value	5	3	2	1

TABLE. XXIV. SCORE OF INTELLECTUAL ASPECT

Kode	AII01	AII02	AII03
A001	4	3	4
A002	5	5	5
A003	4	4	5
A004	5	4	4
Criterion value	5	1	3

TABLE. XXV. SCORE OF ADMINISTRATIVE ASPECT

Kode	AIII01	AIII02	AIII03	AIII04
A001	4	3	3	4
A002	4	3	4	4
A003	4	3	3	4
A004	4	3	4	4
Criterion value	5	1	3	1

Examples of calculation information from Table XXVI are in Table XXIII Alternative code A001 with criteria A101 is the difference from code A001 with the criteria value AI01, namely: weight value of 5 with the gap (gap) criterion value 5 = 0, Code A001 with the value of criterion A102 is the difference from Code A001 with AI02 criteria, namely: weight value 4 with a gap value of criterion value 3 = 1, Code A001 with a criterion value A103 is the difference from code A001 with a criterion of AI03 namely: weight value 4 with a gap (gap) value of criterion 2 = 2 , Code A001 with criteria value A104 is the difference from code A001 with criteria AI04, namely: the weight of the value 4 with the difference (gap) value of criteria 1 = 3 as in Tables XXVI, XXVII and XXVIII.

Example calculation from Table XXVIII that is based on the competency difference in Table IV the gap value weighting. The weight values are taken from Tables XXVI, XXVII and XXVIII. The results can be seen in Tables XXIX, XXX and XXXI.

TABLE. XXVI. CALCULATION OF MAPPING THE GAP OF NATURE AND BEHAVIOUR ASPECT

Kode	AI01	AI02	AI03	AI04
A001	0	1	2	3
A002	-1	1	2	2
A003	-2	1	3	3
A004	-1	1	2	3

TABLE. XXVII. CALCULATION OF MAPPING THE GAP OF INTELLIGENCE ASPECT

Kode	AII01	AII02	AII03
A001	-1	2	1
A002	0	4	2
A003	-1	3	2
A004	0	3	1

TABLE. XXVIII. CALCULATION OF MAPPING THE GAP ADMINISTRATIVE ASPECT

Kode	AIII01	AIII02	AIII03	AIII04
A001	4	3	3	4
A002	4	3	4	4
A003	4	3	3	4
A004	4	3	4	4
Criterion value	5	1	3	1

TABLE. XXIX. WEIGHTING GAP'S SCORE OF NATURE AND BEHAVIOUR ASPECT

Kode	AI01	AI0	AI03	AI04
A001	5	4.5	3.5	2.5
A002	4	4.5	3.5	3.5
A003	3	4.5	2.5	2.5
A004	4	4.5	3.5	2.5

TABLE. XXX. WEIGHTING GAP'S SCORE OF INTELLIGENCE ASPECT

Kode	AII01	AII02	AII03
A001	4	3.5	4.5
A002	5	1.5	3.5
A003	4	2.5	3.5
A004	5	2.5	4.5

TABLE. XXXI. WEIGHTING GAP'S SCORE OF ADMINISTRATIVE ASPECT

Kode	AIII01	AIII02	AIII03	AIII04
A001	4	3.5	5	2.5
A002	4	3.5	4.5	2.5
A003	4	3.5	5	2.5
A004	4	3.5	4.5	2.5

For example, calculation of Table XXXII in code A001 with criterion A101, criterion A102 Criteria A103, Criteria A104 value from Table XXVIII. The NCF value (core factor value or main factor) known in the AI01 code criteria has the greatest weight and is made the main factor in the Aspects of Nature and Behaviour. NSF value (Secondary Factor Value) is the average value of the Criteria AI02, AI03, AI04 = 4.5.3.5.2.5 = 3,500. Total Value = 5,000 (NCF) x 60% (core value Main Percentage of company HRD provisions) + 3,500 (NCF) x 40% (secondary value Supporting percentage of company HRD provisions) = 4,400 as Tables XXXII, XXXIII and XXXIV.

TABLE. XXXII. FACTOR CALCULATION OF NATURE AND BEHAVIOUR ASPECT

Kode	AI01	AI02	AI03	AI04	NCF	NSF	Total
A001	5	4.5	3.5	2.5	5.000	3.500	4.400
A002	4	4.5	3.5	3.5	4.000	3.833	3.933
A003	3	4.5	2.5	2.5	3.000	3.167	3.067
A004	4	4.5	3.5	2.5	4.000	3.500	3.800
	Core	Secondary	Secondary	Secondary			

TABLE. XXXIII. FACTOR CALCULATION OF INTELLIGENCE ASPECT

Kode	AII01	AII02	AII03	NCF	NSF	Total
A001	4	3.5	4.5	4.000	4.000	4.000
A002	5	1.5	3.5	5.000	2.500	4.000
A003	4	2.5	3.5	4.000	3.000	3.600
A004	5	2.5	4.5	5.000	3.500	4.400
	Core	Secondary	Secondary			

TABLE. XXXIV. FACTOR CALCULATION OF ADMINISTRATIVE ASPECT

Kode	AIII01	AIII02	AIII03	NCF	NSF	Total
A001	4	3.5	5	4.000	3.667	3.867
A002	4	3.5	4.5	4.000	3500	3.800
A003	4	3.5	5	4.000	3.667	3.867
A004	4	3.5	4.5	4.000	3.500	3.800
	Core	Secondary	Secondary			

E. Final Calculation

In this step, all aspect values are calculated according to the percentage applicable to each aspect and calculated based on alternatives and their Subcritical Aspects.

Example of Final Results calculation in Table XXXV in Alternative A001 (Cantika Dewi) Aspects of nature and behaviour with a value of 4.4 (Total NCF and NSF results), intellectual 4 (Total NCF and NSF results), Administrative = 3,867 (NCF and NSF total results) values total = 4.4x40 (Percentage determined by the company) + 4x30 (Percentage determined by the company) + 3,867x30% (Percentage determined by the company) = 4.12 and the ranking results can be seen in Table XXXV sorted by Weight as the following:

F. Ranking Result

In ranking, from the results of the final calculation that has been done, then the alternative with the highest value will get the first rank, from the calculation results above, the ranking results obtained in Table XXXVI as follows:

G. Program Implementation

Making a decision support system in selecting prospective employees is designed in the form of a web-based application. The following is a display of the forms used in the prospective employee selection system.

TABLE. XXXV. FINAL CALCULATION

Alternative	Aspect			Score Total
	Nature and Behavior	Intelligence	Administrative	
A001-Cantika Dewi	4.4	4	3.867	4.12
A002- Eprriadi	3.8	4.4	3.8	3.98
A003-Rika Novriani	3.933	4	3.8	3.913
A004-Arif Yulkianto	3.067	3.6	3.867	3.467
Percentage	40 %	30 %	30%	

TABLE. XXXVI. RANKING RESULTS

Alternative	Total Score	Rank
A001-Cantika Dewi	4.12	1
A004 -Arif Yulkianto	3.98	2
A002- Eprriadi	3.913	3
A003- Rika Novriani	3.467	4

1) Display login page: Is the first appearance of the application when it will enter the system, in this form, the user or user is asked to enter a username and password as seen in Fig. 5.

2) Home page display: The home page is the first display when a user or user logs in to the system as seen in Fig. 6.

3) Alternative page: Alternative pages are pages that contain alternative data or prospective employees to be selected as shown in Fig. 7.

4) Aspect page: An Aspect Page is a page that contains aspect names and percentages that will be used in calculations as seen in Fig. 8.

Fig. 5. Login Menu.

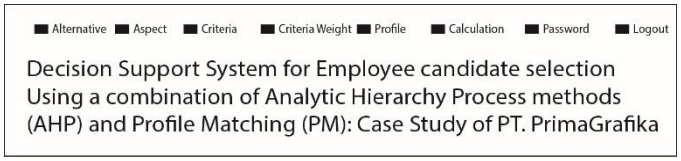


Fig. 6. Homepage.

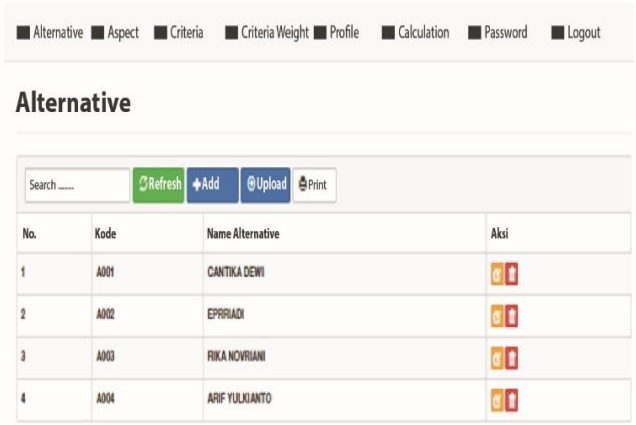


Fig. 7. Alternative Page.

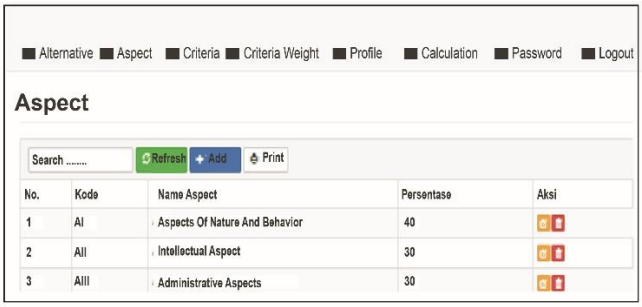


Fig. 8. Aspect Page.

5) Criteria page: Criteria page or also called Sub Aspect is a page that contains criteria that will be used to carry out the employee selection process that can be seen in Fig. 9.

6) Criteria weights page: Criteria Weights Page is the weight that will be used in the process of calculating the selection of prospective employees obtained from the level of importance in the AHP method as seen on Fig. 10.

7) Profile page: Profile pages are alternative values for each criterion (range 1-5). This is the value obtained from the alternative as can be seen on Fig. 11.

8) Calculation page: This Calculation page discusses the calculation process from weighting to the final results in the form of ranking from rank one to the end which shows the quality of prospective employees who will be chosen by the company as seen in Fig. 12.

9) Password page: This application provides a Password page that can be used by Application users to change old passwords with new ones, for the purpose of system security. The page can be seen in Fig. 13.

10)Log out menu: Log out menu is used to signed out from system as seen in Fig. 14.

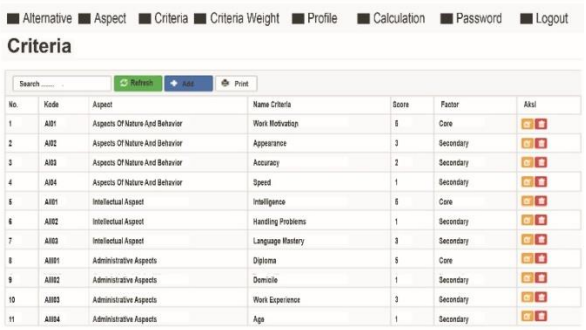


Fig. 9. Criteria Page.

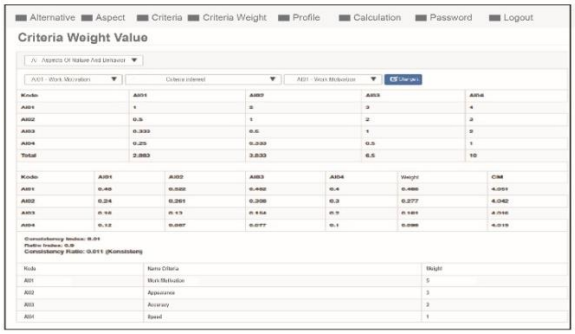


Fig. 10. Criteria Weights Page.

Fig. 11. Profile Page.

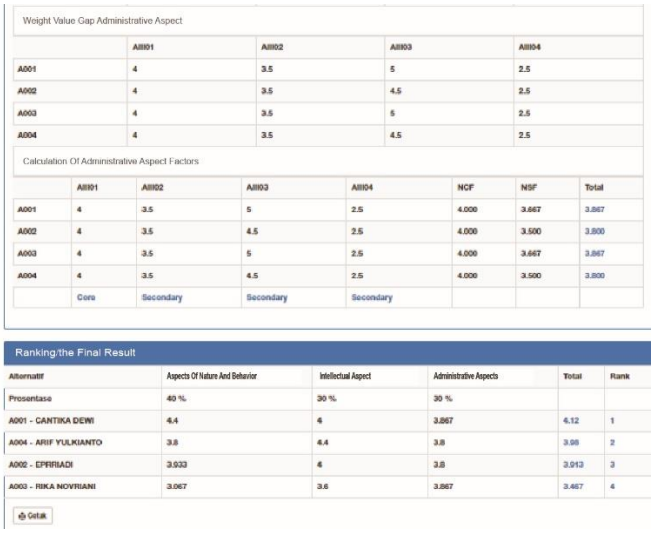


Fig. 12. Calculation Page.

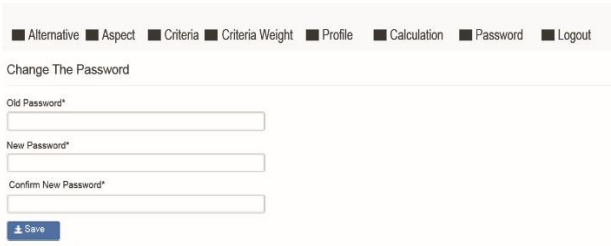


Fig. 13. Password Page.

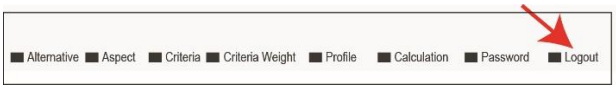


Fig. 14. Log Out Menu.

H. Systems Testing

This validation test was carried out to assess the design specifications as expected or not according to user requirements, validity testing in this study used User Acceptance Testing (UAT) and Delone and McLean Model while testing the ability of the system was carried out to test the software produced, Black Box Testing.

Black Box Testing, called Behavioral Testing, focuses on the functional requirements of the software. That is, the Black Box Testing technique makes it possible to obtain a set of input conditions that will fully implement all functional requirements for a program.

User Acceptance Testing (UAT), which is a system testing conducted by the user of the system to ensure the system is running well according to user needs.

Delone and McLean Model, namely, Delone and Mclean which states that information quality, system quality and service quality will positively influence use and user satisfaction and subsequently will have a positive effect on net benefits or final results.

Back box testing using questionnaire data on 60 respondents can be seen in Table XXXVII as follows:

TABLE. XXXVII. BLACK BOX TEST TYPE

NO	Class test	Test unit	Test result
1	Login testing	Inputting Username & Password	Succeed
2	Alternative testing	Adding alternative data	Succeed
		Delete alternative data	
		Change alternative data	
		Save alternative data	
3	Aspect testing	Add criterion data	Succeed
		Delete criterion data	
		Change criterion data	
		Save criterion data	
4	Criterion testing	Input criterion data	Succeed
		Change criterion data	
		Save criterion data	
5	Criterion's weight testing	Change weight comparison	Succeed
6	Calculation process	Calculation/ranking	Succeed
7	Menu password	Password change	Succeed
8	Menu logout	Logout from system	Succeed

UAT testing from questionnaire results from 10 respondents can be seen on Table XXXVIII.

Testing using Delone and Mclean model using questionnaire from 10 respondents can be seen on Table XXXIX as follows:

The results of the testing using three different types of testing which are Black Box Testing, User Acceptance Testing (UAT) and Delone and McLean Model by giving respondents set of questionnaire yield a satisfying result. Back box testing in the decision support system combination of AHP and Profile Matching can be received and run very well.

TABLE. XXXVIII. CONCLUSION OF USER ACCEPTANCE TESTING *(UAT)*

No	Aspects	Actual score	Ideal score	Total score %
1	Perceived Ease of Use	220	250	88%
2	Perceived Usefulness	219	250	87,6%
3	Attitude Toward Using	216	250	86,4%
4	Behavioral Intention To Use	124	150	82,6%
Total		779	900	86.1%

TABLE. XXXIX. CONCLUSION OF DELONE AND MCLEAN MODEL TESTING

No	Aspect	Actual score	Ideal score	Total score %
1	Information Quality	210	250	84%
2	System Quality	213	250	85.2%
3	Service Quality	130	150	86.6%
4	Use	86	100	86%
5	User Satisfaction	87	100	87%
Total		726	850	85.7%

V. Conclusions and Suggestions

A. Conclusions

Based on the description explained, the conclusions obtained from this paper are as follows:

By using a combination of Analytic Hierarchy Process (AHP) and Profile Matching (PM) methods, the selection process of prospective employees can be carried out well. In the employee selection process a lot of data and the same name can be processed using a database-based system, to avoid data duplication, or file exchange.

Based on system testing, using three different methods of testing resulted that system can be accepted by the company, User Acceptance Testing (UAT) obtained from respondents totalling of 10 people with an ideal score of 900 are given an actual score of 779 or 86.1%, overall the system is acceptable. Delone and McLean Model Test results obtained from 10 respondents with an ideal score of 850 produced an actual score of 726 or 85.7%, overall the quality of this system is running very well. With the results of ranking: A001 - Cantika Dewi = 4.12, A004-Arif Yulkianto = 3.98, A002 Eprriadi = 3,913, A003-Rika Novriani = 3,467.

B. Suggestions

To create a support system for the Analytic Selection Process for Prospective Employees (AHP) and Profile Matching (PM) Methods, better in the implementation phase, it needs support from all parties, both from PT.Prima Grafika or related staff and need to add software testing another to actualize the use of employee selection.

References

[1] M. Azman Maricar ,et al, 2016, Decision Support System of the Employees Acceptance using Analytical Hierarchy Process (AHP) and Multi Factor Evaluation Process (MFEP)International Journal of Engineering and Emerging Technology, Vol. 1, No. 1.

[2] Susilowati, T. et al. 2018. "Using Profile Matching Method to Employee Position Movement." International Journal of Pure and Applied Mathematics 118(7 Special Issue).

[3] Ngurah IG, Partha W, Novenando P, Weking M, Sudarmojo YP. Designing a Decision Support System for the Best Employee Selection Using AHP Method Case Study PT . Z Bali. 2018;3(2).

[4] Moedjiono, et al, 2016, Decision Support Model for User Submission Approval Energy Partners Candidate Using Profile Matching Method and Analytical Hierarchy Process ,Scientific Journal of Informatics Vol. 3, No. 2, ISSN 2407-7658.

[5] (Shiddieq & Septyan, 2017) Comparative Analysis of Ahp and Saw Methods in Employee Performance Evaluation (Case Study in Pt. Grafindo Media Pratama Bandung).

[6] Tata, D., Lumban, S., & Syahrizal, M. (2018). Decision Support System for Selecting Wireless Router using Promethee Method (Case Study: My Republic Medan), 17, 240–244.

[7] ((Sudrajat, 2019) Selection of Outstanding Employees Using the Method Profile Matching.

[8] Heny Pratiwi Decision Support System For Determination Of Employees Perform Use Multi Factor Evaluation Process Method). Journal of Information Systems, September 2014; Vol.5 No.2: p.95-101.

[9] Kristiana, Titin. 2015. Application of Profile Matching For Performance Appraisal Civil Servants (PNS). ISSN: 1978-1946. Jakarta: Nusa Pillar Journal Mandiri Vol. XI, No. September 2, 2015: 161-170.

[10] Shaout, A & Yousif, M. K, (2014) "Performance Evaluation –Methods and Techniques Survey" :International Journal of Computer and Information Technology (ISSN:2279 – 0764), Volume 03 – Issue 05.

[11] Sina Khanmohammadi And Mandana Rezaeiahari. AHP Based Classification Algorithm Selection For Clinical Decision Support System Development. Procedia Computer Science, 2014; p.328-334.

[12] El, M., El, A., Segers, J., & Pais, I. 2016. Computers in Human Behavior Technology adoption in employee recruitment: The case of social media in Central and Eastern Europe. Computers in Human Behavior, 57, 240–249. http://doi.org/10.1016/j.chb.2015.12.043.

[13] Suvendu Chanda Nayak and Chitaranjan Tripathy. Deadline Sensitive Lease Scheduling in Cloud Computing Environment Using AHP. Journal of King Saud University - Computer And Information.

[14] Herman Firdaus, Indra, Gunawan Abdillah, Faiza Renaldi. 2016. Decision Support System for Determination of TerGood Employees Using AHP and TOPSIS Methods. National Seminar on Information and Communication Technology 2016 (SENTIKA 2016) Yogyakarta.

[15] Handayani, Rani Irma (2017). Supporting Decision System for Employee Achievement Selection with Profile Matching Method in Pt. Sarana Inti Persada (Sip). Pilar Nusa Mandiri Journal, Volume 13 No.1 March 2017, Number 28-34.

[16] Kristiana, Titin (2015). Application of Profile Matching for Performance Assessment of Civil Servants (PNS). Pilar Nusa Mandiri Journal, Vol.XI No.2 September 2015, Number 161-170.

[17] Fashoto,et al,2018,Development of A Decision Support System on Employee Performance Appraisal Using AHP Model, Department of Computer Science, University of Swaziland Vol 2 No 4, ISSN: 2549-9610.

[18] ((Sudrajat, 2019) Selection of Achievement Employees Using the Profile Matching Method.

[19] E Cruz-Cunha, M. M., Silva, J. P., Gonçalves, J. J., Fernandes, J. A., & Ávila, P. S. (2016). ERP Selection using an AHP-based Decision Support System. Information Resources Management Journal, 29(4), 65–81. doi:10.4018/irmj.2016100105.

[20] Rubayet Karim, and C. L Karmaker, "Machine Selection by AHP and TOPSIS Methods." American Journal of Industrial Engineering, vol. 4, no. 1 (2016): 7-13. doi: 10.12691/ajie-4-1-2.

[21] Xiaohui Liang 2013. "Definisi Profile Matching .Jakarta.

Analyzing the Impact of Forest Cover at River Bank on Flood Spread by using Predictive Analytics and Satellite Imagery

Muhammad Aneeq Yusuf[1], Dr. Muhammad Khalid Khan[2], Dr. Tariq Mahmood[3]
Muhammad Umer[4], Rafi Ullah Afridi[5]
College of Computing and Information Sciences
Karachi Institute of Economics and Technology, Karachi, Pakistan[1, 2, 4, 5]
Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan[3]

***Abstract*—Floods have been a recurring problem for a number of countries around the world including Pakistan. It is believed that densely populated forests at river banks can prevent floods from spreading towards settlements and farmlands. The role of forest in flood spread has been an area of research for a while but the role of predictive modeling in this area is yet to be investigated in detail. In this study, we have used predictive analytics and satellite imagery to develop an environmental model that can predict the flood spread by considering forest cover at river bank and month of the year as parameters. We have used the satellite images of an area situated in the northern region of Pakistan i.e. Dera Ghazi Khan from the USGS's Land Sat program. These images comprised of a section of the Indus River and its adjoining areas. We want to analyze the forest bank at various section of the Indus River. We developed and trained our predictive model by using the satellite imagery data and tested it on a separate dataset to determine error percentage. The model showed significant promise and predicted the flood spread with an average accuracy of above 80%.**

Keywords*—Floods; forests; predictive modeling; satellite imagery; environmental modeling

I. Introduction

Natural disasters are known to adversely affect population on a massive scale. It has been well documented that the impact of these disasters is more severe in developing countries, as compared to the better developed counterparts because of the disaster recovery services [1]. In such countries, large portion of the population lives below the poverty line and the impact of natural disasters is tremendous and everlasting [1]. Natural disasters can be of many types but in this study we are more focused on the impact of flood spread.

Floods have been a recurring problem for a number of countries around the world, particularly ones where farmers rely on irrigation for agriculture, such as Pakistan. Even though the country has a hot and dry climate, it receives ample rainfall during the months of July-August and December-January [2]. This often results in floods, as has been seen between the years 2003 and 2014. In a report by the Asian Development Bank, Pakistan is expected to experience increased variability in river flows. This is due to rise in temperature and melting of glaciers. Moreover, the report also states that average annual rainfall is not expected to have any long-term trend [2]. This tends to create a high level of uncertainty for the country, with respect to floods. As such, Pakistan needs to take measures to prevent high-river flows from becoming disastrous.

The country has already experienced a devastating flood in 2010, which has been followed by a few smaller floods, in terms of impact [3]. The floods in 2010 alone were responsible for over 1,500 deaths and internal displacement of another 15 million people [3]. The main cause for this disaster was a few days of severe rainfall, at the end of July. It has been documented that this was mainly due to anomalous behavior, within the upper-atmospheric air. Moreover, it has been established that such atmospheric events, although rare, are a part of climatic events that occur over Pakistan [3]. However, this does not explain the floods that followed in subsequent years. As such, it would be correct to state that Pakistan is ill-equipped to deal with high-flows in its rivers, in times of heavy precipitation.

Furthermore, recently developed models of climatic conditions, particularly those of Pakistan modeled over a period of 30-years, have showed that the country is likely to experience wetter conditions. However, the coastal areas and the mountain ranges are expected to have a weak drying signal [3]. As the world moves towards a warmer climate, as a result of rising levels of greenhouse gases, the river-flows in the country are also expected to rise, as a result of faster glacial melting [4]. The importance of flood-risk assessment and flood prediction is of significant importance in such a hostile environment. The government has also been investing heavily in technology to help predict such disasters and provide better management skills, in case of such an event [4]. It is important to note here that other countries, with river systems, also face similar conditions [17].

It has been documented that the high-level of deforestation has been one of the major factors, contributing to the recurrence floods [1]. The country's population makes use of the wood for timber and fuel. However, the uncontrolled manner in which this is done, tends to greatly reduce the water retention capacity of the forests [1]. The forests which aren't dense enough fails to prevent the water from reaching the rural

populace. In addition to this, deforestation further leads to an increased surface water runoff and soil erosion, further clearing the way for the water to reach populated areas, in times of high-flows [1].

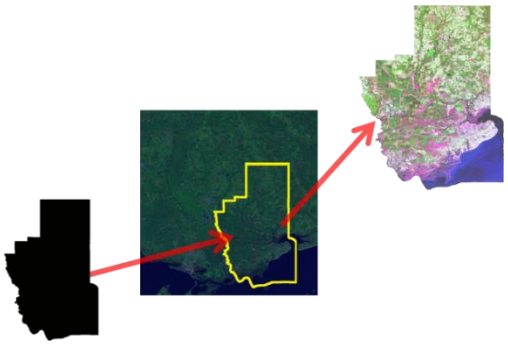
II. Approach

By using various image processing and machine learning techniques, we can predict the flood spread. Here we introduce few techniques that are used in this work. Image masking technique is used to extract the area of interest from the satellite image (Fig. 1(a)).

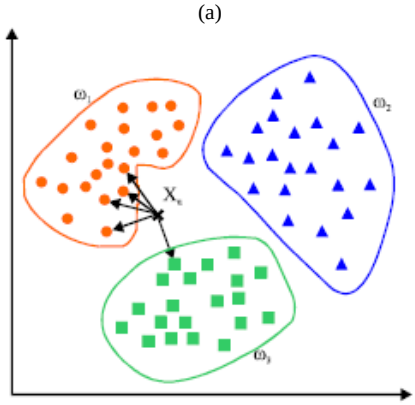
This is done by identifying the start and end points of the area of interest, in the form of pixel-row and pixel-column number, while setting the values for all other pixels to 0 [5]. This tends to keep the pixel values of the area of interest intact, while neglecting the effect of everything else in the image. The technique is particularly useful for feature extraction and the processing of very large images. Next, the classification of image pixel is required and K-Nearest Neighbor Algorithm is used to achieve it.

K-Nearest Neighbor (Fig. 1(b)) classifies data sets based on their similarities with their neighbors and the k stands for number of data classes that are considered for classification [6]. This algorithm is based on supervised learning the basic idea of its working is:

- An object which is a new instance will be classified from majority votes given by its neighbors [7].
- It is measured by distant function so that the object is assigned to the most common class among its K nearest neighbors [7].



(a)



(b)

Fig. 1. (a) Image Masking Visualization (b) KNN Visualization.

At the end, a model is developed by using predictive analytics techniques. Predictive Analytics can be defined as a set of processes that make use of data-mining and probability, to forecast outcomes [8]. This is done by identifying a set of predictors, also termed as known variables, and using them to develop a statistical model [8]. It should be noted here that once developed, the model can be extended and improved, once additional data, relating to the problem becomes available. Moreover, some models, excluding computational models, can be defined using a mathematical equation. This can then be used further to map out how exactly the predictors effect the outcome of a phenomenon, over time [8]. Computational models do not have a single defined equation, since they are developed using a number of different simulations, until a certain level of accuracy is reached.

It is clear that floods tend to be a devastating problem for Pakistan, recovery from which takes long-periods of time. It is also evident that various image processing and predictive analytics algorithms can help in analyzing the impact of forest cover at river banks. In this study, we have used predictive analytics and satellite imagery to develop a model that can predict the flood spread by considering forest cover at river bank and month of the year as parameters. We developed a model to predict how far the next flood would spread if given the area of the forest cover at river bank. In the subsequent section, we will present the related work and Section 4 explains the methodology in more detail. Sections 5 and 6, will discuss the results and conclusion respectively.

III. Related Work

Given Pakistan's history, with respect to floods, the natural disaster is well documented and extensively researched. However, the majority of these studies have been centered on determining the causes of the disaster, ways of better management, if and once the disaster strikes and determining the damage and patterns of floods in the past. Similarly, a number of researches have been conducted, focused on different approaches and machine learning methods for image processing and detecting and observing change in land-cover. A lot of this work is focused on detecting, quantifying and managing deforestation but currently, no known model exists, which makes use of forest cover around river bank to predict the risk of floods.

Author in [18], served as an example of how temporal changes, in land cover, are monitored using satellite imagery, from Landsat. The authors made use of imagery spread across a period of 40 years and performed a descriptive analysis of the changing land cover. However, in this case the researchers made use of cellular automata to develop a predictive model, for the Bakhtegan Lake. The authors clarified that the change in the ecology of the lake was mainly due to extreme climate change conditions. It should be noted here that, even though the imagery was taken between 1973 and 2013, only 12 images were used for the actual model development and analysis.

In [9], authors focused on assessing the usability and accuracy of developing linear and generalized linear mixed models (GLMM), using various machine learning techniques on freely available datasets. The study established that

although Artificial Neural Networks tend to outperform GLMM in certain areas, its accuracy suffers in other regions. On the other hand, Bayesian Networks were found to perform similarly across different test zones. Similarly in [10], assessment is made about the usability of clustering algorithms for the automatic detection of cultivable land, within a satellite image. The authors made use of a technique called 'Partitioning Around Medoids' for this purpose. The training and testing was done on single band images, with a particular focus on determining which bands are best suited to detect cultivable land, using unsupervised learning. It is important to point out here that the images used for our research were composed of images from the three visible bands (Red, Green and Blue), overlayed on each other, in a single image. However, the data was obtained from the Landsat 8 missions, similar to the ones used in this study. It is also important to note here that the authors identified the need for improved techniques for feature selection, in order to improve the accuracy of clustering algorithms.

When it comes to classification problems, machine learning is usually the de-facto approach used. However, accuracy of such classifications is crucial, in order to ensure that the results carry weight and are true in the specified situations. As a result, we made use of [19] to determine how accuracy of satellite imagery is performed. The authors made use of a data-assisted labeling approach, over unsupervised classified images. This was done by generated N-spectral class maps, using the ISODATA approach, followed by the development of maps with reference data and then assessment and selection of maps with the highest accuracy. It is important to mention here that the authors were able to report success with this approach and a similar approach was used in our own research, as well.

How ANNs are being successfully used in various image processing applications, many of which require a high-level of accuracy is discussed in [11]. The authors of this paper made use of the 'Sigmoidal function', for transfer and learning of the network. However, most of the application areas identified by the authors have access to a large amount of computational power. This was something that we lacked and therefore were not able to utilize ANNs, or any other form of Neural Networks, for our application. It should be noted here that the examples used in this paper do not identify the size of the images or number of pixels per image, which are processed in the mentioned applications. In our case, each of the Landsat images had approximately 56 million pixels, on average. As such, training a network using the computing power available was infeasible. Similarly, [21] was studied to determine the feasibility and benefits of using neural networks for satellite image classification. This paper was particularly selected due to the fact that the authors used Landsat imagery for testing purposes. The authors made use of a number of differently configured neural networks, in order to observe the changes of adding and removing nodes and layers of the network. The research resulted in the discovery of an optimal pairing, which was capable of delivering rapid convergence and better classification. The results were compared with a maximum likelihood classifier and the neural network was found to be equally good, when identifying urban areas. It is important to note here that the authors made use of a section of the Landsat image. In our case, a machine learning technique was used, instead of neural networks, due to the processing requirements of classifying entire Landsat images, through a neural network.

A significant amount of work has also been done, during the last decade, to develop change detection models, for land cover mapping, using satellite imagery from MODIS and Landsat. It is important to note here that the spatial resolution of MODIS is $1km^2$, while that of Landsat is $30m^2$. In [12], the authors made use of the C5.0 algorithm (a variant of decision trees), to perform classification of regional land-covers across Central Asia, using time-series data from MODIS.

Furthermore, the authors analyzed data from the years 2001 and 2009, to observe seasonal features and how they change over-time. In addition to this, they also made use of high-resolution, remotely sensed imagery, to test their classifications. The results showed an accuracy of over 90%. In the following year, another study surfaced [13], where the authors used satellite imagery from Landsat 5 Thematic Mapper, to collect data from the Yellow River Delta, which is considered to be one of the fastest growing river deltas in the world. The research team made use of scenes from 1995, 2004 and 2010 to perform a comparative analysis of the area and detect any changes. Moreover, a per-pixel classification approach was employed using the C5.0 decision tree algorithm. By performing statistical analysis of the processed images, the research team was able to identify significant changes in the landscape of the delta. Moreover, the team was also able to determine intensified farming and urban sprawl as the main reasons for the said changes.

In [14], authors looked at the variations in stream flow and observed a number of records with anomalously high stream flow, which has resulted in flooding. It is important to note here that the hydrology of the region is dominated by snow accumulation and following melting processes, which result in times of high stream flow and low stream flow. The publication also pointed out that there was an imperfect understanding of the factors that contributed to the anomalies. The authors made use of the "Variable Infiltration Capacity" model to identify the key factors that contribute to this phenomenon, while also determining the influence of a set of predictors on these factors. The resulting model was then used to predict future stream flow in the basin. Similarly in [15], it is aimed to identify the synergies between the different types of changes in land-cover and how they affect the 'Catchment Scale'. Moreover, it was clarified that cross-disciplinary, long-term field studies are required to ensure major advances. It should be noted here that the research identifies forest management as a key-factor in drivers of flood-regime changes, which is also the main variable in our own prediction model. In addition to this, since we made use of a regression based approach in our research, [20] was reviewed to determine the outcome of using various types of regression techniques, to map landslide susceptibility. As is the case with most regression techniques, an inventory of the event occurrence and factors leading to that occurrence were used to develop the mapping models. The researchers identified the selection of events as an important factor in the accuracy of

the resultant models. A lot of emphasis, in this research, has been imposed on random sampling for model fitting. Following this, in our own research, we made use of sequential data, with both positive and negative points for flood events, to train our regression model.

In [16], authors focused on the use of large ensemble climate model simulations, in order to determine the causes of floods, in England. The authors were able to identify increased capacity for moisture in the atmosphere and anthropogenic warming as the main causes for increased westerly flow, in the Thames River. Combining these results with the catchment sensitivity results in a slight increase in properties, with respect to riverine flooding. However, this is also accompanied by a significant level of uncertainty. Furthermore, in [22] a group of researchers made an effort to improve the optimum-path forest clustering algorithm, by altering one of the characteristics of the main algorithm. Through this technique, the authors had aimed to make improvements to image segmentation techniques, for satellite imagery. It should be noted here that image segmentation is a key characteristic, when extracting features for target identification and object description.

It is evident that majority of the current work done is limited to observing the effects of floods and developing better mechanisms for flood management. However, significant progress is yet to be made, with respect to developing models for predicting future floods. Hence, predicting the flood spread by analyzing the forest cover at river bank is logical step in this direction.

IV. Methodology

The main objectives of this research is to convert information from satellite images into a quantifiable form and then use that data to develop and test a predictive model. We have used the satellite images of an area situated in the northern region of Pakistan i.e. Dera Ghazi Khan from the USGS's Land Sat program. These images comprised of a section of the Indus River and its adjoining areas as depicted in Fig. 2 but the problem is that these images covers lots of area which is of no interest to us and increase the processing burden. We used an image masking technique to extract our area of interest as depicted in Fig. 3. To get the pixel wise classification of the image, we used KNN algorithm and the output is depicted in Fig. 4. It should be noted here that in Fig. 4 the 'Water Class' is represented by blue, the 'Forest Class' has been represented by green and the 'Barren Land Class' has been represented by red.

The entire process was divided into a series of sequential steps, which are defined as follows; a graphical representation of the steps can be seen in Fig. 5:

- Feature Extraction: Once a satellite image is loaded into the system, the user identifies the X and Y coordinates of the training pixels, to be used for each of the three classes, namely, Water, Vegetation and Non-vegetation areas. This is done for each image.
- Image Slicing: Following the feature extraction step, the user would identify the limits of the X and Y coordinates, for his area of interest (the river in question and its immediately surrounding areas). Using these values, only the area of interest is extracted from the overall image, significantly reducing the image size and number of pixels to classify.
- Pixel Classification: This step makes use of the values identified in the first step and trains the pixels extracted in the second step, color coding them into the three classes required. This is done using the K-Nearest Neighbors classifier and results in a uniformly colored image.
- Pixel Count: The output from the previous step is then passed to the Pixel Count module, which identifies the number of pixels in each class. The values are then multiplied by the spatial resolution of the satellite image (mentioned in section 1.4), to provide the area covered by each of the three classes, in meters. This results in the information from the satellite imagery to be available in a numerical form, which can now be modelled.
- Data Analysis and Predictive Modeling: Now that the data is available in a quantifiable form, it can be modeled using a regression model. The model is then tested for accuracy and used for future predictions.

It was discussed in the related work that neural networks are infeasible to be used with large sale images; we have decided to use machine learning technique to conduct pixel wise classification of the satellite images. Further, clustering also didn't work for large scale images on average computer systems; we decided to make use of supervised learning technique, rather than unsupervised learning to maintain a higher degree of accuracy and speed-up the learning process. The K-Nearest Neighbors algorithm was used, keeping in mind the sheer size of the satellite images and the computational power available. Each of the images was passed through the first four steps, mentioned above, with the numerical data being compiled in a tabular format; the resultant table can be seen below, Table I. It is important to note here that the feature extraction and classification was done on each of the images, so as to maintain a high-level of accuracy. As such, we were able to obtain over 85% classification accuracy, in almost every image.

Once the data was converted to a numerical form, it was then visualized graphically to determine the patterns and scatter in the data, the resultant plots can be seen in Fig. 6, 7, 8 and 9. The results of the visualizations are given in section 4. After careful analysis of the visualizations, it was decided to make use of the 'Ridge's Regression' model due to scatter of data.

It is important to state here that given the spread of the data points in the scatter plots, it can be inferred that a linear regression model would not be an ideal fit. This is mainly due to the fact that fitting a linear regression line to such a plot would mean greater variations in error bars. However, Ridge's Regression fits a curve that is closer to the data points, resulting in lower error bars and hence a more accurate predictive model.

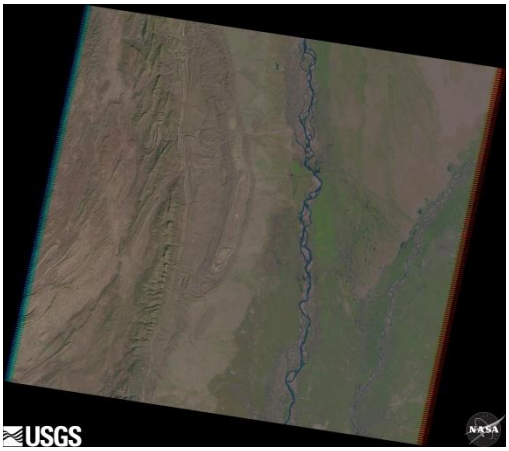


Fig. 2. Raw Satellite Image of North of Dera Ghazi Khan, Pakistan Showing River Indus and Associated Areas (Landsat).

Fig. 3. Extracted Area of Interest using Image Masking Technique.

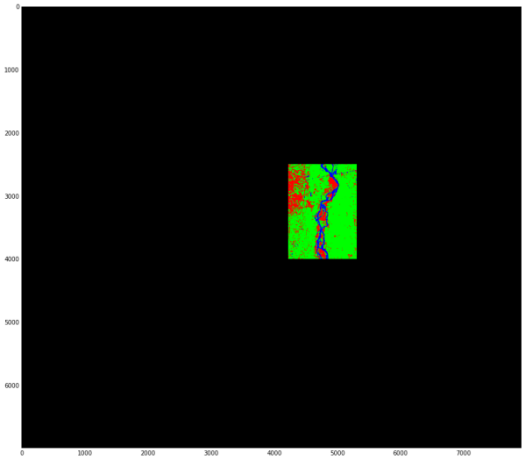


Fig. 4. Image Classification using KNN Algorithm.

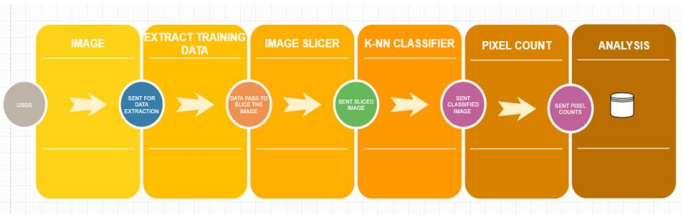


Fig. 5. Data Pipeline Diagram.

TABLE. I. DATA CREATED FROM SATELLITE IMAGES

	Date	Month	Soil Area (meters)	Vegetation Area (meters)	Water Area (meters)
1	2009-05-21	5	789138000	567756900	98784000
2	2009-06-06	6	787755600	538948800	128974500
3	2009-06-22	6	714393900	606996900	134288100
4	2009-07-08	7	414722700	887888700	153067500
5	2009-08-09	8	165952800	1142939700	146786400
6	2009-08-25	8	161583300	1096844400	197251200
7	2009-09-10	9	200644200	1138447800	116586900
8	2009-09-26	9	339753600	1019788200	96137100
9	2009-10-12	10	599943600	767756700	87978600
10	2009-10-28	10	739434600	643428000	72816300
11	2009-12-15	12	362314800	1043644500	49719600
12	2010-02-17	2	368703000	1025220600	61755300
13	2010-04-06	4	1193170500	205023600	57484800
14	2010-04-22	4	660550500	669238200	125890200
15	2010-05-24	5	268866000	1103776200	83036700
16	2010-06-09	6	1192885200	171452700	91341000
17	2010-07-11	7	625924800	557286300	272467800
18	2010-08-28	8	248742000	707883300	499053600
19	2010-11-16	11	267040800	1101443400	87194700
20	2010-12-02	12	244593000	1135185300	75900600
21	2011-01-19	1	345753900	1066625100	43299900
22	2011-02-20	2	591701400	753785100	110192400
23	2011-03-08	3	371706300	988533000	95439600
24	2011-03-24	3	738674100	637821900	79182900
25	2011-04-25	4	1260012600	110147400	85518900
26	2011-07-14	7	576993600	605738700	272946600
27	2011-10-02	10	354651300	979584300	121443300

In addition to this, a total of 27 data points are available to us for the training purpose as shown in Table I. During the analysis it was also revealed that forest cover around a river bank alone, would not be sufficient to accurately predict when and by how much a flood is expected to occur. This was mainly due to the fact that rivers do not flood their banks, unless there is heavy rainfall or extreme glacial melting at its source. As such a second independent variable was identified, in the form of "Estimated Rainfall". However, rainfall data for our specific area wasn't readily available. We made use of the "Month", when the image was taken, to serve as a supplemental variable for "Estimated Rainfall". This was mainly to accommodate seasonal variations, as closely as possible.

A total of three models were developed, two uni-variable regression models and one multivariate regression model. The aim here was to compare the three models and select the most optimal one for future use. Moreover, this allowed for determining the overall effect each variable has on the final model. Finally, another 6 data points were created, using the same techniques as identified earlier, for testing purposes. The accuracy of each of the three models was determined and averaged-out for better comparison. Table II shows the end results of this analysis and testing phase.

A. Training Data Blue (Water) and Green (Vegetation)

The histogram in Fig. 6 shows a graphical representation of the data extracted from the processing images, from the training set. There was an abundance of vegetation in the concerned area, as compared to water. However, during the exploratory data analysis, it was seen that readings no. 16 and 17 show an increase in water area, whilst there was a decline in vegetation around the river bank. However, this cannot be taken as a causation effect and further analysis and modeling is necessary.

B. Testing Data Blue (Water) and Green (Vegetation)

A similar histogram was developed for the test dataset and the results are shown in Fig. 7. As can be seen, despite the decline in vegetation, in readings no. 3 and 4, there was not a significant change in the area of water. This is why it was necessary to add another variable, in the form of months. It should be noted here that this was done to include the effect of seasonal change.

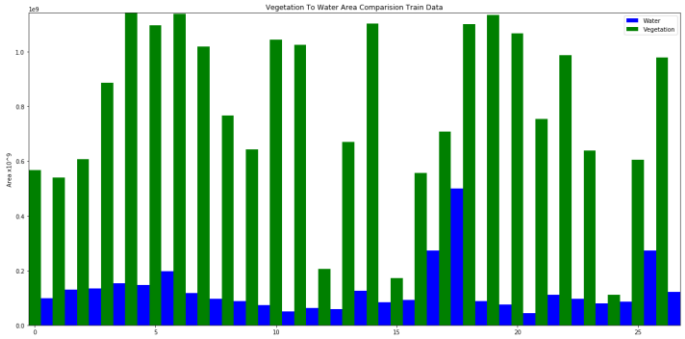


Fig. 6. Train Data Visualization.

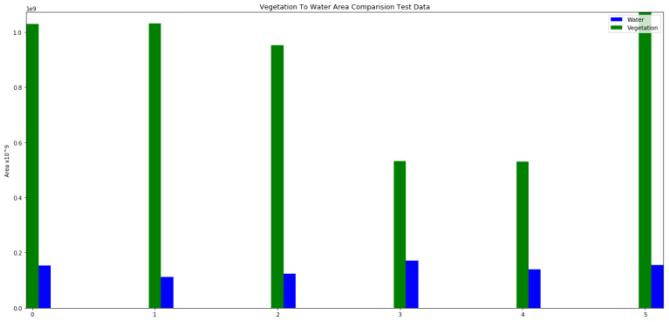


Fig. 7. Test Data Visualization.

C. Scatter Plot (Water - Vegetation)

In order to further analyze the data, a scatter plot (Fig. 8) of the dependent and independent variables was also visualized. It was revealed that the data was skewed towards the y-axis, with just three outliers. This meant that there were just three instances in the data which indicated flood conditions in the region. However, there was significant variation in vegetation.

D. Scatter Plot (Water - Month)

Much like the plot above, the scatter plot of months against water area (Fig. 9) showed little variability. It is important to note here that for the purpose of analysis the water area in similar months was averaged out; hence there are just 12 data points, instead of 27, for the training data. Further analysis of the plot revealed that seasonal variations do seem to contribute to the phenomenon and are mostly active between June and September.

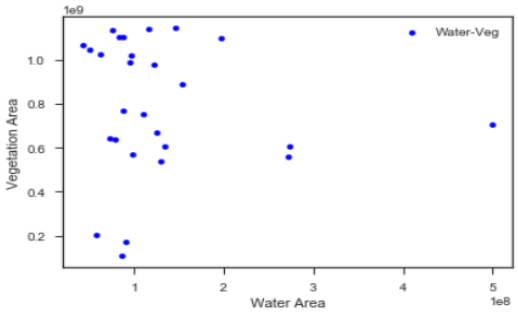


Fig. 8. Scatter Plot Vegetation Area against Water Area.

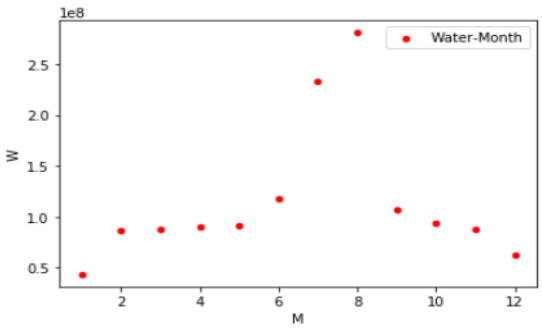


Fig. 9. Scatter Plot Water Area against Month.

V. Results

The following table (Table II) highlights the results of each of the six test cases. The percentage accuracy of each model, for every data point can be seen. As is visible, there was significant variation in the predicted and actual values of the test points, in all of the three models. The final row contains the percentage average error of each of the three models. When averaged out, the multiple-ridge-regression model was the most accurate one. However, to justify the model, an R-Squared test was also conducted on each of the three models. The 'Ridge Vegetation' model had an R-Squared score of 0.0082, as compared to 0.0180 for 'Ridge Month' and 0.0223 for 'Ridge Multivariable' models. Although the R-Squared values were quite low for each of the three models, it can largely be attributed to the small number of data points available.

Moreover, it was also the model with the least errors for individual tests, with the exception of reading no. 1. It can also be concluded from the given table, that month alone cannot be used as a predictor variable, since it has a high level of variation, in percentage error. Moreover, it can also be concluded that forest cover alone, can be used as a predictor variable, but it will contain a high degree of error. Therefore, a secondary variable needs to be added to make the model more accurate.

TABLE. II. Percentage Error in Predictive Models; RV=Ridge Vegetation, RM=Ridge Month, MVM=Multivariate Model

Date	Vegetation Area (1000's)	Water Area (1000's)	% Error (RV)	% Error (RM)	% Error (MVM)
23/08/14	10301163	153527	21.23%	13.35%	30.30%
24/09/14	1032350	111824	8.09%	22.53%	3.43%
10/10/14	952204	123246	0.14%	14.41%	6.40%
07/06/15	532970	171257	21.42%	26.97%	17.99%
26/08/15	530257	139664	3.60%	4.75%	0.56%
27/09/15	1072591	156467	23.46%	12.43%	10.24%
Mean Error:			12.99%	15.74%	11.49%

VI. Conclusions

From the previously stated results and careful analysis, it can be concluded that the "Multiple-Regression-Model" does show a lot of promise for accurately predicting how far a flood is expected to spread, given the amount of forest cover on the river bank and seasonal conditions. It is important to note here that there have not been any significant flood conditions in the region, following the floods in 2010. As such, it is yet to be tested if the model would hold true, in such a condition.

However, given that the number of data points used is quite small, further improvements can be made with the model, by fitting it over a larger dataset. It should be noted here that with the amount of computing power available to us, it would have been difficult to create more datapoints in a quick and efficient way. As such, datapoints from a 3-year period was used to perform model fitting, which could have resulted in over-fitting of the data. It is recommended that datapoints from a 10-year period be used for the creation of a more accurate model.

Moreover, given the budget constraints, it was not possible to make use of high-resolution satellite imagery, which led to the use of intuitions, rather than concrete data. This could have also resulted in over or even under-fitting of the data. Added to this, the constraints also made it difficult to obtain the complete rainfall data for the said 3-year period. It is recommended that satellite imagery with a spatial resolution of at least 10x10 meters be used for improving the model. In addition to this, the use of rainfall data, instead of months, would help improve the quality of the results.

References

[1] Sayeeda Amber Sayed, Pedro Arcos González. (2014). Flood disaster profile of Pakistan: A review. Science Journal of Public Health. 2 (3), 144-149.

[2] Asian Development Bank (2017). Climate Change Profile of Pakistan. Chaudry, Qamar Uz Zaman.

[3] G. van der Schrier, L. M. Rasmijn, J. Barkmeijer, A. Sterl, W. Hazeleger. (2018). The 2010 Pakistan floods in a future climate. Climatic Change. 148, 205-218.

[4] Muhammad Iqbal Javed Akhter, Muhammad Irfan, Naeem Shahzad, Rehan Ullah. (2017). Community Based Flood Risk Reduction: A Study of 2010 Floods in Pakistan. American Journal of Social Science Research. 3 (6), 35-42.

[5] "Area/mask processing methods", Trucco, Chapter 3.

[6] Tavish Srivastava. (2014). Introduction To K-Nearest Neighbors. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/. Last accessed 11th Nov 2017.

[7] David Cournapeau. (2007). Nearest Neighbors. Available: http://scikit-learn.org/stable/modules/neighbors.html. Last accessed 2nd Oct 2017.

[8] Math Works. Create models and forecast future outcomes. Available: https://www.mathworks.com/discovery/predictive-modeling.html. Last accessed August 15, 2018.

[9] Helen Mayfield, Carl Smith, Marcus Gallagher, Marc Hockings. (2017). Use of freely available datasets and machine learning methods in predicting deforestation. Environmental Modelling & Software. 87, 17-28.

[10] R.B. Arango, A.M. Campos, E.F. Combarro, E.R. Canas, I. Díaz. (2016). Mapping cultivable land from satellite imagery with clustering algorithms. International Journal of Applied Earth Observation and Geoinformation. 49, 99-106.

[11] Alexandrina-Elena Pandelea, Mihai Budescu, Gabriela Covatariu. (2015). Image processing using artificial neural networks. Universitatea Tehnică „Gheorghe Asachi" din Iași. 61 (65).

[12] Igor Klein, Ursula Gessner, Claudia Kuenzer. (2012). Regional land cover mapping and change detection in Central Asia using MODIS time-series. Applied Geography. 35 (-), 1-16.

[13] Marco Ottinger, Claudia Kuenzer, Gaohuan Liu, Shaoqiang Wang, Stefan Dech. (2013). Monitoring land cover dynamics in the Yellow River Delta from 1995 to 2010 based on Landsat 5 TM. Applied Geography. 44, 53-68.

[14] Charles L. Curry, Francis W. Zwiers. (2018). Examining controls on peak annual streamflow and floods in the Fraser River Basin of British Columbia. Hydrology and Earth System Sciences. 22, 2285-2309.

[15] M. Rogger, F. Gallart, M. Agnoletti, A. Alaoui, J.C. Bathurst, G. Bodner, M. Borga, V. Chaplot, G. Glatzel, J. Hall, J. Holden, L. Holko, R. Horn, A. Kiss. (2017). Land-use change impacts on floods at the catchment scale – Challenges and opportunities for future research. Water Resources Research. 53 (7), 5209-5219.

[16] Nathalie Schaller, Alison L. Kay, Rob Lamb, Neil R. Massey, Geert Jan van Oldenborgh, Friederike E. L. Otto, Sarah N. Sparrow, Robert Vautard, Pascal Yiou, Ian Ashpo. (2016). Human influence on climate in

the 2014 southern England winter floods and their impacts. Nature Climate Change. 6, 627-634.

[17] Top River Flood countries accessed at https://www.wri.org/blog/2015/03/world-s-15-countries-most-people-exposed-river-floods.

[18] Taghi Jokar Arsanjani, Reza Javidan, Mohamad Jafar Nazemosadat, Jamal Jokar Arsanjani, Eric Vaz. (2015). Spatiotemporal monitoring of Bakhtegan Lake's areal fluctuations and an exploration of its future status by applying a cellular automata model. Computers & Geosciences. 78 (-), 37-43.

[19] Ruili Lang, Guofan Shao, Bryan C.Pijanowski, Richard L.Farnsworth. (2008). Optimizing unsupervised classifications of remotely sensed imagery with a data-assisted labeling approach. Computers & Geosciences. 34 (12), 1877-1885.

[20] Arzu Erener, A. AbdullahSivas, A. Sevtap Selcuk-Kestel, H. Sebnem Düzgün. (2017). Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. Computers & Geosciences. 104 (-), 62-74.

[21] C.H.Jarvis, N.Stuart. (1996). The sensitivity of a neural network for classifying remotely sensed imagery. Computers & Geosciences. 22 (9), 959-967.

[22] Siya Chen, Tieli Sun, Fengqin Yang, Hongguang Sun, Yu Guan. (2018). An improved optimum-path forest clustering algorithm for remote sensing image segmentation. Computers & Geosciences. 112 (-), 38-46.

A Review of Feature Selection and Sentiment Analysis Technique in Issues of Propaganda

Siti Rohaidah Ahmad[1], Muhammad Zakwan Muhammad Rodzi[2], Nurlaila Syafira Shapiei[3]
Nurhafizah Moziyana Mohd Yusop[4], Suhaila Ismail[5]
Department of Science Computer, Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia, Sungai Besi, Kuala Lumpur, Malaysia

***Abstract*—Propaganda is a form of communication that is used in influencing communities, or people in general, to push forward an agenda for a certain goal. Nowadays, there are different means used in distributing propaganda including postings on social media, illustrations, cartoons and animations, articles, TV and radio shows. This paper is focused on election propaganda. Candidates in elections would use propaganda as a form of communication to channel and deliver messages through social media. Sentiment analysis (SA) is then used in identifying the positive and negative elements within the propaganda itself, through analysing the related documents, social media, articles or forums. This paper presents the various techniques used by previous researchers in issues of propaganda using SA, which include feature selection to remove irrelevant features and sentiment methods to identify sentiment in documents or others. Feature selection is a dominant side in sentiment analysis due to content of textual has a high measurement classification that can jeopardize SA classification interpretation. This paper also explores several SA techniques to identify sentiments in issues of propaganda. This study has also attempted to identify the use of swarm algorithms as a suitable feature selection method in SA for propaganda issues.**

Keywords*—Sentiment analysis; feature selection; swarm algorithm; propaganda

I. Introduction

Social media networks, such as Facebook, Twitter, and Instagram can be accessed online to facilitate a user's interaction with another user, such as sharing news, reading, discussing important events, having political discussions, and sharing thoughts. Smartphones have also helped to spread the use of social media to unlimited duration and location. This technology has facilitated in spreading information quickly and accurately to users. Propaganda would normally consist of accurate or semi accurate information, which is used to influence people to spread the various agendas to achieve a specific goal.

According to [1], propaganda is a special kind of social communication that involves three conditions, namely, the nature of its target society, communication experience, and the propagandist. In [2], propaganda refers to any communication used to assist or support the objectives of the government or the opposition in persuading, and influencing the emotion, behaviour, and actions of an individual, or any groups for the benefit of the propagandist, either directly or indirectly. Propaganda is used to influence the minds of millions of people as they go about their daily lives using various modes of communication (Lasswell, 1927). Although propaganda is often linked to the dissemination of negative messages [3] that are aimed at gaining or turning the support for the opposing party towards the propagating party, propaganda can also be used to spread positive messages [4], such as championing women's rights, racial equality, community health, encouraging people to vote, and so on. In conclusion, propaganda can be used to convey information that contains either positive or negative elements. Sentiment analysis (SA) is used to extract positive, negative, or neutral sentiments from text data. The results of this analysis can help various organisations in various fields, such as politics, business, productions, and marketing to make decisions. Thus, sentiment analysis can be utilised by certain parties to analyse and determine the status of a message, whether it contains positive or negative propaganda issues. This study aims to extensively analyse the effectiveness of SA for issues of propaganda, especially in political focus during elections.

The paper structure is as follows; Section 2 will display and expose the generic information about propaganda. Section 3 discusses about SA and how it can be linked to propaganda. In Section 4, feature selection and SA techniques are discussed in relation to issues of propaganda. Section 5 explains about a survey of swarm algorithms as a feature selection technique in SA. Finally, the last section will focus on the conclusions and possible directions for future research development plan.

II. Propaganda

Propaganda can be defined as an act of disseminating rumours, anecdotes, stories and gossips, whether it is based on a true or false allegation, with the sole purpose of persuading and influencing the public. The normal practice is that propaganda is used to influence people's mind in a misleading way by providing fabricated facts. In literature writing, propaganda is often referred to as a technique in controlling and manipulating the society's beliefs. History shows that propaganda is used in shaping the public's perception and discernment in order to achieve the influencer's political or economic goal [5]. During the First World War at around the early twentieth century, the word 'propaganda' began to undergo changes in its role and task, from religious indoctrination to political aspect or views, which reflect the shift in societal power from Church to State [6]. It was during this period that propaganda emerged as being political and partisan in nature, with its attempt to coerce or persuade a

mass audience to conform to a particular opinion or point of view. Based on the above definitions, propaganda can only be created through deliberate planning involving current sentiments from propagandists. Several questions have been raised regarding why propaganda exists. In [7], the objective of a propaganda may be to impress people to accept beliefs and attitudes pertaining to those of the propagandist or to bemuse certain patterns of behaviour, such as contribution of money, joining groups or blast-off demonstration for a cause. They also explained another objective of propaganda, which is to protect and maintain the authority of the institution or organisation it represents to ensure the legitimacy of its campaigns and events. Thus, it can be concluded that propaganda is created only to achieve the propagandist's dogma. Hence, propaganda activities are important to shape people's mind set in accepting the propagandist's ideology.

III. Sentiment Analysis

Sentiment analysis (SA) involves analysing texts that contain opinions and emotions. Any opinion or emotion expressed in the form of a text would contain a negative, positive, or neutral element. The information or opinion exchange and propaganda procreative often used in the internet [8]. According to [8], although different types of text classifications are available, SA was chosen for two main reasons. First, web pages are rich in emotion-related opinions and content. Second, text analysis relates to the use of propaganda on the web because text content contains opinions or directives that could influence public perception in decision making. [8] Also argued that SA is capable of detecting the presence of propaganda, and anger or hatred embodied in community channels. In addition, [8] concluded that SA can help produce an effective analysis and better understanding of the use of the web by extremist groups for disseminating information and propaganda. In the field of politics, [10] – [13] used SA to analyse sentiments contained in speeches of candidates for elections, issues, and prediction of election results. Propaganda is also used by the military, media, politicians, advertisers, governments, businesses, during election, marketing and by other mediums. This study will examine how previous studies utilised SA to analyse the propaganda contained in the speeches of candidates during elections.

IV. A Review of Feature Selection and Sentiment Analysis Techniques

The purpose of this study is to identify two main points in selected literatures discussed in this paper. The first step was to identify the feature selection or feature extraction techniques used to select important and useful features. The second step was recognising the methods used to identify the sentiments in the speech of an electoral candidate. According to [13], a feature is an aspect or attribute data [14] that users commented on. For example, a Panasonic electrical appliance, such as a radio is the product feature. Meanwhile, sentiment includes the words used to describe negative or positive sentiments, such as bad, good, beautiful, amazing, and so forth. In [13], feature can be categorized in two groups. First, in the form of unigram, bi-gram or tri-gram from a list of word in a document. Second using a part of speech tagging to identify each word in the sentence whether the word is a noun, adjective, adverb, verb, determiner or preposition.

Ramteke et al. [9] suggested two level or stage of framework in order to create a labelled training dataset. In stage one; they used hashtag clustering to get several tweets that consists similar hashtag. For example, the official hashtag for the American presidential candidate, Donald Trump, was #MakeAmericaGreatAgain. Therefore, anyone who tweeted using this official hashtag was considered to support Trump and labelled as a positive tweet. In stage two, they used Vader to analyse sentences as the input and to produce a percentage value for three categories, namely, positive, negative, and neutral and compound for the overall polarity of the sentence. In this experiment, they used Naïve Bayes and Support vector Machine (SVM) as the supervised learning algorithm in order to adjudicate the contrariety of tweets and predict election outcomes. They collected twitter data for two U.S. Presidential candidates, which were Donald Trump and Hillary Clinton, from 16 - 17 March 2016. They used TF-IDF (term frequency-inverse document frequency) to recognise terms that were applicable towards specific sentiments. Meanwhile, [10] did not use feature selection techniques or feature extraction to choose useful features. They used Syuzhet [15] methods to identify sentiments in the speeches presented by election candidates. A machine-based method and a structural topic-modelling [16] were used to analyse themes in candidate speeches. The results of the 50 main terms used in speeches were identified. In the final step, the word2Vector was used to sketch the candidates' words from specific thematic terms. The dataset collected from Clinton's speech included 89 scripts; with a total of words are 286,899. As for Trump's speech, 74 scripts were collected, with a total of 276,212 words. This dataset was collected from 1 April 2015 to 9 November 2016 for Clinton, whereas the dataset for Trump was collected starting from 16 June 2015 to 9 Nov 2016. Experimental results showed that a combination of machine-based methods of stm, syuzhet, and word2vec was very effective and quick to identify sentiments and themes in speeches. According to the authors, the combination of stm and word2vec had helped them understand the deeper perspective of the candidates' attitudes, views, and perspectives on the issues being reviewed. These researchers have also assumed that qualitative analysis is required to ensure that important information in the speech of a candidate is not lost.

Researchers [11] were able to classify texts related to the presidential and gubernatorial candidates for the Sao Paulo 2014 elections in Brazil. A dataset of 131 online news articles related to the 2014 elections was collected from *Veja*, *Estadao*, *Folha*, *GI*, and *Carta Capital*. They aimed to classify every paragraph in each article into either a positive, negative, or neutral group. Three learning algorithms, namely, Naïve Bayes, Support Vector Machine, and MaxEnt were used for sentiment classification. They conducted two experiments; first, they used the combination of unigram and bigram as a feature extraction technique to evaluate the best paragraph representative. Second, they also used feature selection method such as Chi-square, Categorical Proportion Difference, and Categorical Probability Proportion Difference

(CCPD) for selecting features. This second experiment was conducted to assess the effect of feature selection on learning algorithms. Experimental results showed that the MaxEnt learning algorithms, which use unigrams and bigram combinations, and was selected by CCPD that acts as feature selection technique was the most effective method for classifying text.

Researchers [12] used the Storm-based Real-time Analytic Service (SRTAS) designed to excerpt political-related tweet data from public tweets. This study had also analysed propaganda and sentiment rates using SRTAS and Kafka, namely, the Big Data System. Data tweets went through data processing, and each tweet was classified using the SVM module. Tokenizer was used to identify meaningful terms, and the stemming process was conducted to reduce the token that was relevant to the individual token. The process of classifying sentiments is often based on word subjectivity. Each word need to compare with predefined word-net libraries (AFINN-111) by researchers. However, their study did not specify the measurements used to measure the effectiveness of SRTAS. They also did not use feature selection techniques to choose relevant features from public data tweets.

Research in [17] conducted sentiment analysis on 42,235 twitter archive data in Hindi language using Dictionary-based, Naïve Bayes, and SVM algorithms. They analysed the sentiments of twitter users towards five political parties during India's general state election in 2016. Through supervised and unsupervised approaches, they built a classifier that could classify the collected data into positive, neutral, and negative groups. The limitation of their research was that they failed to consider the emoticons used in a message, which are relevant when defining the polarity of a tweet. Results of their analysis using SVM and Naive Bayes techniques showed more positive sentiments for the Bhartiya Janta Party (BJP), while the Dictionary-based approach showed more positive sentiments for the Indian National Congress. The SVM also read that BJP has a 78.4% chance to win the election based on the positive tweets. However, the final results showed BJP had only won 60 from 126 constituencies during the general election. The authors [18] also performed sentiment analysis on 343,645 twitter messages during the Austrian Presidential election in 2016. The dataset included tweets from the two presidential candidates, Van-Derbellen and Norbertghofer. The researchers used a combination of network science and sentiment analysis methods. They ran the SentiStrength algorithm during data preparation regarding on a lexicon of sentiment words, idioms, and emoticons in order to separate positive and negative data. Besides, they also occupied the NRC emotion-word lexicon through the tweets, and then stored all the identified emoticons in the tweets. Next, they conducted data analysis for the final version using sentiment analysis, network analysis, text mining techniques, and quantitative data analysis. Some limitations were observed when using SentiStrength and the NRC emotion-word lexicon. Thus, the limitations show some scores were inappropriate when assigned by these tools. In [19], the researchers used the lexicon and Naive Bayes Algorithm Learning Machine (NBALM) to totalize the political sentiments expressed in tweets for 100 days earlier from the election day. However, NBALM was only able to identify sentiments related to certain hashtags compared to lexicons of analysis. These researchers labelled the tweets manually and automatically. As a result, labelled tweets automatically had better precision than lexicon analysis. Therefore, the labelling of tweets automatically saved working hours, improved accuracy, and discarded any potential tendencies.

In 2017, the authors [20] used Twitter to determine sentiments in elections. The dataset was collected based on hashtags by two presidential candidates (Trumph and Hillary) starting from 24 April till 28 November 2016. Analysis of Variance (ANOVA) was accomplished using IBM SPSS to facilitate the validation of sentiment analysis results. This study was able to identify positive and negative sentiments that could be influenced by the debates, controversies, interviews, and other revelatory events. Nonetheless, they did encounter several limitations when contextualising the sentiments found in these tweets: limited number of characters for each tweet (only 140 characters are permissible); only a small number of tweets were used within the timeframe of their study; Twitter API was used to support composing or gaining the tweet (thus, difficulties in interpreting the exact meaning of a tweet); limited hashtag usage for both candidates; and no geographical mapping of the sentiments were provided.

V. A Survey on Feature Selection using Swarm Algorithm in Sentiment Analysis

Swarm Intelligence involves a simple assemblage of agents interacting locally and externally in their corresponding circumstances [21]. The individual agent is unintelligent, but the overall system of Swarm Intelligence commands the intelligent comportment with interactions between the agents and their environment. Swarm Intelligence algorithms are broadly used in optimisation problems where the calibre of result could be scaled. These algorithms will improve the quality of the solution by working in several iterations or repetitions and knowledge application of previous iterations on a selection of recent values. A comparison between various Swarm Intelligence techniques has shown that exactitude with much reduced feature sets can be achieved. In this comparison, three Swarm Intelligence techniques were used, namely, the Artificial Bee Colony (ABC), the Ant Colony Optimization (ACO), and the Particle Swarm Optimization (PSO). This comparison experiment used datasets from product reviews, internet movie database (IMDb), governmental decisions data, Twitter data, and restaurant reviews. The ABC and PSO algorithms, being puissant optimisation techniques, are widely used for working out hybrid optimisation challenges. These two methods have been used for optimising feature selection subsets, and for improving the accuracy of classification and clustering. The PSO algorithm, when amalgamated with sentiment classifiers (SVM and CRF), can enhance the classification accuracy by 4.25%. Meanwhile, the ABC algorithm often applies SVM, Naïve Bayes, FURIA, and RIDOR as classifiers, with accuracy increments of 9.04% as concluded in this research, Swarm Intelligence algorithms have generated better results based on sensitivity, specificity, and accuracy. It could also improve a classifier's performance in each SA.

Researchers in [22] reported that feature selection methods are necessary when selecting relevant feature vectors. One of the most consequential processing steps in classification is feature selection. This step is important for choosing useful features that would not affect the quality of the classification. They proposed the fitness proportionate selection-based binary particle swarm optimisation (F-BPSO) as a feature selection method in their study. To manufacture some changes in the experiment, these researchers used fitness sum rather than average of fitness in the fitness proportionate selection step. This modified approach is known as the fitness sum proportionate selection binary particle swarm optimisation (FS-BPSO). The FS-BPSO method did the final changes to make it more appropriate for the feature selection specialty, which is oriented towards sentiment classification. This further modified method is called the SCO-FS-BPSO, where SCO refers to "sentiment classification oriented". Experimental results showed that the SCO-FS-BPO can produce higher accuracy and obtain better quality features compared to the traditional Binary Particle Swarm Optimization (BPSO) and F-BPSO. This study used two benchmark datasets, namely, the Madelon and the Semeion Handwritten Digit (shortened to Semeion for simplicity) to test the effectiveness of the proposed method.

Humans around the world are spending multitudinous time using the internet, especially when surfing various social networks (Web 2.0), such as Facebook, Twitter, and Instagram. Hence, these social media which produce many opinions had influenced almost every single human in this world [23]. Thus, [23] have tried to understand this phenomena by using the revolutionary swarm intelligence algorithms, which might be the best and the most similar to the interactions between online users as members of a swarm. They used two datasets from Twitter and Reddit API. Twitter data format must include 'User_id', 'id_str', 'created_at', 'favourite_count', 'retweet_count', 'followers_count', and 'text'. Meanwhile, Reddit data consists of simple plain texts, with no limit to the length of posts and paragraph usage. In addition, Reddit only stored the data for all post posted by user which contains a string of text. Ten different subsets were divided from the dataset. Nine subsets were used for training and the final subset was used for testing. Two braces of heuristic arrays and pheromone continued to update their previous values and predictions during training. These arrays were called positive pheromones, negative pheromones, positive heuristics, and negative heuristics. The progression of the ant colony system for this experiment was trailed by these four arrays, which was also used to assist the ants knows how to structure the right conclusion. Other than that, all sentiments of posting were evaluated using natural language processing (NLP) techniques. The value for correct and incorrect predictions will increase if a post contains sentiments and the prediction matches of the sentiment. The results showed that the prediction accuracy for Twitter data was higher than for Reddit data. This research had also presented a graph of increasing positive heuristic and decreasing negative heuristic. Thus, it was concluded that one of the advantages of a swarm intelligence algorithm is its ability to perform faster and better than traditional algorithms for Sentiment Analysis, partly because it closely resembles human behaviour. Besides, it is a good algorithm in order to break down sentiments in social sites. Nonetheless, when sentiment changes rapidly and drastically like in group of chats, this algorithm does not perform well.

Feature selection is the main step in classification system that selects a subset from original features. In [24], four features selection were conducted in the comparative study for text categorization that include Information Gain (IG), CHI-square, Genetic Algorithm (GA) and PSO in order to optimize classification result. The performance of this propose method have been tested using Reuters-21578 dataset. Besides that, in order to show effectiveness of this experiment, they used a simple classifier, K-Nearest Neighbour (KNN) that effects the performance of categorization. After analysing the precision and recall, the PSO-based algorithm showed a better reading accuracy value than other three of feature selections. The percentage of selected features achieve until 12% in micro-F1 and macro-F1 measures when tested with PSO-based algorithm. Hence, PSO-based algorithm is the best compare to GA, IG and CHI with reading in Macro-F1 is 78.8564 and Micro F-1, 89.5684 [24]. PSO is quickest in locating optimal solution and able to get the determination within 10 iterations. Several advantages that can be identified from this propose method such as PSO has the capability to assemble quickly where it has strong searching skills, able to identify the problems in the environment and can effectively search the slightest feature subset within the sentence [24]. The drawback which can be listed is all the values from population size, maximum number iterations and range of weight are not optimal values. This disadvantage can be discussed in further research.

The growth of social media contributes to big numbers of user generated content including opinions, client reviews and comments. So, it is one of the opportunity to evolve an intelligent system that spontaneously cluster and classify them into positive and negative [25]. In this experiment, two classifiers were used, SVM and the hybridization of two classifiers, SVM and PSO (SVM-PSO). The multiplication of TF and IDF will calculate the score of every sentence in source document which word based on adjective word excerpted from Parts of Speech Tags using movie review dataset [25]. The approaches of sentiment analysis in [25] for sentiment analysis involve a two-level method, firstly, recognize the parts of document in order to provide the positive or negative sentiments. Secondly, connecting these parts of document in ways that boost up the odds of document falling into one of these two types of categories. This research has categorized data into four groups, True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). The result of SVM-PSO in precision, recall and accuracy test is better than SVM [25]. One of advantages can be concluded in this research is the hybridization of SVM-PSO increases the accuracy reading in sentiment analysis. Besides, this research help in making decision towards client reviews, comments and opinions and can save more consume of time in analysis [25]. Despite, the drawback in this research is the small number of classifiers which has only two. In order to yield the best result, more classifiers should be involved in experiment.

TABLE. I. A SUMMARY OF FEATURE SELECTION AND SENTIMENT ANALYSIS TECHNIQUES

Author	Feature Selection	Sentiment Analysis Techniques
[9]	Not mentioned in paper.	Used tools (VADER) to identify the percentages of positive, negative, and neutral categories. They also used Multinomial Naive Bayes and Support Vector machines to determine the polarity of tweets.
[10]	Not mentioned in paper.	Used syuzhet as sentiment lexicon to detect sentiments in every sentence in a document.
[11]	Chi-Square, Categorical Proportional Difference, and Categorical Probability Proportion Difference	Used Naïve Bayes, Support Vector Machine (SVM), and MaxEnt to classify each labelled paragraph of Corpus Vi´es into either positive, negative or neutral.
[12]	Mentioned, but feature selection was used in reference to sentiment classification and extraction.	Classified the words into positive, negative or neutral based on the subjectivity of the word.
[17]	Not mentioned in paper.	Used Dictionary-based, Naive Bayes, and SVM algorithms as classifiers in order to assist classification of test data as positive, negative, and neutral.
[18]	Not mentioned in paper.	Used SentiStrength tool and the NRC dictionary for extracting sentiment and emotion based on their polarities and vectors.
[19]	Not mentioned in paper.	Used Lexicon and Naïve Bayes algorithm to identify sentiments.
[20]	Not mentioned in paper.	Used context-based customized dictionaries.

Researchers in [26] have hybridised ant colony optimisation (ACO) and K-nearest neighbour (KNN) as a feature selection technique to produce an optimum feature set that can help to yield high classification and clustering accuracy. They used electrical product data from Nokia, Canon, Apex, Creative, and Nikon. This study used the grammatical structure method which consist the type of reliance and post tagger to identify sentiment word and its relationship with the features contained in the user's review. The combination of ACO-KNN as feature selection has yielded an optimised feature subset and improved classification accuracy.

VI. CONCLUSION

Based on Table I, most of the reviewed studies did not specify their feature selection method. As for sentiment analysis techniques, several studies have used tools, such as Vader, syuzhet, dictionary-based, and sentiment classification techniques, such as Naïve Bayes, and support vector machines. Based on this literature review, the use of swarm algorithm technique for feature selection is common in different domains, such as product, movie, Twitter, restaurant, and online communication. The results of several experiments in this review have shown that the swarm algorithm technique can produce relevant and high quality feature subsets, which could improve the accuracy of sentiment classification. The results of this review suggest that swarm algorithm technique should be used for feature selection in propaganda domains because this technique has the potential to produce a subset of quality and relevant features. This can help increase the accuracy of sentiment classification. Therefore, more detailed studies and experiments on the use of the swarm algorithm technique for feature selection in propaganda domain need to be conducted to prove that this technique is capable of producing a subset of quality features that are of relevance; thus, increasing the accuracy of the classification of sentiments.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Pertahanan Nasional Malaysia and the Skim Geran Penyelidikan Jangka Pendek Fasa 2/2018 for supporting this research project through grant no. UPNM/2018/GPJP/2/TK/14.

REFERENCES

[1] C. R. Wright, Mass communication: A sociological perspective. 1959.

[2] Ron D. (Ed). McLauri, Military propaganda: Psychological warfare and operations. New York: Praeger; First Edition edition, 1982.

[3] R. Di Tella, S. Galiani, and E. Schargrodsky, "Reality versus propaganda in the formation of beliefs about privatization," Journal of Public Economics, 2012.

[4] M. Ginsberg, "Propaganda Art as a Powerful Weapon for Promoting Nationalism, Patriotism and Hatred Towards the Enemy," in Inside the World's Major East Asian Collections: One Belt, One Road, and Beyond, 2017.

[5] LiteraryDevices Editors, "Propaganda," 2014. [Online]. Available: https://literarydevices.net/propaganda/.

[6] E. W. Fellows, "' Propaganda :' History of a Word Author (s): Erwin W . Fellows Published by : Duke University Press Stable URL : http://www.jstor.org/stable/454039 C HANGES IN THE MEANING attached to a word may be associated with," vol. 34, no. 3, pp. 182–189, 2017

[7] Jowet, "How To Analyze," pp. 269–288, 2009.

[8] A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web forums," ACM Transactions on Information Systems, vol. 26, no. 3, pp. 1–34, 2008.

[9] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using Twitter sentiment analysis," in Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016, 2017.

[10] D. Liu and L. Lei, "The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election," Discourse, Context and Media, 2018.

[11] Caio Magno Aguiar Carvalho, Hitoshi Nagano, and Allan Kardec Barros, "A Comparative Study for Sentiment Analysis on Election Brazilian News," in Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, 2017, pp. 103–111.

[12] M. K. Akhmedov Khumoyun, Y. Cui, and H. Lee, "Storm based Real-time Analytics Service on Propaganda and Sentiment Analysis of Political Tweets on Cloud Computing Environment," in International Conference on Engineering Technologies and Big Data Analytics, 2016, pp. 61–65.

[13] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection techniques in sentiment analysis," Intelligent Data Analysis, vol. 23, no. 1, pp. 159–189, 2019.

[14] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[15] M. Jockers, "syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text," The R-Project, 2016.

[16] M. E. Roberts et al., "Structural Topic Models for Open-Ended Survey Responses," American Journal of Political Science, 2014.

[17] P. Sharma and T. S. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," in Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, 2016.

[18] E. Kušen and M. Strembeck, "Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections," Online Social Networks and Media, 2018.

[19] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," in 2017 IEEE MIT Undergraduate Research Technology Conference, URTC 2017, 2018.

[20] S. Hari, D. Kolagani, and M. B. A. Student, "Identifying Trending Sentiments in the 2016 U . S . Presidentialelection: A Case Study Of Twitter Analytic," Issues in Information Systems, vol. 18, no. 2, pp. 80–86, 2017.

[21] A. Kumar, R. Khorwal, and S. Chaudhary, "A Survey on Sentiment Analysis using Swarm Intelligence," vol. 9, no. October, 2016.

[22] L. Shang, Z. Zhou, and X. Liu, "Particle swarm optimization-based feature selection in sentiment classification," Soft Computing, vol. 20, no. 10, pp. 3821–3834, 2016.

[23] 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.

[24] M. H. Aghdam and S. Heidari, "Feature Selection Using Particle Swarm Optimization in Text Categorization," Journal of Artificial Intelligence and Soft Computing Research, vol. 5, no. 4, pp. 231–238, 2015.

[25] J. P. Nirmala Devi K, "Sentiment Classification Using SVM And PSO," International Journal of Advanced Engineering Technology E- Int J Adv Engg Tech VII/Issue II, pp. 411–413, 2016.

[26] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "Ant colony optimization for text feature selection in sentiment analysis," Intelligent Data Analysis, vol. 23, no. 1, pp. 133–158, 2019.

DesCom: Routing Decision using Estimation Time in VDTN

Adnan Ali[1], Jinlong Li[2], Aqsa Tanveer[3], Maryam Batool[4], Nimra Choudhary[5]
School of Computer Science and Technology, University of Science and Technology China, Hefei, Anhui, China[1, 2]
Department of Computing and Information Technology, University of Gujrat, Sialkot, Pakistan[3, 4, 5]

***Abstract*—VDTN was proposed as a disrupting network which is established on the paradigm of the delay-tolerant network. VDTN uses vehicular nodes to convey messages as, it permits sparse opportunistic network connectivity, which is considered by the low node density where the vehicular traffic is sporadic, and no end-to-end paths exist between nodes. The message bundle is directed from the sender to the receiver node based on the routing protocol decision. While Routing protocols take decisions based on different metrics like Time to live, Location, Remaining Buffer Size, meeting probability, etc. In this paper, a routing protocol named *DesCom* is proposed for Vehicular Delay-Tolerant Network under a highly suppressed and sparse environment. DesCom takes the decision based on Message TTL, Transmission rate, and Estimation time. Estimation time is calculated in our previous work. The protocol defines whether to direct the message to the requested node or search the other more suitable node to carry that data bundles. After compiling multiple simulations with different numbers of vehicles and comparing DesCom with other routing protocols it is concluded that DesCom has the least buffer time with low latency along with good delivery probability.**

***Keywords*—Estimation time; VDTN; routing; vehicular delay-tolerant network; ONE simulator**

I. INTRODUCTION

The vehicular ad-hoc network (VANET) [1] is the enhancement of the Mobile ad-hoc network with some distinctive characteristics**.** VANET is known for high mobility which leads to discontinuous communication to the areas consisting of low traffic and congestion in the area of high traffic. These factors disturb the performance of the network. To overcome these difficulties, the vehicular network may carry messages using the store, carry and forward paradigms of DTN as shown in Fig. 1 and, its extension VDTN is introduced. The idea of Delay Tolerant Networks[2] is to dig out from Interplanetary Network (IPN) which was started in 1970 to establish a communication between two planets (mars and earth). DTN consists of a message-oriented layer called "Bundle Layer" that is deliberate for unstable communication.

Datagrams (IP packets) are accumulated in wavering length data packets, termed as Bundles. The source node generates a data bundle and stores its information until it encounters another vehicular node. On meeting with the node, the message is transferred to the receiving nodes and it keeps happening until its TTL (Time to Live) [3] is expired.

Globally, smart cities are immoderately increasing day by day [4] due to the advancement of technology. Vehicular communication is essential for the smooth working of cities. The vehicular delay-tolerant network plays a significant role in communication where the network is not much dense. In a real-time scenario, each vehicle travels with different speed with unpredictable movement. Speed of vehicles vary due to the following scenarios [5]:

1) The road at which vehicle travel may be indented.
2) A throng of traffic inside the city is different from outside the city.
3) Highway speed is different from rural roads.
4) The type of vehicle which carried out the material may be different.
5) Timing is another factor in daylight, the speed of the vehicle is different from the night.

This paper derives in the sequence of previous work[5][6] in which authors explain connection time and calculate estimated time respectively. *Estimation Time* is a predictable time and calculated based on vehicle speed and distance (covered by a specific vehicle). Instead, *Connection Time* is the time between connection establishment and termination. The simulation results indicate that DesCom is better than some compared routing protocol of VDTN as it has the lowest buffer time with low average latency having a high delivery rate as compared to epidemic, prophet and direct delivery. This shows that the routing decision taken based on estimation time may lead to better decision making.

The remainder of this paper is organized according to the following sequence: Section I Introduction, Section II Vehicular delay Tolerant Network, Section III Literature Review, Section IV Methodology, Section V Results and Discussion, Section VI Conclusion and Future Work, and Section VII Acknowledgment.

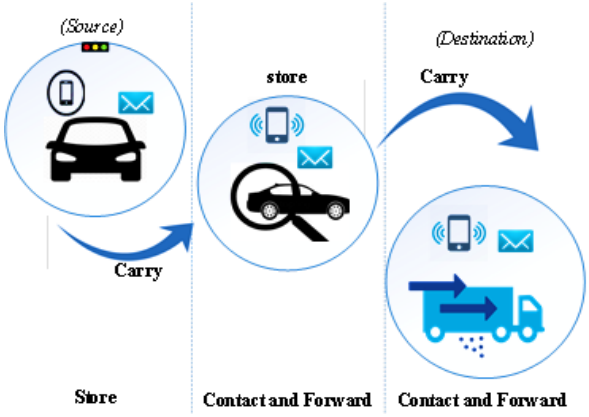


Fig. 1. DTN Paradigm.

II. Vehicular Delay-Tolerant Network

DTN addresses the technical challenges which exist during communication between different devices as it loses its continuous connectivity due to high mobility. This network continues to advance the data even when from the source node to the destination node there is no complete identifiable path. DTN consists of a unique feature of intermittent connectivity and opportunistic communication [7] in which each node communicates with each other in the predefined time. The architecture of DTN was extended to transit networks, called Vehicular DTN or simply a VDTN. VDTN supports the concept of VANET with the Delay-Tolerant Network concept in network connectivity to support long disruption. It extends the capabilities of the delay-tolerant network for the vehicular environment [8]. VDTN is the extension of DTN in which vehicles interconnect with each other and with stationary nodes in order to place the message to the destination nodes travel along the road [9]. Most of the problem in the vehicular network is due to mobility as the speed of the vehicle as each vehicle travel with variant speed. VDTN is composed of mobile nodes that are wirelessly interconnected to one another despite connectivity issues. VDTN is growing the challenging field of DTN containing mobile nodes called vehicles. The idea behind this theory is that the node store message called bundle in its buffer until they find the opportunity to handover it to its destination or relay node [10]. The layered architecture of VDTN was initiated in [11] as illustrated in Fig. 2.

The bundle layer is specified as a message-oriented overlay layer in delay tolerant network as well as in vehicular delay-tolerant network. In DTN bundle layer occurs beyond the transport layers. Application data units are converted into protocol data units at the bundle layer and then data units are called “bundles” which are sent by DTN nodes according to the Bundle Protocol. The Bundle layer comprises a node-to-node transference of message having the responsibility for reliable delivery. In VDTN instead of positioning beyond the transport layer, the Bundle Layer is positioned inferior to the network layer because of communication problems like high error rates, high latency, long delay, sparse and discontinuous connectivity, and even end-to-end connectivity doesn't exist. The core objective is to convey the vast message bundles instead of IP packets over the network. VDTN architecture categorizes the two logical planes:

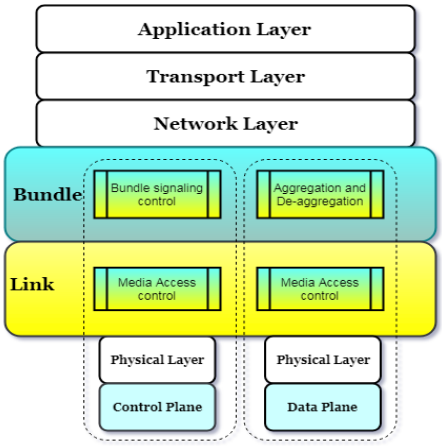


Fig. 2. VDTN Architecture.

1) Control plane.
2) Data plane.

The control plane uses short bandwidth, less-power, and vast-range link, as it continuously assists to let node discovery. It is responsible for a signaling message, node localization, resource reservation and routing of data bundles from the sender to the receiver node. The data plane utilizes excessive bandwidth with high-power and short-range link and its activation take place at estimated time duration, otherwise, the connection of the data plane is not activated. This approach is considered vital because it is significant for energy-constrained vehicular nodes in the network such as stationary nodes as well as guarantees the optimization of the available resources of the data plane. It is responsible for traffic classification, buffer management, scheduling, forwarding bundles among each other and data aggression and de-aggression.

Bundle layer in VDTN is logically divided into two sub-layers which are also illustrated in the above figure:

a) Bundle Signaling Control (BSC)
b) Bundle Aggregation and De-aggregation (BAD)

Bundle signaling control executes the functions of the control plane such as routing, signaling message, and message exchanging and resource reservation. Bundle aggregation and de-aggregation execute the functionality of the data plane which consists of queuing, storage management and scheduling and traffic classification. VDTN provides a promising environment for velar communication with no network infrastructure.

In VDTN, vehicles act as a moving node to carry the message from source to destination. Different protocols are used that perform the paradigms of the store, carry and forward [12]. The Vehicular Delay-Tolerant Network scenario is elaborated in Fig. 3. Applications of VDTN includes:

1) Notification of traffic jam
2) A weather condition report (rainy, windy, stormy)
3) Road accidents
4) Vehicles collision avoidance
5) Message loss.

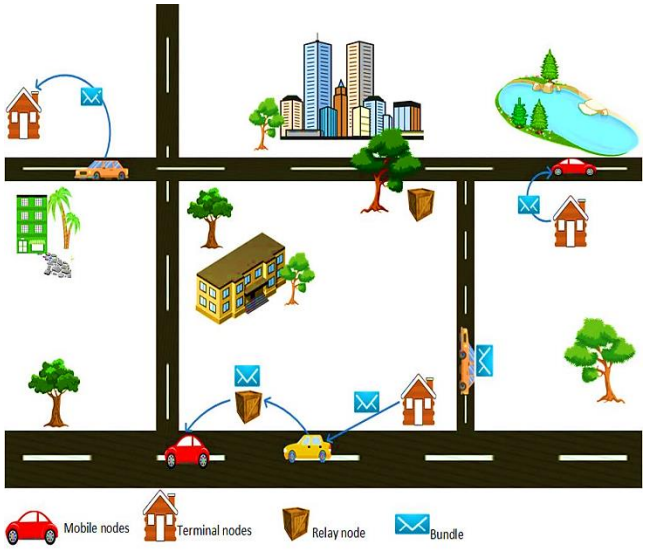


Fig. 3. Scenario1.

Mobile nodes are vigorously allowed to gather and broadcast message bundles through the VDTN. They move along roads and transfer the message bundle that must be conveyed to the terminal nodes. Terminal nodes can be characterized as fixed and mobile nodes. Relay nodes are stationary nodes with capabilities of DTN (store-and-forward) located at road crossings. Mobile nodes interrelate with relay nodes to placing and then picking up message bundles.

III. Literature Review

Many routing protocols are published in the literature that carries the data over the network. Here we explain some of the routing protocols and their working so it will be clear how our routing protocol is different and have distinguished features than other available in the literature. Table I elaborates the functionality and some other specifications of routing protocols. The routing protocols are categorized as follow:

MaxProp: (Maximum Priority) [13] is one of the well-known routing protocols of VDTN. It uses hop count in packets as a measure of network reliability on the way to a short distance. The MaxProp established on the base of store carry and forward paradigms which are usually employed in delay tolerant network. When two vehicles communicate with each other, they exchange the message in a definite order. If the node which is encountered, is the destination node having a message packet then the message is delivered.

Direct Delivery: In the Direct Delivery routing algorithm the message (bundle) is directly sent to the destination nodes [14]. The protocol does not require any information regarding the network so, it is flooding in nature. When the source node generates the message also called bundle then it waits for the destination node to receive it.

Epidemic: It is a flooding-based technique. In Epidemic routing [15] each node replicates its message to every other node until to meet it so far. The message replication is happening after the maintenance of the summary vector. The summary vector is the retained for each individual node that stores the message information which is delivered to the encountered node or stored in its buffer. This flooding base technique wastes the network resources for the sake of message delivery to one destination.

Spray And Wait: Spray And Wait [15] protocol is the advancement of an epidemic protocol. According to this protocol, nodes do not replicate the message to every other node but optimize the number of nodes on which the source node conveys the message. It contains two phases:

1) The Spray Phase
2) The Wait Phase.

In the Spray Phase, the source node performs message replication to the x nodes and then x nodes are further spread the message to y nodes. In the Wait Phase, if a node has a single copied message left then the sender node directly transmits it to the destination node.

ProPHET: This protocol uses an approximation of delivery to conclude the performance measures such as Delivery probability. It uses the matric called Delivery Probability to evaluate the probability of the nodes conveying the message to the destination nodes.

First Contact: A routing protocol in which only a single copy of the message is available over the network. It uses a single path to transfer the message and if the connection is not available then the node stores the data and then transfers it to the first contacted node. It does not delete the message but removes the old messages when there is no space in its buffer.

TABLE. I. Routing Protocols

Routing Protocol	*Abbreviation*	*Estimation Based*	*Copy*	*Type*	*Function*
Epidemic	Epidemic	No	Unlimited	Blind flooding	Rapid propagation of data
SprayAndWait	SnW	No	n_copy	Controlled flooding	Set the limit to the number of copies
ProPHET	Prophet	Yes	Unlimited	Probabilistic	Forward based on encounter history
DirectDelivery	DD	No	Single	Direct	Node moves and delivers the bundle
MaxProp	Maxprop	Yes	Unlimited	Maximum Probability	Delivery likelihood
FirstContact	FC	No	Single	Probabilistic	A packet is delivered in the result of random walk
DesCom	DC	Yes	Single	Routing Decision	Decide either packet is sent or not and ensure high delivery rate

Author in [16] describes the scheduling and dropping policy of packets that implement traffic planning across the network nodes. According to which, the message is scheduled using one of the following policies: Round Robin, Priority Greedy and Time threshold. According to the dropping policy, the message with the least priority and the lowest remaining TTL is discarded first. The research work [17] describes the influence of nodes in the situation of delivering the message according to which, nodes are categorized as "cooperative", "non-cooperative" and "partly cooperative". Author in [18] illustrates the willingness of cooperating nodes or the ability to participating in message forwarding. Two types of cooperating models are introduced: Upon the reception of message, the node either drop the copy of message packet or retain it in its buffer and follow the rules of the routing algorithm or upon the reception of node, store the copy of message packet in its buffer and forward it to the receiver node. A routing scheme was introduced called Road Accident Prevention (RAP), which introduced the Early Warning to make the promising decision, i.e. chose the alternate route, slow down the speed of the vehicle of change the lane. The paper [19]elaborate buffer management strategy which not only increases delivery rate but also decreases high priority by using the ProPHET routing protocol [20]. According to which, among multiple nearest nodes, the node with maximum delivery probability and situated near to destination is defined as the closest node to the destination due to which delivery rate may be increased. This work [21] describes a message dropping policy which is established on the messages priority and message TTL. According to this policy, the message having the lowest-priority and minimum message TTL are deleted first [22]. Some protocols are also available which specifies the Global Positioning System GPS [23]which are intended to use in the situation where communication is based on intermittent connection, continuous link disconnection or reconnection. Author in [24] provides a congestion control solution to optimize the message delivery probability and reduce the chance of message loss in VDTN and improve the awareness for the drivers of the network on the road. This work [25][26] defines the misbehaving of nodes in terms of the number of copies, such as single, n and an unlimited number of copies on the set of VDTN and defer the forwarding matric due to which delivery rate is reduced. Author in [6] introduces a parameter named connection_time which can be calculated when the connection between nodes gets started and terminated.

IV. Methodology

Vehicles travel in urban and rural areas may vary from one another due to their speed such as car travel, having different speeds as compared to heavy road vehicles. Vehicles speed may also vary due to traffic density and everyday timings such as inside the city, traffic density may be higher than outside the city, at office time car traveling speed may be different than night time. A study [27] shows, if two cars communicate with each other via IEEE 802.11g passing at 20Km per hour is around 20s and at 40Km per hour is approximately 15s and at 60Km per hour is almost 11s. By keeping this in consideration, if TCP is used at 60Km per hour then the goodput is low, around 4 out of 10 experimentations, no data is relocated to the destination. UDP provides a better result as compared to TCP because it is a connectionless protocol.

According to our previous work[5], estimation time is predicted time which can be measured by the speed of the vehicle and distance covered by the vehicle. Vehicle speed and position can be visualized using ONE Simulator [28]. Routing decisions are conducted using estimation time [29]. The source node generates a message bundle and searches for a node to carry that message bundle. In this work, when the vehicle carries the message bundle for the sake of message transference and establishes a connection with any other node. Both nodes exchange some information with each other like each other speed, remaining buffer size, and distance from each other. Then the estimated time is calculated from this information.

There is some other information. Like, every message has Time to Live (TTL), after this time message will be terminated and will no longer be available. Transmission time is time to deliver the message from one node to another node. This time depends on the link condition and data rate. We assume the link is good with this (625Kbps) data rate. Now we have 3 times. First is Estimation Time (ET), second is TTL and third is Transmission Time (Tt).

If Estimation time is greater than TTL it means message may expire soon and sending it may not be useful as a greater TTL message. If Transmission Time is greater than Estimation Time again then there is no point to send the message. Let's suppose Message size is 6250Kb, in the ideal scenario it will take 10 seconds to send the message, but the estimated time is 9 seconds. So preferred is that a smaller message should be chosen. There are some other conditions and scenario which affect communication. The vehicle carries the message bundle and moving on the road than there could be the following possible scenarios for the vehicle traveling on the road:

1) ***Scenario 1:*** One vehicle is stationary, and another vehicle is moving. In this case, the moving vehicle consists of a value greater than zero as it contains some positive value and the stationary vehicle has a value equal to zero. Because of this estimation time will be greater comparatively.

2) ***Scenario 2:*** Both vehicles are at stationary conditions having a speed equal to zero. Multiple simulations are established but, this kind of scenario is not available in which both vehicles are stationary having communication, this situation can be possible but having very low probability.

3) ***Scenario 3:*** Mutually moving vehicles crossing each other. When two interconnecting vehicles are moving in some directions and communicate with each other than there are further sub scenarios which are as follow:

a) When both vehicles are traveling in a similar direction.

b) When two vehicles are crossing each other but in the opposite direction.

c) When two vehicles are crossing over each other at the juncture.

Scenario a: If mutually vehicles are moving in the same direction then it is harder to calculate estimation time because

if Vehicle 1 is moving with speed 100km/h and Vehicle 2 is moving with speed 45km/h then there are two possibilities:

- We never know when Vehicle 1 overtakes Vehicle 2 due to variations in their speeds.
- Which vehicle takes a turn and when.

Scenario b and c: According to both scenarios, Vehicle 1 and Vehicle 2 are moving toward each other having a speed value greater than zero. In this case, the meeting time of both vehicles is limited as compared to Scenario A but, can be estimated.

V. Result and Discussions

A. Simulation Setup

For visualization of the simulation environment, a java based tool is used named, ONE[28] (Opportunistic Network Environment). ONE simulator consists of a configuration file termed as "default _setting.txt" in which matrices can be altered according to the need. For the simulation environment, we use a small part of the Helsinki city which is a map-based model having the shortest path available as shown in Fig. 4 in which vehicles move to the destination. Table II shows the values used for simulation.

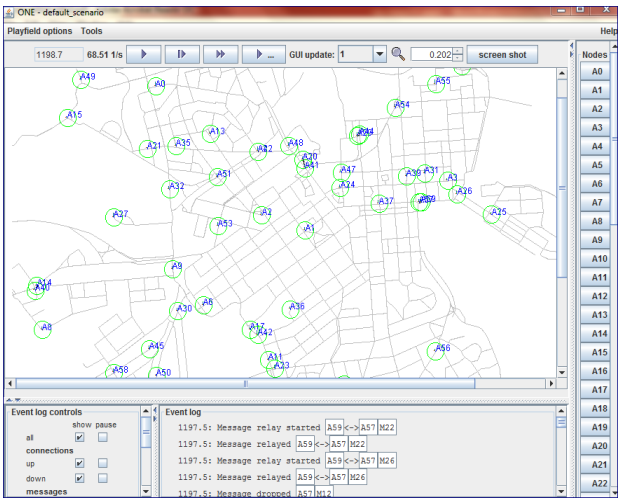


Fig. 4. Simulation Environment.

TABLE. II. Simulation Setup

Sr.No	Matrices	Values
1	sim_time	21600
2	No. of host	20, 40, 60
3	Transmit speed	625kbps
4	Transmit range	60 meters
5	Buffer_size	10M
6	Vehicle speed	10-50 km/h
7	Message TTL	30
8	Message size	100kB, 1MB
9	World size	4500, 3400
10	Wait time	0, 120

B. Simulation Results

As described earlier, DesCom can conduct the decision, whether the message packet should be sent to the encountered node or find the other suitable node to carry that message packet. Parameters that are excluded are as follows:

1) Delivery probability: It is the relation between the numbers of successfully conveyed messages along with the total number of messages created by the source node.

$$D_S = Successfully\ delivered\ messages$$

$$Total_M = Total\ number\ of\ messages$$

$$Delivery\ Probability = \frac{D_S}{Total_M}$$

2) Overhead ratio: Overhead ratio is one of the most vital metrics which shows how competent a protocol is, in terms of precise relay decisions. It indicates the total number of message copies that are relayed in the network before reach to its actual destination.

$$Relay_M = Transmitted\ messages\ that\ are\ not\ delivered$$

$$Total_M = Total\ number\ of\ messages$$

$$Overhead\ ratio = \frac{Relay_M}{Total_M}$$

3) Average latency: It is the average time reserved by the messages from the formation to their first delivery at the actual destination node. It is the average time between message generation and received towards the destination node

$$D_T = Time\ when\ message\ is\ delivered$$

$$C_T = Time\ when\ message\ is\ created$$

$$Total_D = Total\ number\ of\ delivered\ messages$$

$$Average\ Latency = \frac{D_T - C_T}{Total_D}$$

4) Buffer time: It is the average time of buffer taken by a vehicle to carry the message packet in its storage. It defines how long a message will remain in the vehicle's buffer. The decrease in average buffer time means that more messages are carried by the encountered vehicles.

TABLE. III. Parameter Checking with different Number of Hosts

Sr.No	Vehicle speed	Parameter	No. of Host		
			20	40	60
1	10-50	***Delivery Probability***	0.8425	0.8712	0.8712
2		***Average Buffer-time***	297.8803	151.0947	104.1008
3		***Average Latency***	640.0966	453.166	415.3612
4		***Overhead ratio***	35.5187	148.4858	330.2909

So, if the number of nodes is changed then the delivery rate still gave goodput along with low latency and buffer time respectively. Fig. 5, Fig. 6 and Fig. 7 demonstrate the comparison of DesCom with Epidemic, MaxProp, Direct Delivery, First Contact and Prophet Router having parameters that are defined in Table III. In the y-axis, there are values that show the performance of the protocol. In the x-axis, there are predefined parameters in the order of average latency, buffer time average, the number of transported messages and overhead ratio.

From Fig. 5 it can be clearly seen DesCom has the lowest latency and buffer time among all other protocols. While Delivered messages are only less than MaxProp. Here the overhead ratio is also greater than others.

Fig. 6 and Fig. 7 tends to have sort of same trend which is little different than Fig. 5, Here in all of the aspects except overhead ratio, Descom is better than all of the protocols except MaxProp.

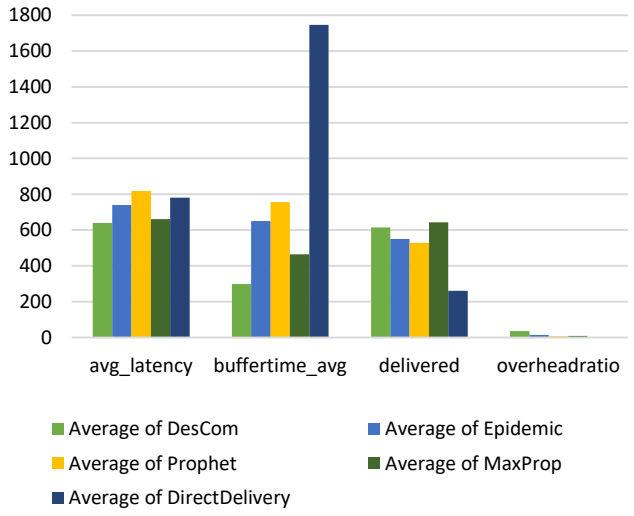


Fig. 5. Comparison with 20 Nodes.

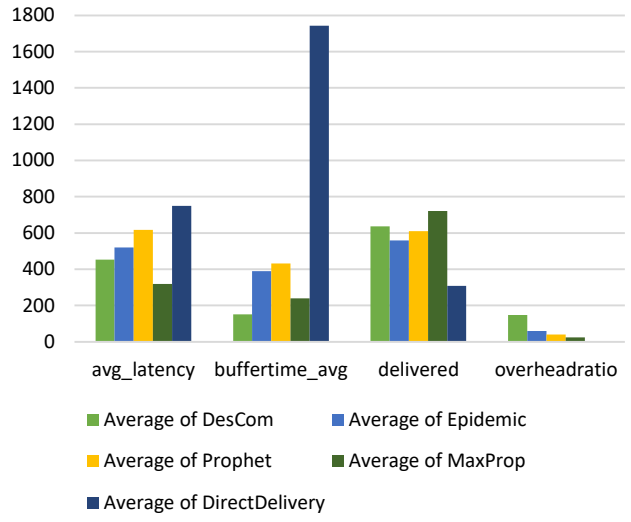


Fig. 6. Comparison with 40 Nodes.

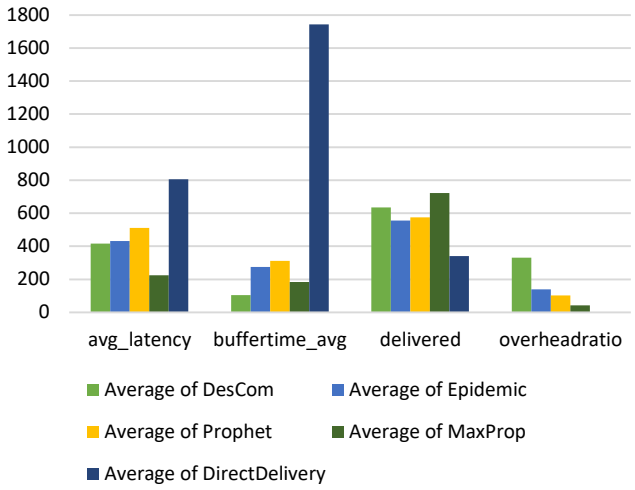


Fig. 7. Comparison with 60 Nodes.

VI. Conclusion and Future Work

In this paper based on our previous work we made a routing protocol to prove that estimation time can be used as metric for future routing protocols of such environment where we can calculate the estimation of a connection. In DesCom, routing decisions are conducted using metrics such as TTL, Estimation Time and Transmission Time. Comparisons with other protocols and results show that the routing decisions, taken at the time of message delivery are better which ensures the high delivery rate along with a reduction in buffer time and latency with the increasing number of vehicles. Many simulations were run in ONE simulator with the different number of vehicles to attest that "DesCom" conducts accurate decisions with a variant number of nodes. As Future work, we will suggest and expect more protocols based on estimation time. This estimation time could be a big help to reduce buffer issues. It may also help to reduce average latency and the delivery rate can be increased. We will further work to improve this protocol.

Acknowledgment

This Research is supported by the National Key Research and Development Program of China (Grant No. 2017YFC0804001) and the National Natural Science Foundation of China (Grant No. 61573328), sponsored by CAS- TWAS President's Fellowship for International Ph.D. Students at School of Computer Science and Technology, University of Science and Technology China.

References

[1] A. Cavalcanti, Elmano & Rodrigues, Anderson & Spohn, Marco & Cézar, Reinaldo & Fabiano, "VANETs' research over the past decade: Overview, credibility, and trends.," ACM SIGCOMM Comput. Commun. Rev., 2018.

[2] H. W. V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall, "Delay-Tolerant Networking Architecture," 2007.

[3] B. Zyl, Ignus & Rudman, L & Irwin, "A review of current DNS TTL practices," 2015.

[4] 2018. page 48 - 54. 10.31058/j.ud.2018.12005. Akinyemi, Lateef & Makanjuola, Tunji & Shoewu, Oluwagbemiga & Folorunso, Oluwaseyi. (2018). Smart City and Vehicle Pollution Monitoring Using Wireless Network System. VOLUME 1, "Smart City and Vehicle Pollution

Monitoring Using Wireless Network System," vol. Vol.1, p. page 48-54, 2018.

[5] A. Ali, M. Shakil, H. Rafique, and S. M. Cheema, "Connection Time Estimation between Nodes in VDTN," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 339–345, 2019.

[6] A. Ali, N. Sarwar, H. Rafique, I. Hussain, and F. N. Khan, "Connection Time for Routing Decisions in Vehicular Delay Tolerant Network," in Intelligent Technologies and Applications, 2019, pp. 679–690.

[7] K. Keranen, A., OTT, J., "The ONEsimulator for DTN protocol evaluation. In Simutools," in Proceedings of the 2nd International Conference on Simulation Tools and Techniques ,ICST (Institute for Computer Sciences, Social- Informatics and Telecommunications Engineering), 2009, pp. 1–10.

[8] J. 2009 V. N. G. J. Soares, F. Farahmand, and J. J. P. C. Rodrigues, "Improving Vehicular Delay-Tolerant Network Performance with Relay Nodes," 2009 Next Generation Internet Networks, pp. 1–5, "Improving Vehicular Delay-Tolerant Network Performance with Relay Nodes," in Next Generation Internet Networks, 2009, pp. 1–5.

[9] P. R. Pereira, "From Delay-Tolerant Networks to Vehicular Delay-Tolerant Networks," IEEE Commun. Surv. Tutorials, vol. Vol. 14, no. No. 4, p. Page No.-1166-1182, 2012.

[10] F. F. V. N. G. J. Soares, J. J. P. C. Rodrigues, "Performance assessment of a geographic routing protocol for vehicular delay-tolerant networks," IEEE Wirel. Commun. Netw. Conf., pp. 2526–2531, 2012.

[11] S. Rana, "VDTN Layered Architecture," 2014.

[12] J.-M. Benamar, Maria & AHNANA, SARA & ZAHRAE SAIYARI, FATIMA & Benamar, Nabil & El Ouadghiri, Driss & Bonnin, "Study of VDTN routing protocols performances in sparse and dense traffic in the presence of relay nodes," J. Mob. Multimed., 2014.

[13] S. F. A. Shah, M. H. Zafar, I. Andonovic, and T. Jan, "Hybrid routing scheme for Vehicular Delay Tolerant Networks," in 2016 8th Computer Science and Electronic Engineering (CEEC), 2016, pp. 158–163.

[14] R. S. P. Puneet Dwivedi, "Study of Routing Protocols for Delay Tolerant Networks," in International Conference on Contemporary Technological Solutions towards fulfillment of Social Needs, 2018, p. Page – 144.

[15] J. G. F. Antonio S. S. Vieira et al., "VDTN-ToD: Routing Protocol VANET/DTN Based on Trend of Delivery," in The Ninth Advanced International Conference on Telecommunications, 2013, p. pg 135-141.

[16] V. N. G. J. Soares, F. Farahmand, and J. J. P. C. Rodrigues, "Scheduling and drop policies for traffic differentiation on vehicular delay-tolerant networks," SoftCOM 2009 - 17th Int. Conf. Software, Telecommun. Comput. Networks, pp. 353–357, 2009.

[17] D. Benamar, Maria & Benamar, Nabil & El Ouadghiri, "The effect of cooperation of nodes on VDTN routing protocols," international Conference on Wireless Networks and Mobile Communications (WINCOM), Marrakech, 2015, pp. 1-7.

[18] X. L. and P. Hui, "An Energy-Efficient n-Epidemic Routing Protocol for Delay Tolerant Networks," in IEEE Fifth International Conference on Networking, Architecture, and Storage, Macau, 2010, pp. 341-347.

[19] K. Yao Théophile, GBALLOU & Bi, GOORE & Brou, "Dynamic Weight Dropping Policy for Improve High-Priority Message Delivery Delay in Vehicular Delay-Tolerant Network," Int. J. Adv. Comput. Sci. Appl., vol. Vol.9, no. Issue.10, 2018.

[20] Seung Deok Han and Yun Won Chung, "An Improved PRoPHET Routing Protocol in Delay Tolerant Network," Sci. World J., vol. Volume 201, p. 7 pages, 2015.

[21] A. PENURKAR, Milind et GAJIPARA, "Scheduling and Dropping Policies in Vehicular Delay Tolerant Network," Int. J. Eng. Sci., 2013.

[22] H. Rafique, F. Anwer, A. Shamim, B. Minaei-Bidgoli, M. A. Qureshi, and S. Shamshirband, "Factors Affecting Acceptance of Mobile Library Applications: Structural Equation Model," Libri, vol. 68, no. 2, pp. 99–112, 2018.

[23] V. N. G. J. Soares and J. J. P. C. Rodrigues and F. Farahmand, "GeoSpray: A geographic routing protocol for vehicular delay-tolerant networks," Inf. Fusion, vol. 15, pp. 102–113, 2014.

[24] C. O. and M. Radenkovic and U. K. School of computer science, University of Nottingham, "Congestion Aware Spray and Wait Protocol: A Congestion Control Mechanism for the Vehicular Delay Tolerant Network," Int. J. Comput. Sci. Inf. Technol., vol. Vol, no. No 6, p. pg 83-95, 2015.

[25] P. R. P. and M. P. C. N. Magaia, "Nodes' Misbehavior in Vehicular Delay-Tolerant Networks," in Conference on Future Internet Communications (CFIC), Coimbra, 2013, 2013, pp. 1-9.

[26] A. Ali, H. Rafique, T. Arshad, M. A. Alqarni, S. H. Chauhdary, and A. K. Bashir, "A Fractal-Based Authentication Technique Using Sierpinski Triangles in Smart Devices," Sensors, vol. 19, no. 3, p. 678, 2019.

[27] M. G. R. et Al, "Measuring the capacity of in-car to in-car vehicular networks," IEEE Commun. Mag., vol. 47, no. 11, pp. 128-136, 2009.

[28] P. K. L. B. Er. Richa Thakur, "Delay Tolerant Networks: An Analysis of Routing Protocols with ONE Simulator," in I. J. Computer Network and Information Security, 2016, pp. 51–58.

[29] H. Rafique, A. Shamim, and F. Anwar, "Investigating acceptance of mobile library application with extended technology acceptance model (TAM)," Comput. Educ., p. 103732, 2019.

LCAHASH-1.1: A New Design of the LCAHASH System for IoT

Anas Sadak[1], Bouchra Echandouri[2], Fatima Ezzahra Ziani[3], Charifa Hanin[4], Fouzia Omary[5]
Faculty of Science, Mohammed V University
Rabat, Morocco

***Abstract*—The present paper represents an extension of LCAHASH system. LCAHASH is a previously developed lightweight hash algorithm. It is based on cellular automata. It was developed as an alternative to existing hash functions to ensure data integrity and to meet the security requirements of the Internet of Things devices. Due to the limited amount of storage and the limited computation capabilities of these devices, the algorithms used by these devices should be as efficient and as robust as possible. In this contribution, we propose an enhanced version of the original LCAHASH algorithm to improve its efficiency and its robustness. A description of the system proposed along with a security analysis and the results of a statistical battery of tests (Dieharder) are included. These results show that the system proposed exhibits good statistical and cryptographic features.**

Keywords*—Information security; hash function; cellular automata; IoT; data integrity

I. Introduction

In the past few years, our societies have witnessed a rapid growth in technology due to the development of the Internet and the technological advances in electronics and software. Soon, every device we possess will be connected to the Internet. These devices range from toothbrushes to homes and include cameras, smartphones and other appliances.

These advances help to improve our daily lives. However, most of these connected devices suffer from a lack of security. This can be due to either a negligence of the manufacturer or the lack of adapted security standards for these constrained devices. Examples of these security breaches are man-in-the-middle attacks, denial of service (DoS), data forging or physical attacks. This explains the increasing need to develop lightweight cryptographic primitives adapted to the computing power and the storage capacity of these devices.

In this context, LCAHASH, a lightweight hash system based on cellular automata, was previously developed to ensure the integrity of RFID tags data. In this paper, we modified the original LCAHASH design to make it more robust, more efficient and more suited for the Internet of Things (IoT) devices [1].

This paper is organized as follows: In Section II, we present cellular automata. In Section III, we mention some related works. In Section IV, we describe the new design for LCAHASH. Afterwards, we present the experimental results in Section V. Finally, we conclude with the conclusion in Section VI.

II. Background on Cellular Automata

Cellular automata (CA) are dynamic systems made of cells that take a state from a defined set of states [2]. In the case of $S=\{0,1\}$, the CA is called a Boolean CA. Cellular automata evolve according to a local transition rule [2]. The local transition rules can be represented by a truth a table or a logical function defining the relation between the present cell's state and its next state in relation with its neighbor(s). The whole system evolves according to a global function. Depending on the number of neighbors, the boundary conditions, the rules applied and the ruleset, a cellular automaton can be one-dimensional or more, linear or chaotic, uniform or hybrid, etc.

Cellular automata are simple structures that yield a complex and unpredictable behavior. This feature makes them easy to implement both in hardware and software and makes them good candidates to use in the context of lightweight cryptography [2].

III. Related Works

We present here some other existing lightweight hash functions and some existing cellular automata based hash functions.

With the exception of DM-PRESENT [3], which is based on the Davies-Meyer construction, most of the existing lightweight hash function are based on a sponge construction [4] rather than a Merkle-Damgård construction. Using the sponge construction, a fixed length output or digest is obtained from an arbitrary length input. This is made possible by the use of fixed length permutation function. Examples of these lightweight hash functions are KECCAK [5] (winner of the SHA-3 competition), PHOTON [6], SPONGENT [7], QUARK [8], GLUON [9] and Hash-One [10].

In [11], Damgård proposed three methods to design collision resistant hash functions. Among these methods, one uses cellular automata as a building block. This method was later on attacked in [12]. Daemen et. al also proposed two hash functions based on cellular automata in [12] and [13]. Those constructions were shown to be vulnerable by Chang in [14]. A family of hash functions was proposed by Mihaljevic in [15]. However, in this proposition, no rules and no neighborhood configuration were specified. Newer constructions comprise [16], [17] and [18]. 2D cellular automata were used by Hirose and Yoshida in [19].

IV. Description of LCAHASH 1.1

In this section, we describe the new design of LCAHASH.

The LCAHASH 1.1 process starts with splitting the input M into blocks of 128 or 256 bits, depending on the version of the algorithm. Padding is applied to the last block M_n if necessary. In the next step, a block is chosen at random (M_{index}) and XORed with IV_1, an initial vector of size 128 or 256 bits generated randomly. Following this step, a 7 or 13 bits prime number N is generated at random and R_is are computed following:

$$R_i = M_i \bmod N \quad (1)$$

The R_is are then concatenated into M'. M' is in turn split into blocks of 128 or 256 bits and padding is applied to the last block M_k' if necessary. M_{evol} is then obtained as follows:

$$M_{evol} = IV_2 \oplus M_1' \oplus M_2' \ldots \oplus M_k' \quad (2)$$

where IV_2 an initial vector of size 128 or 256 bits generated randomly. Finally, using the hybrid cellular automaton with the rule set {30,90}, M_{evol} is evolved for 128 or 256 iterations depending on the version of the algorithm. The obtained sequence is the output or the digest of our hash function system.

Algorithm 1. LCAHASH 1.1 Algorithm

Input: M, IV_1, IV_2, *index*, N
Output: Digest

```
1   Split M into n blocks of 128 or 256 bits;
2   If M_n is not a multiple of 128 or 256 then
3           Pad M_n;
4   End If
5   M_index ← M_index ⊕ IV_1
6   For i=1 to n do
7           R_i ← M_i mod N
8   End For
9   M' ← R_1 || R_2 ||...|| R_n
10  Split M' into k blocks of 128 or 256 bits;
11  If M_k' is not a multiple of 128 or 256 then
12          Pad M_k';
13  End If
14  M_evol ← IV_2 ⊕ M_1' ⊕ M_2' ⊕ ... ⊕ M_k'
15  Digest ← Evol_{30,90}( M_evol , 128 or 256 iterations)
```

Fig. 1 shows the different steps of LCAHASH 1.1.

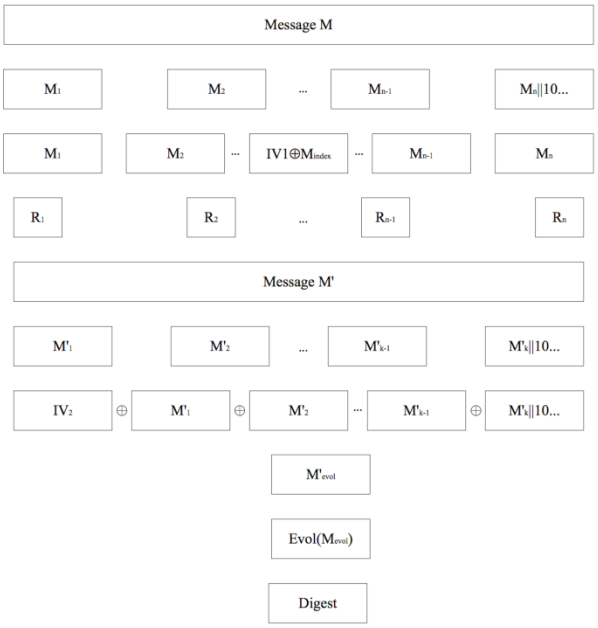


Fig. 1. LCAHASH 1.1 steps

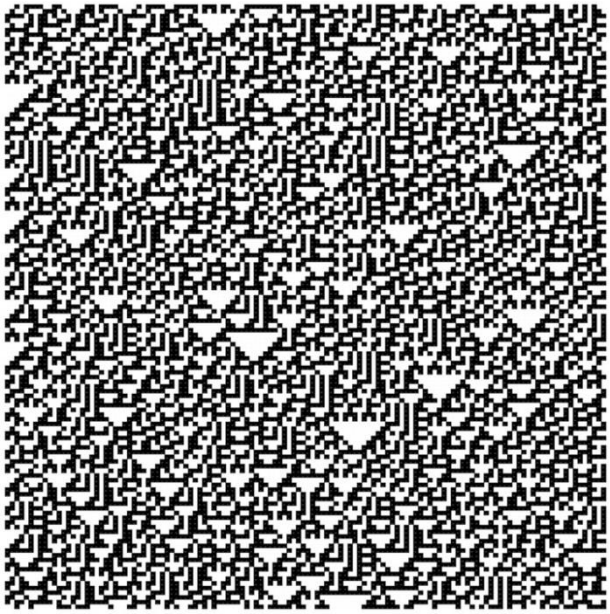


Fig. 2. Sample Evolution of M_{evol}.

Fig. 2 shows a sample evolution of M_{evol} using the non-uniform cellular automaton with ruleset {30,90}.

V. Security Analysis, Statistical Tests and Performance

Before proceeding to the security analysis of our modified design of LCAHASH, we consider first the major properties expected from a cryptographic hash function. These properties represent theoretical measures and are used in the field of cryptography to measure the level of security of hash functions and their ability to resist the most known cryptanalytic attacks.

Ideally, a cryptographic hash function is [20]:

- Deterministic: the same output (*h*) is always produced from the same input (*m*).
- Easy and fast to compute.
- Pre-image resistant: given an output *h*, it is computationally hard to find any input *m* such that $h = f(m)$. This property prevents an attacker who has a hash value *h* from finding the input *m*.
- Second pre-image resistant: given an input *m*, it is computationally hard to find another input *m'* such that $f(m) = f(m')$. This property prevents an attacker who has an input (*m*) and its corresponding output (*h*) from replacing the original input with another input (*m'*).
- Collision resistant: it is computationally hard to find two inputs *m* and *m'* that generate the same output such that $h = h'$. This property prevents an attacker to find two inputs with the same output. Note that a hash function which is collision resistant is also second pre-image resistant.

A. Security Analysis

1) Complexity: In general, a hash function should be easy and fast to compute. In order to evaluate this property, we try to approximate the complexity of our algorithm.

Splitting the message M into n blocks and applying padding requires at most $s - \frac{L}{s}$ steps (where *s* is the block size (128 or 256 bits) and *L* is the message length). XORing IV_1 and M_{index} requires *s* modulo 2 additions. Calculating the R_is requires $n \times s$ modulo 2 divisions. Splitting the message M' into *k* blocks and applying padding requires at most $s - \frac{L'}{s}$ steps (where *L'* is the length of the message *M'*). XORing M'_is with IV_2 requires *k* modulo 2 additions. Finally, evolving M_{evol} requires s^2 steps.

Overall, the complexity of our algorithm is $O(s^2)$.

2) Pre-image and second pre-image resistance: The security of the proposed system lies in the global transition rule of the cellular automaton as IV_1, IV_2, *index* and *N* are known parameters. Depending on the version of the algorithm, 128×2^{128} or 256×2^{256} operations are needed in order to find the global transition rule. Therefore, our proposed system is pre-image and second pre-image resistant.

3) Collision resistance: For each of the cellular automaton evolutions, a new state is obtained with $2^s - 1$ possible sequences. If we consider the birthday attack [21] , the complexity upper bound for breaking the collision resistance of our system is $O(2^{s/2})$. It means that depending on the version used, $2^{128/2}$ or $2^{256/2}$ operations are required to find a collision.

4) Avalanche effect: In cryptography, a function that displays the avalanche effect property is a function for which a small change in the input causes a greater change in the output [22]. Preferably, one bit changed in the input should change half of the bits in the output (strict avalanche criterion) [23].

To evaluate the avalanche effect property of our system, we took a sample of a hundred 1024 bits messages. For each message, we changed the original message bit per bit and calculated the Hamming distance between the hash value of the original sample message and the modified messages. We then calculated the average percentage values based on the hundred sample messages. We evaluated the avalanche effect using the 128-bit version of our system. The results of our setting are shown in Fig. 3 and Table I.

B. Statistical Tests

In order to test LCAHASH against statistical attacks, we used the DIEHARD test suite [24]. This test suite consists of a series of tests that evaluate the randomness of an algorithm. The benefit of this test suite is to show that the output of our algorithm is statistically indistinguishable from the output of a true random source and that predicting the output is computationally hard.

Table II shows the results of the DIEHARD test suite.

From the table below, we can see that LCAHASH 1.1 128-bit version passed all the tests of the DIEHARD test suite. LCAHASH 1.1 has then a good random behavior and its outputs are statistically indistinguishable from those generated by a true random function.

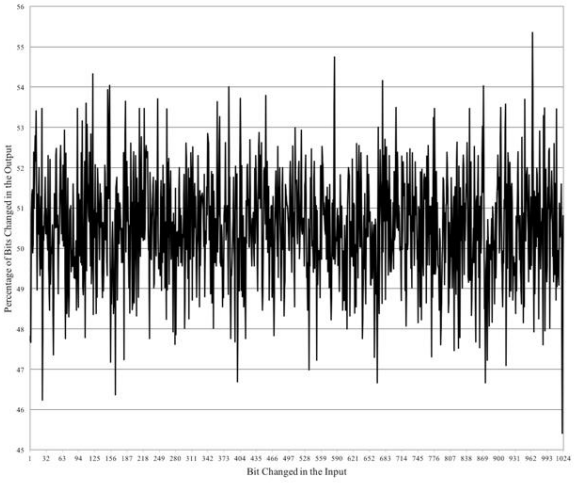


Fig. 3. Avalanche Effect for LCAHASH (128-Bit Version).

TABLE. I. Min, Max and mean Average Hamming Distances

Min	45.41%
Max	55.35%
Mean	50.51%

TABLE. II. DIEHARD TEST SUITE

Test	P-values	Interpretation
Diehard birthdays	0.39440811	Passed
Diehard operm5	0.53313565	Passed
Diehard_rank_32x32	0.57604853	Passed
Diehard_rank_6x8	0.72206931	Passed
Diehard_bitstream	0.94834578	Passed
Diehard_opso	0.13756532	Passed
Diehard_opso	0.92199512	Passed
Diehard_dna	0.63962068	Passed
Diehard_count_1s_str	0.49117533	Passed
Diehard count1s byt	0.93944197	Passed
Diehard parking lot	0.22547510	Passed
Diehard 2d sphere	0.64639589	Passed
Diehard 3d sphere	0.79712585	Passed
Diehard squeeze	0.83155292	Passed
Diehard sums	0.17939475	Passed
Diehard runs	0.79737984	Passed
Diehard craps	0.47732566	Passed
Marsaglia tsang gcd	0.85305736	Passed
Sts monobit	0.95959990	Passed
Sts_runs	0.64144586	Passed
Sts_serial	0.50158333	Passed
Rgb_bitdist	0.46256434	Passed
Rgb_minimum_distance	0.63795066	Passed
Rgb_permutations	0.59225124	Passed
Rgb_lagged_sum	0.49443699	Passed
Rgb_kstest_test	0.69428829	Passed
Dab_bytedistrib	0.56457052	Passed
Dab_dct	0.42650444	Passed
Dab_filltree	0.41658555	Passed
Dab_filltree2	0.54588999	Passed
Dab_monobit2	0.67956689	Passed

C. Performance

To measure the software implementation of LCAHASH 1.1, the Java source code was turned on an Intel Core i3-4010 32-bit processor clocked at 1.7 GHz with 4 Go of RAM. The results are presented in Table III. The performance of LCAHASH 1.1 is compared to LCAHASH 1.0 and other lightweight hash functions [9]. Those results show that LCAHASH 1.1 has a better performance than LCAHASH 1.0. Furthermore, LCAHASH has a satisfying performance when compared to other well established lightweight hash functions.

TABLE. III. SOFTWARE PERFORMANCE OF LCAHASH 1.1

Hash function	Output size (bit)	Cycle per byte (cpb)	Clock (GHz)
LCAHASH 1.0	128	324	1.7
	256	374	1.7
LCAHASH 1.1	128	995	1.7
	256	2844	1.7
GLUON	112	1951	2.66
U-QUARK	128	43373	2.66
D-QUARK	160	53103	2.66
S-QUARK	224	25142	2.66
PHOTON	80	1243	2.66

VI. CONCLUSIONS

In this paper, we presented a modified version of LCAHASH, a previously developed lightweight hash function. In our proposed design, we used a non-uniform cellular automaton with the rule set {30,90} to evolve the variable length input of our algorithm in order to generate a 128-bit or a 256-bit digest.

The new design preserves the security and statistical properties of the original design and has a better performance.

In future work, we should investigate non-uniform cellular automata more deeply and we should implement our algorithm in constrained devices to evaluate its hardware performance.

VII. CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Hanin, B. Echandouri, F. Omary, and S. E. Bernoussi, “L-CAHASH: A Novel Lightweight Hash Function Based on Cellular Automata for RFID,” Ubiquitous Networking Lecture Notes in Computer Science, pp. 287–298, 2017.

[2] S. Wolfram, A New Kind of Science. Champaign, Ill: Wolfram Media, 2002.

[3] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. B. Robshaw, Y. Seurin, and C. Vikkelsoe, “PRESENT: An Ultra-Lightweight Block Cipher,” Cryptographic Hardware and Embedded Systems - CHES 2007 Lecture Notes in Computer Science, pp. 450–466.

[4] G. Bertoni, J. Daemen, M. Peeters, and G. V. Assche, “On the Indifferentiability of the Sponge Construction,” Advances in Cryptology – EUROCRYPT 2008 Lecture Notes in Computer Science, pp. 181–197, 2008.

[5] E. B. Kavun and T. Yalcin, “A Lightweight Implementation of Keccak Hash Function for Radio-Frequency Identification Applications,” Radio Frequency Identification: Security and Privacy Issues Lecture Notes in Computer Science, pp. 258–269, 2010.

[6] J. Guo, T. Peyrin, and A. Poschmann, “The PHOTON Family of Lightweight Hash Functions,” Advances in Cryptology – CRYPTO 2011 Lecture Notes in Computer Science, pp. 222–239, 2011.

[7] A. Bogdanov, M. Knežević, G. Leander, D. Toz, K. Varıcı, and I. Verbauwhede, “Spongent: A Lightweight Hash Function,” Cryptographic Hardware and Embedded Systems – CHES 2011 Lecture Notes in Computer Science, pp. 312–325, 2011.

[8] J.-P. Aumasson, L. Henzen, W. Meier, and M. Naya-Plasencia, "Quark: A Lightweight Hash," Journal of Cryptology, vol. 26, no. 2, pp. 313–339, Oct. 2012.

[9] T. P. Berger, J. D'Hayer, K. Marquet, M. Minier, and G. Thomas, "The GLUON Family: A Lightweight Hash Function Family Based on FCSRs," Progress in Cryptology - AFRICACRYPT 2012 Lecture Notes in Computer Science, pp. 306–323, 2012.

[10] P. M. Mukundan, S. Manayankath, C. Srinivasan, and M. Sethumadhavan, "Hash-One: a lightweight cryptographic hash function," IET Information Security, vol. 10, no. 5, pp. 225–231, Jan. 2016.

[11] Damgård, I. B. (1989). A Design Principle for Hash Functions. Advances in Cryptology — CRYPTO' 89 Proceedings Lecture Notes in Computer Science, 416-427. doi:10.1007/0-387-34805-0_39.

[12] Daemen, J., Govaerts, R., & Vandewalle, J. (1991). A Framework for the Design of One-Way Hash Functions Including Cryptanalysis of Damgård's One-Way Function Based on a Cellular Automaton. Advances in Cryptology — ASIACRYPT 91 Lecture Notes in Computer Science,82-96. doi:10.1007/3-540-57332-1_7.

[13] Daemen, J., Govaerts, R., & Vandewalle, J. (1992). A Hardware Design Model for Cryptographic Algorithms. In: Deswarte Y., Eizenberg G., Quisquater JJ. (eds) Computer Security — ESORICS 92,648, 419-434. doi: https://doi.org/10.1007/BFb0013911.

[14] Chang, D. (2006). Preimage Attacks on CellHash, SubHash and Strengthened Versions of CellHash and SubHash (Vol. 2006, p. 412, Rep. No. 2006/412). IACR Cryptology ePrint Archive.

[15] Mihaljevic, M., Zheng, Y., & Imai, H. (1999). A Family of Fast Dedicated One-Way Hash Functions Based on Linear Cellular Automata Over GF (q). IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 40-47.

[16] Jeon, J. (2012). One-Way Hash Function Based on Cellular Automata. IT Convergence and Security 2012 Lecture Notes in Electrical Engineering, 21-28. doi:10.1007/978-94-007-5860-5_3.

[17] Kuila, S., Saha, D., Pal, M., & Chowdhury, D. R. (2014). CASH: Cellular Automata Based Parameterized Hash. Security, Privacy, and Applied Cryptography Engineering Lecture Notes in Computer Science, 59-75. doi:10.1007/978-3-319-12060-7_5.

[18] Hanin, C., Echandouri, B., Omary, F., & Bernoussi, S. E. (2017). L-CAHASH: A Novel Lightweight Hash Function Based on Cellular Automata for RFID. Ubiquitous Networking Lecture Notes in Computer Science, 287–298. doi: 10.1007/978-3-319-68179-5_25.

[19] Hirose, S., & Yoshida, S. (1997). A One-Way Hash Function Based on A Two-Dimensional Cellular Automaton. The 20th Symposium on Information Theory and Its Applications (SITA97), 213-216.

[20] P. Rogaway and T. Shrimpton, "Cryptographic Hash-Function Basics: Definitions, Implications, and Separations for Preimage Resistance, Second-Preimage Resistance, and Collision Resistance," Fast Software Encryption Lecture Notes in Computer Science, pp. 371–388, 2004.

[21] M. Bellare and T. Kohno, "Hash Function Balance and Its Impact on Birthday Attacks," Advances in Cryptology - EUROCRYPT 2004 Lecture Notes in Computer Science, pp. 401–418, 2004.

[22] A. F. Webster and S. E. Tavares, "On the Design of S-Boxes," Lecture Notes in Computer Science Advances in Cryptology — CRYPTO '85 Proceedings, pp. 523–534.

[23] R. Forrié, "The Strict Avalanche Criterion: Spectral Properties of Boolean Functions and an Extended Definition," Advances in Cryptology — CRYPTO' 88 Lecture Notes in Computer Science, pp. 450–468.

[24] G. Marsaglia, DIEHARD Statistical Tests: http://www. stat. fsu. edu/pub/diehard/.

An Empirical Comparison of Machine Learning Algorithms for Classification of Software Requirements

Law Foong Li[1], Nicholas Chia Jin-An[2]
Department of Computing
School of Computing and Creative Media
UOW Malaysia KDU University College
40150 Glenmarie Shah Alam, Malaysia

Zarinah Mohd Kasirun[3]
Department of Software Engineering
Faculty of Computer Science and Information Technology
University of Malaya
50603 Kuala Lumpur, Malaysia

Chua Yan Piaw[4]
Institute of Educational Leadership
Faculty of Education, University of Malaya
50603 Kuala Lumpur, Malaysia

***Abstract*—Intelligent software engineering has emerged in recent years to address some difficult problems in requirements engineering. Requirements are crucial for software development. Moreover, the classification of natural language user requirements into functional and non-functional requirements is a fundamental challenge as it defines the fulfillment criteria of the users' expected needs and wants. Therefore the research of this article aims to explore and compare random forest algorithm and gradient boosting algorithm to determine the accuracy of functional requirements and non-functional requirements in the process of requirements classification through the conduct of experiments. Random forest and gradient boosting are ensemble algorithms in machine learning that combines the decisions from several base models to improve the prediction performance. Experimental results show that the gradient boosting algorithm yields improved prediction performance when classifying non-functional requirements, in comparison to the random forest algorithm. However, the random forest algorithm is more accurate to classify functional requirements.**

Keywords*—Machine learning; ensemble algorithms; requirements classification; functional requirements; non-functional requirements

I. Introduction

Requirements are introductory building blocks for developing software projects. They are often classified into functional and non-functional requirements [1], [2]. In definition, functional requirements describe the system functionality whilst non-functional requirements describe system properties and constraints. This distinction has determined how requirements are being handled in practice; during elicitation, documentation, and validation [3].

Additionally, requirements are crucial in determining the success of a project; as it establishes a formal agreement between client and software provider working towards the same goal. However, the task of requirements categorization normally expends significant human effort and time when performed manually [4], [5]. The field of software engineering (SE) has witnessed remarkable progress in the past two decades attributable to the advancement of machine learning [6] and natural language processing.

Machine learning in natural language processing has become ever more accessible, leading to more innovations in software engineering. Many techniques and algorithms have been created and adapted into different systems, which has improved performance and overall computational efficiency. Numerous attempts have been made to construct automation for the assistance of extraction and classification of requirements using supervised [7], [8], [9] and semi-supervised learning techniques [10].

This paper aims to explore and compare the machine learning algorithms of random forest algorithm and gradient boosting algorithm. Both algorithms are employed to predict respective labelled data of the functional and non-functional requirements.

This paper is organized as follows: Section II describes the background and Section III discusses the related works. Followed by, Section IV presents the research methodology used. Section V exhibits the results of the study. Section VI presents the findings of the study. Section VII highlights the limitations of the study. Section VIII outlines future work to be undertaken. Finally, Section IX concludes the presented work.

II. Background

There are two primary types of learning schemes in machine learning: supervised learning, where the output has been given a priori labelled or the learner has some prior knowledge of the data; and unsupervised learning, where no prior information is given to the learner regarding the data or the output [11]. The following terms and tools employed in the study are briefly described as follows:

A. Python Pandas

Pandas is a software library written for the Python programming language intended for data manipulation and analysis. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames store and manipulate tabular data in rows of observations and columns of variables [11].

B. Scikit Learn

Scikit Learn is one of the most popular Python toolboxes. It provides a wide selection of supervised and unsupervised learning algorithms. It has advanced functions not commonly offered by other libraries including ensemble methods. Both random forest and gradient boosting are ensemble methods [11]. Both algorithms predict (regression or classification) by combining the outputs from individual trees.

C. Label

Label, also known as target array, defined as the number of categories the machine learning algorithm has to predict. It is generally contained in a NumPy array or Pandas Series [12]. For the purpose of this study, the algorithm of A1 only has two (2) labels which are functional or non-functional requirements. The algorithm of A2 has more than two (2) labels, also known as multiple labels, such as security, performance, usability, etc.

D. n-gram

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n-items from a text or speech corpus [11]. The items can be phonemes, syllables, letters, words or base pairs. The n-gram, n=1 is referred to as a "unigram"; n=2 as a "bigram"; and n=3 as a "trigram". For instance:

n-gram 1 = the, phone, rang
n-gram 2 = the phone, phone rang
n-gram 3 = the phone rang

E. Accuracy

Accuracy is one metric for evaluating classification models. This is the measure of the correct number of classifications divided by the total number of classifications [11].

III. Related Works

Machine learning has increasingly gained attention in software engineering. However, there are insufficient research works available in scholarly literature regarding carrying out an accurate comparison of machine learning algorithms for classification of software requirements, which could be used as a reference to conduct similar works in the future. These studies [7], [9], [10] concentrate on the classification of software requirements by tackling the machine learning approach through different models. Furthermore none of these studies will use the ensemble approach.

Kurtanović and W. Maalej [7] classified requirements as functional and non-functional requirements using support vector machine, abbreviated as SVM algorithm. Most present studies focuses on the classification of either functional or non-functional requirements. For example, Slankas and Williams [9] evaluated multiple classifiers to identify non-functional requirements and found the support vector machine had the highest effectiveness.

The approach proposed by Casamayor et al. [10] for the non-functional requirements identification is focused on semi-supervised text classification. The accuracy rates are above 70% for this proposed approach, significantly higher than the results obtained through the supervised method of using the standard collection of documents.

Instead, this study provides an implication for deciding accurate algorithm to categorize functional and non-functional requirements individually by attempting the ensemble approach which makes allowance for better predictions compared to a single model in order to close the research gap as mentioned earlier in this section.

IV. Research Methodology

A mixture of natural language processing algorithms and machine learning algorithms was used in the study. The machine learning algorithms were used to predict functional and/or non-functional labelled data, as well as labels such as security, usability, efficiency, etc. Alternatively, the natural language processing algorithms were used to generate a sentence(s) from the user requirements.

The main emphasis of this article is the machine learning algorithms. Therefore, no natural language processing algorithms will be discussed. In this article, algorithm A1 represents the algorithm that classifies the functional requirements and algorithm A2 represents the algorithm that classifies the non-functional requirements. Experiments are conducted to determine the machine learning algorithms. The random forest and gradient boosting algorithms was used to predict the respective labelled data of functional and non-functional requirements.

A. Data Preparation

The raw data was formatted and vectorised before passed into the random forest algorithm and the gradient boosting algorithm to perform model fitting and prediction. Then, the data frames were sorted with the relevant data to be used by the natural language processing algorithms at the next stage.

Once the file was read, it began to format the input data into a standard format. A few steps were taken to convert the raw data into a standard format as follows:

1) Remove all punctuations from the text.
2) Convert the text into lowercase.
3) Add a full-sentence column into the data frame.
4) Remove stop words.

The following vectorization steps were conducted to prepare the data for use by the algorithm (after the raw data has been formatted):

1) Create a new data frame that has only the required columns for the algorithm.
2) Split the training and test data into x and y coordinates respectively.
3) Vectorise the data to be used by the algorithm. This means that the words would be converted into unique

identifiers for the algorithm since the algorithm only accepts numerical data. For example, yes => 0 and no => 1.

4) Concatenate any features into the vectorised data.

B. Experimental Instruments

Python Pandas is well suited for different kinds of data. Furthermore, the DataFrame of Python Pandas can be created by loading datasets from existing external storage such as a SQL database, CSV files, list of dictionary, etc. Thus, the Python Pandas DataFrame is designated for the data manipulation in this study.

On the other hand, Scikit Learn was also chosen as a tool in this study because it is built on top of common data and Python Math Libraries. The design makes for ease of integration, whereby numpy arrays and pandas data frames can pass directly to the machine language (ML) algorithms of Scikit. In addition, it features numerous classification, regression, and clustering algorithms including support vector machines, random forest, gradient boosting, etc. Hence, two (2) ML algorithms in Python, which are random forest and gradient boosting, have been chosen for this study to classify functional requirements and non-functional requirements in turn.

C. Experimental Procedure

This section outlines the experimental procedure that classifies software requirements into two (2) different types, which are functional and non-functional requirements.

Fig. 1 illustrates the model fitting outline before model prediction. First, the system will read the selected file and check for errors in the file type chosen. When errors are found, it will prompt the user with an error message and allow for file re-selection. Once all errors are expunged, the system will format and prepare the data for use by algorithm A1 and A2. Following, the system will perform the model fitting, followed by the model prediction. Finally, the model predictions of algorithm A1 and algorithm A2 will be saved. The system will then return to the user, the trained or fitted model of algorithm A1 and algorithm A2 to make the prediction.

The purpose of conducting experiments on both algorithms is to determine which algorithm; random forest or gradient boosting is more accurate for algorithm A1 and algorithm A2. As mentioned at the beginning of this section, algorithm A1 represents the algorithm that classifies functional requirements and algorithm A2 represents the algorithm that classifies non-functional requirements.

There will be two (2) experiments in this study. Each experiment will have different sets of data. For example, a file that comprises a mixture of functional requirements and non-functional requirements will be used in conducting the first experiment to determine which algorithm, random forest or gradient boosting, is more accurate for algorithm A1. Objectively, the purpose of conducting the second experiment is to determine which algorithm, random forest or gradient boosting, is more accurate for algorithm A2. Hence, a file that consists of only non-functional requirements with their sub-category(s) will be employed to realize the purpose of the second experiment.

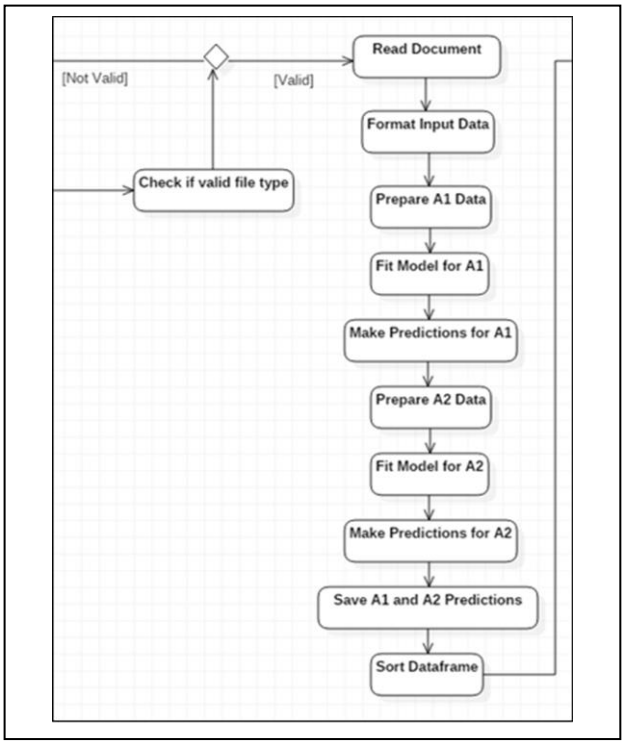


Fig. 1. Experimental Procedure of this Study.

In practice, a two (2) test per model configuration change is conducted in each experiment. The two (2) tests are as below:

1) The first test will be conducted five (5) times for each n-gram range and the averages are collected to find out which of the n-gram ranges are more accurate.

2) The second test will be conducted ten (10) times for each n-gram range and the averages are collected to reduce any variances that could come from the data splitting processes.

Nonetheless, there might be the circumstance of a new setting that leads to retesting the n-gram range to determine which n-gram range is most accurate. In such circumstance, more tests, to be conducted repetitively, was required. Such circumstance could also be a test to find a new setting, such as the number of estimators, also known as the number of decision trees built by the algorithm, or the depth of the decision trees. So, whenever a change is made in the number or depth of estimators, for example, the n-gram test needs to be conducted again to see the implications on the n-gram range.

V. Results of Study

The results of algorithm A1 that applied random forest and gradient boosting technique with two parameters which are the number of trees and maximum depth of the trees are shown respectively in Table I and Table II. Separately, Table III and Table IV illustrates the results of algorithm A2 that applied the same technique and parameters as presented in Table I and Table II. There are some null values of maximum depth of the trees appears in Table I and Table III which means no limits or infinite depth of the trees.

The final results of algorithm A1 and A2 are revealed in Table V and Table VI. Table V shows that the random forest technique has an advantage in the average accuracy compared with the gradient boosting algorithm. The random forest algorithm achieved a higher average accuracy which is 0.826 in comparison with the gradient boosting algorithm which is 0.789. Thus, the measured values deviated by the range of 0.037. It can be concluded that the random forest is more accurate for algorithm A1, which is responsible to classify functional requirements.

In contrast, Table VI illustrates that the gradient boosting algorithm is more accurate for algorithm A2. The gradient boosting algorithm obtained a higher average accuracy which is 0.591 compared to the random forest algorithm which only has 0.582 of the average accuracy.

It is noteworthy to indicate that the number of labels of raw data has influenced the average accuracy of the algorithm significantly. In this study, the random forest and gradient boosting algorithms was used in the first experiment to predict functional and non-functional labelled data. Meanwhile, the random forest and gradient boosting algorithms was used in the second experiment to predict non-functional labelled data such as security, usability, efficiency, and so on. The efforts to predict multiple labels of raw data in the second experiment is tougher than that of the binary labels in the first experiment. The machine learning algorithms had to predict binary labels of raw data such as functional requirements and non-functional requirements in the first experiment, which would give 50% accuracy through blind guessing. In contrast, multiple labels of raw data from different categories of non-functional requirements, for instance, ten (10) labels, needed to be predicted in the second experiment which gave 10% accuracy through the prediction process. This is because the algorithm was required to choose from the available labels rather than to perform blind guessing. It also sets the baseline for comparing the accuracy of the algorithms.

In summary, the results of this study shows that the random forest algorithm is more accurate for algorithm A1 whereas the gradient boosting algorithm is more suited for algorithm A2 due to accuracy.

TABLE. I. RESULT OF ALGORITHM A1 USING RANDOM FOREST WITH NUMBER OF TREES AND MAXIMUM DEPTH OF THE TREES PARAMETERS SETTING

Number of Trees	Maximum Depth of the Trees	Precision	Recall	Accuracy	Predict Time
50	40	0.892	1	0.915	0.106
200		0.892	1	0.915	0.103
100	40	0.900	1	0.894	0.104
50	30	0.868	1	0.894	0.108
100	30	0.868	1	0.894	0.104
100		0.889	0.970	0.894	0.104
200	40	0.868	1	0.894	0.104
50	30	0.878	1	0.894	0.104
100		0.878	1	0.894	0.104
200		0.919	0.944	0.894	0.104

TABLE. II. RESULT OF ALGORITHM A1 USING GRADIENT BOOSTING WITH NUMBER OF TREES AND MAXIMUM DEPTH OF THE TREES PARAMETERS SETTING

Number of Trees	Maximum Depth of the Trees	Precision	Recall	Accuracy	Predict Time
150	11	0.914	0.970	0.915	0.001
150	3	0.889	0.970	0.894	0.001
150	11	0.912	0.939	0.894	0.001
300	3	0.889	0.970	0.894	0.001
300	11	0.912	0.939	0.894	0.001
450	11	0.889	0.970	0.894	0.001
450	3	0.848	1.000	0.894	0.001
450	3	0.935	0.879	0.872	0.001
150	7	0.886	0.939	0.872	0.001
300	7	0.910	0.909	0.872	0.001

TABLE. III. RESULT OF ALGORITHM A2 USING RANDOM FOREST WITH NUMBER OF TREES AND MAXIMUM DEPTH OF THE TREES PARAMETERS SETTING

Number of Trees	Maximum Depth of the Trees	Precision	Recall	Accuracy	Predict Time
50	20	0.586	0.586	0.586	0.125
50	40	0.621	0.621	0.621	0.109
50		0.621	0.621	0.621	0.123
100	20	0.586	0.586	0.586	0.110
100	40	0.621	0.621	0.621	0.121
100		0.655	0.655	0.655	0.120
200	20	0.621	0.621	0.621	0.109
200	40	0.655	0.655	0.655	0.110
200		0.586	0.586	0.586	0.109
50	20	0.621	0.621	0.621	0.109

TABLE. IV. RESULT OF ALGORITHM A2 USING GRADIENT BOOSTING WITH NUMBER OF TREES AND MAXIMUM DEPTH OF THE TREES PARAMETERS SETTING

Number of Trees	Maximum Depth of the Trees	Precision	Recall	Accuracy	Predict Time
150	11	0.586	0.586	0.586	0.001
300	11	0.621	0.621	0.621	0.001
450	11	0.621	0.621	0.621	0.001
600	11	0.517	0.517	0.517	0.001
750	11	0.586	0.586	0.586	0.002
900	11	0.655	0.655	0.655	0.002
1050	11	0.552	0.552	0.552	0.002
150	11	0.517	0.517	0.517	0.001
300	11	0.586	0.586	0.586	0.001
450	11	0.586	0.586	0.586	0.001

TABLE. V. FINAL RESULTS OF ALGORITHM A1

Setting	Average Precision	Average Recall	Average Accuracy	Predict Time
Random Forest	0.802	0.965	0.826	0.111
Gradient Boosting	0.809	0.882	0.789	0.658

TABLE. VI. FINAL RESULTS OF ALGORITHM A2

Setting	Average Precision	Average Recall	Average Accuracy	Predict Time
Random Forest	0.582	0.582	0.582	0.105
Gradient Boosting	0.591	0.591	0.591	0.131

VI. DISCUSSION

There are several tuning parameters important for random forests and gradient boosting algorithm, however, only two parameters which are the number of trees and the tree depth were chosen in this discussion. Generally, a higher number of trees increased the performance and made the predictions more stable which could result in better accuracy. In comparison, more trees also meant more computational cost and after a certain number of trees, the improvement was negligible. The results in Table I show that there was no significant improvement on the accuracy; for example, the first and second record obtained the same accuracy rate which was 0.915, but both had different number of trees 50 and 200 for the first and second record respectively. It is concluded that as the number of trees grows, it does not always mean the performance of the forest is significantly better than previous forests which had fewer trees. The results of Table I indicates that the addition of trees is insignificant. This result is aligned with the previous studies finding [14] revealing that the smallest number of trees is sufficient to obtain the same level of accuracy. Barman et al. [15] also discovered that there is no significant difference between using a number of trees, and larger number of trees in a forest will not significantly improve the performance but in contrary, it will increase its computational cost.

Table II shows that the first record gained the highest accuracy which was 0.915 with 150 number of trees and 11 depth of the trees. The results in Table II yet again shows that larger number of trees has no significant improvement on the accuracy.

Table III displayed that the eighth record listed has the highest accuracy value of 0.6555 in contrast to the seventh and ninth record listed which possesses the same number of trees, but have differing depth of trees which are 20, 40 and infinite depth of trees respectively. The results indicate that depth of the trees has a significant effect on the accuracy. As the depth increases, the stability of prediction will decrease as each model tends to cause overfitting [16]. The depth of the tree meaning length of tree. Larger tree helps to convey more information whereas smaller tree gave less precise information. Hence, there needed to be a balanced ratio within the depth of trees to gain a better performance.

Alternatively, the findings in Table IV indicate the number of trees will determine the accuracy of prediction when the depth of the trees value is constant. The sixth record listed in Table IV gained the highest accuracy which was 0.655.

The parameters in random forest and gradient boosting are either to increase the predictive power of the model or to make it easier to train the model. Important parameters to fine tune would be the number of trees, the depth of trees and the number of features used for a split. However, the number of trees and the depth of trees were selected to perform the fine tuning the models of this study. Optimistically this article has given essential understanding to begin using the random forest and gradient boosting on projects.

VII. LIMITATION OF STUDY

As with all research, there were a few limitations to this study that must be acknowledged. Firstly, the system is limited by the amount of data due to the issue of confidentiality; similar to that encountered in similar research [13]. As more quality data is collected, the accuracy of algorithm A1 and algorithm A2 will increase. Secondly, the system has a limitation on the format of data read, as it requires data to be in a specific format to function properly. It is also limited by its inability to determine how many separate requirements there are in a sentence, as well as its inability to read from .docx and .txt files.

VIII. FUTURE WORK

With the limitations of study stated in Section VII. For future work, the following aspects are identified:

1) Collect more data to be used in experiments.

2) Generalize the reading format to allow the reading of data in a more general format.

3) Implement a function to determine how many different requirements exist in a sentence.

IX. CONCLUSION

This article explores and compares the random forest algorithm and the gradient boosting algorithm to discover which is more accurate to classify functional requirements and non-functional requirements, by conducting experiments. Among the investigated machine learning algorithms, the results of this study have shown empirically that the gradient boosting algorithm yields better prediction performance in terms of accuracy when sorting non-functional requirements, in comparison to the random forest algorithm.

For future work, more machine learning algorithms will be investigated by engaging the ensemble strategy in order to improve the overall classification accuracy.

ACKNOWLEDGMENT

This project has received funding from the KDU Research Grant under grant no. KDURG/2017/010.

REFERENCES

[1] G. Kotonya, & I. Sommerville, Requirements engineering with viewpoints, Software Engineering Journal, 11(1), 5-18, 1996.

[2] R. Y. Lee, Requirements Elicitation Software Engineering: A Hands-On Approach (pp.81-102). Paris: Atlantis Press, 2013.

[3] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc. , “The detection and classification of non- functional requirements with application to early

aspects". In 14th IEEE International Requirements Engineering Conference (RE'06). IEEE, 2006. pp. 38-49.

[4] J. Cleland-Huang, R. Settimi, X. Zou, & P. Solc, Automated classification of non-functional requirements. Requirements Engineering, 12(2), 103-120, 2007.

[5] A. Rashwan, O. Ormandjieva, & R. Witte, "Ontology-based classification of non-functional requirements in software specifications: a new corpus and svm-based classifier". In 2013 IEEE37th Annual Computer Software and Applications Conference. IEEE, July 2013. pp. 381-386.

[6] D. Zhang, & J. J. Tsai, Machine learning and software engineering, Software Quality Journal, 11(2), 87-119, 2003.

[7] Z. Kurtanović and W. Maalej, "Automatically Classifying Functional and Non-functional Requirements Using Supervised Machine Learning," In 2017 IEEE 25th International Requirements Engineering Conference (RE). IEEE, September 2017, pp. 490-495.

[8] M. Lu, & P. Liang, Automatic classification of non-functional requirements from augmented app user reviews. In Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering. ACM, June 2017. pp. 344-353.

[9] J. Slankas, & L. Williams, Automated extraction of non-functional requirements in available documentation. In 2013 1st International Workshop on Natural Language Analysis in Software Engineering (NaturaLiSE). IEEE, May 2013, pp. 9-16.

[10] A. Casamayor, D. Godoy, & M. Campo, Identification of non-functional requirements in textual specifications: A semi-supervised learning approach, Information and Software Technology, 52(4), 436-445, 2010.

[11] I. Laura and S. Santi, Introduction to Data Science A Python Approach to Concepts, Techniques and Applications. Springer, Switzerland, 2017.

[12] V.Jake, Python Data Science Handbook. O'Reilly Media, Inc, Sebastopol, United States, 2016.

[13] B. Athuraliya and C. Farook, ""Revyew" Hotel Maintenance Issue Classifier and Analyzer using Machine Learning and Natural Language Processing," In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018, pp. 274-280.

[14] P. Latinne, O. Debeir, & C. Decaestecker, Limiting the number of trees in random forests. In International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, July 2001. pp. 178-187.

[15] T. M. Oshiro, P. S. Perez, & J. A. Baranauskas, How many trees in a random forest? In International workshop on machine learning and data mining in pattern recognition .Springer, Berlin, Heidelberg, 2012, July. pp. 154-168.

[16] C. B. Liu, B. P. Chamberlain, D. A. Little, & Â. Cardoso, Generalising random forest parameter optimisation to include stability and cost. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham. 2017, September. pp. 102-113.

PathGazePIN: Gaze and Path-based Authentication Entry Method

Bayan M. AlBaradi[1], Amani M. AlTowayan[2], Maram M. AlAnazi[3], Sadaf Ambreen[4], Dina M. Ibrahim[5]
Information Technology Department, College of Computer, Qassim University, Qassim, Saudi Arabia[1, 2, 3, 4, 5]
Computers and Control Engineering Dept., Faculty of Engineering, Tanta University, Egypt[5]

***Abstract*—In these days, smartphones are being widely used, people use them to store sensitive and private information. The password authentication is the most used authentication method, however, the password can be disclosure through shoulder surfing attack, since the users write their password in public places and they tend to make the password easy to remember which increase the vulnerability to attacks. Many authentications schemes were proposed to prevent shoulder surfing attack, but they still vulnerable to shoulder surfing attack or have accuracy problems or lack in usability. In this paper, we proposed a gaze-text based PIN entry method, which we called PathGazePIN. It will utilize the user's eye movement to enter the password. The main idea is allowing the user to authenticate by following numbers that move along fixed paths on the screen by using two authentication interfaces: random interface and sorted interface. The results represented that the proposed system will increase the security against shoulder surfing attack as well as usability and accuracy.**

***Keywords*—User authentication; shoulder surfing attacks; PIN; gaze-based; security; accuracy**

I. Introduction

The smartphone has become in large-scale utilization because it can be carried anywhere easily and a lot of people have sensitive information stored on their devices, however, the smartphone may be vulnerable to many attacks, especially in crowded places. Accordingly, the password authentication came to ensure security for the user smartphone. Shoulder surfing attack is one of the dangerous attacks [1]. It constitutes a growing danger of device security. The attacker can obtain sensitive information by two ways, first at close range by directly looking over the person's shoulder either by side attack; when the user near to the attacker or even can see the device screen of a user or iterative attack; the attacker monitors the user when he/she enters the password and notes the repeated movements at several times to detect [2], second from long range by using a camera, a pair of binoculars or other vision magnifiers [3].

Shoulder surfing attack is kind of the nontechnical attacks on social engineering used by an attacker to obtain sensitive and private information such as password, personal identification number, and other confidential information, while the user is unaware [4]. The attacker can easily infiltrate to the victim when they are using devices such as Automated Teller Machine (ATM), smartphones or any other devices in crowded places to steal the password without user perception. This project will use gaze-based authentication that allows a user to enter the passwords using his eyes instead of touch to prevent shoulder surfing.

The rest of this paper is organized as follows; Section II presents background on the authentication mechanisms used to prevent shoulder surfing attacks. Our proposed PathGazePIN system is illustrated in Section III followed by the methodology and the implementation of the proposed PathGazePIN system in Sections IV and V, respectively. In Section VI, testing of the proposed system is presented while in Section VII we evaluate our results. Analysis and discussion for our system is given in Section VIII. Finally, we draw our conclusions and present issues for future work in Section IX.

II. Background

There are many authentications schemes were proposed to develop the authentication mechanisms and to prevent shoulder surfing attack, like textual password, graphical password, and biometrics password.

In the graphical password, three schemes were proposed to prevent shoulder surfing attack [5], which are triangle scheme, movable frame scheme, and intersection scheme. The main idea of these schemes is making the screen so crowded to confuse attackers, the main drawback of these schemes are they require long login time since the user needs to find the pass object on the screen, and if a small number of objects is shown on the screen the system will be vulnerable to attacks. To overcome of triangle scheme drawback Convex Hull Click Scheme [6] has been proposed to improve security and usability, this scheme was proved to be effective against Shoulder Surfing attacks, but it also requires long login time. To decrease login time ColorLogin scheme was proposed in [7], in this scheme they use a color background, which will decrease login time and multiple colors are used to confuse the attacker, while not burdening the users.

In textual password, in [8], they proposed a text-based authentication scheme that prevents Shoulder Surfing attacks using Randomized Square Matrix Virtual Keyboard. This scheme designed to be used in websites that require authentication, this scheme is required long login time due to, the two clicks for each character and the randomization. In this study [9], the scheme has proposed to preventing Shoulder Surfing attack by use of augmented reality. In this system, the user sees a keyboard layout that differs from the actual keyboard QWERTY through augmented reality device. Because the adversary does not have a knowledge of Strategies used consequently, the system will be more secure [10] [11].

On the other hand, typing on a randomized augmented reality keyboard requires more time and it produces errors during writing, which affects usability [12].

In Gaze-based authentication systems, The EyeDent authentication scheme was proposed in [13]. It uses an on-screen keyboard and users use gaze to enter the password, this scheme uses a numeric keypad and also used a QWERTY layout. A novel gaze-based authentication scheme was presented in [14], which is depends on a graphical password by using cued-recall on a single image, this scheme uses a computational model of visual attention to mask the area of images that most user select and it most likely to attract visual attention as called silence map. The study has shown that is images with saliency mask is significantly was more helpful to increase password security [15]. A secure multimodal authentication using gaze and touch PIN called Gaze Touch Cross PIN is presented on [16]. They proved that Gaze Gesture based system provides the most accuracy for authentication.

Smooth pursuit oculomotor control kit (SPOCK) was proposed in [17] it is based on smooth eye movement, selection, and activation to limits the Midas touch problem. Midas touch problem is excessive inadvertent clicking that place intended content outside the clickable area. In this study [3], Gaze-touch pin technology has proposed to cover two models of threats, first is side attack model and second is the iterative attack model. It uses touch and gazes' gestures as an input method. It is also used randomization of the digit to make the system more secure, this system was highly secure against Iterative attack. The authors in [18] has proposed a gaze-based authentication system. The system uses movable shapes and the user should tracks this shapes in the screen for authentication, the shapes have a minimal number to achieve complexity but do not clutter the interface, their study showed that there is no user have an issue with following moving shapes, due to the natural ability of the humans. Moreover, it does not add any cognitive load in human memory. The system has a limitation such that the user should adjust his position to let the Eye Tracker work efficiently and it also requires additional hardware. In [19], they proposed a gaze gesture-based authentication scheme. In this scheme, the interface consists of ten colored circles and each circle is moving in a random path and the user selects four colors as his password. This scheme was secure against video iterative attacks and has less security against dual video iterative attacks. The scheme has few limitations, it is not usable for colorblind users and also it requires long login time.

III. The Proposed PathGazePIN System

This research will provide a new idea for mobile authentication that is the combination of two previous studies [17][19]. By applying textual method and gaze-based technique that enable users to enter their password using their gaze and using SPOCK novel gaze interaction method which will avoid the Midas touch problem and increase accuracy. The system consists of three certain circular paths and ten numbers from 0-9 placed across these paths, the numbers on inner and on outer circular paths are moving, while the numbers on middle circular paths are static. These numbers are randomized across the circular paths but in increasing order for each authentication. The user selects 4 numbers as his/her password and on the authentication, the user either dwells in the static number or follow the moving number to enter each number, as illustrated in Fig. 1.

The static and dynamic approach is used instead of pure dynamic to avoid overwhelmed by the visual cluttering and to avoid more attention requires from the user which affects the usability, also certain circular paths are used to avoid the problem of two numbers with a similar path that may be wrongly recognized by the system which affects the accuracy of the system. Because of the gaze-based entering authentication, the system will be secure against side attack, furthermore, the randomizing of the number is used to completely prevent iterative attack and video analysis attacks. Ordered numbers technique is used to since it reduces the time of finding numbers so it will reduce login time. Also, the nature of human ability to follow moving objects and static and dynamic approach leads to more usability. Moreover, using certain circular paths and Gaze Pursuit-based Authentication will increase the accuracy.

Fig. 1. Application Interface of the Proposed PathGazePIN System.

IV. PathGazePIN Methodology

This system will implement for mobile devices by using gaze-based technique in Android platforms. By apply empirical study to collecting and analyzing data, we check the result whether it is achieving high accuracy, usability, and security from the result of user's experiment. The selected method is an empirical study about the effectiveness of gaze-based technique against shoulder surfing attacks which will focus on the security, usability, and accuracy. We are following the waterfall model, in planning phase we sought about anti-shoulder surfing technique, then analyzed the collected data from the previous research paper to define an appropriate solution, after that, we designed the proposed idea and conduct system implementation, testing procedures and result analyzing.

For testing, we collected a group of participants to measure the system accuracy and effectiveness. The requirements of this study are participants requirement, hardware and software requirements. The participants are a group of victims and

attackers. They must have a background of using smartphones and none of them should be blind. The hardware requirement is a smartphone with an Android operating system to implement the application on Android platform and it does not need for any additional hardware. The application is based on gaze gesture that will detect it by using front camera of the mobile, and software requirement is will analyze the user's eyes movements by using the MATLAB tool. Also, the application will be implemented on Android Studio Development Environment. The data will be collected by participants who will use the application after implementation, a database will be used to store authentication information; login time, the number of wrongs PIN to determine the accuracy and PIN. And other information will be collected manually; the usability and effectiveness against shoulder surfing attack. For analyzing collected data, the Microsoft Excel application will use.

V. PathGazePIN Implementation

A. Application Interface

In this application, Java Android studio is used to design the interface, which consists of animated numbers moving in a circular path and video capture to take a video of user's face while he tracks the numbers. The animated numbers are created by using Canvas and Paint class, and video capture is created by using MediaRecorder class.

Similarly, two types of interface are created which are random interface and sorted interface, as in Fig. 2. In the random interface, the application will randomly generate different arrange of numbers for every start. In the sorted interface, the numbers will appear same arrange for every start.

B. Eye Tracking System

For eye tracking system the MATLAB environment is used, eye Tracking system is consisting of three phases, first is eye extraction, in this phase a video of one eye will be created from recorded video. This video will be used in eye tracking phase to detect and track the pupil and save its X, Y and Z data, then these data are used in the PIN identification phase to determine desired number, for that a different calculations are used for each path, based on the feature of the data on each path.

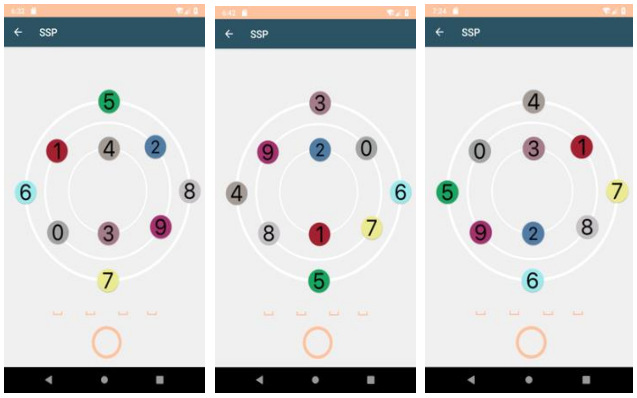


Fig. 2. Random Interfaces for different Login Times.

Outer path, this path consists of four numbers that are moving clockwise, to determine the desired number, first this path has been divided into two halves and then these halves are further divided into two quarters which results as four quarters, namely Right up, Right bottom, Left up and Left bottom. To determine if the gaze moves right or left the values that on y-axes and z-axes are used. The basic idea on this step is to get the highest and lowest y-axes value, and it correspond z-axes value (frame number), after that, we can determine if Y value goes from high to low or from low to high. On 'Left' case, the lowest Y value will have corresponding Z value bigger than the highest Y value, which is mean the plot goes from up to bottom. On 'Right' case, the lowest Y value will have corresponding z value less than the highest Y value which means the plot goes from bottom to up.

To determine if the gaze goes up or bottom the values that on x-axes and z-axes are used. The same idea of determining left and right are used, first get the highest and lowest x-axes value, and it correspond z-axes value (frame number), after that, we can determine if x value goes from high to low or from low to high. On 'Up' case, the lowest X value will have corresponding Z value bigger than the highest X value, which is mean the plot goes from right to left. On 'Bottom' case, the lowest X value will have corresponding z value less than the highest X value. Which is mean the plot goes from left to right. Finally, by combining these two steps we can determine if the gaze goes Right up, Right bottom, Left up or Left bottom.

Middle path: This path consists of four numbers that are static. To determine the desired number this path has been divided into four quarters, Right up, Right bottom, Left up and Left bottom. To determine if the path is static or not the y-axes value is used the biggest y-axes value is subtracted from lowest y-axes value to measure the distribution of the plot. After that, the median values of y-axes and x-axes are used with center point values that we get it from the previous phase to determine the desired quarter.

Inner path: This path consists of two numbers that are moving anti clockwise. Therefore, to determine the desired number this path has been divided into two halves Right and Left. The anti-clockwise path is used to increase the accuracy of the system. To determine if the gaze goes right or left the value that is on y- axes and z-axes is used. On 'Right' case, the lowest Y value will have corresponding Z value bigger than the highest Y value, which is mean the plot goes from up to bottom. On 'Left' case the lowest Y value will have corresponding Z value less than the highest Y value, which is mean the plot goes from bottom to up. to distinguish the inner path from the outer path the x-axes value and y-axes value are used. The main idea is in the inner path the lowest or the highest x-axes value will be in the middle, wherein lowest or the highest x-axes value in the outer path is in the end or in the beginning. Examples of outer left up plot and inner right plot are shown in Fig. 3 and Fig. 4, respectively.

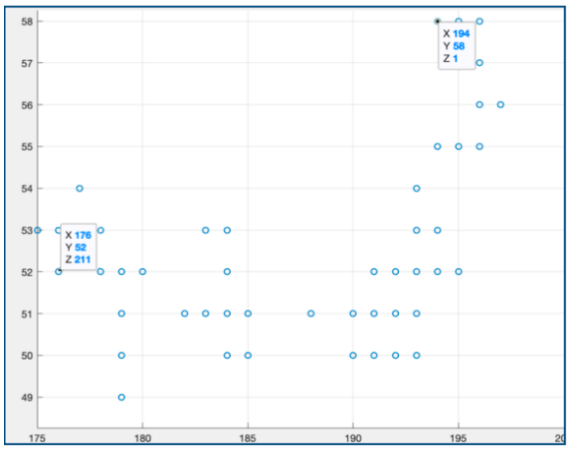


Fig. 3. Outer Left up Plot.

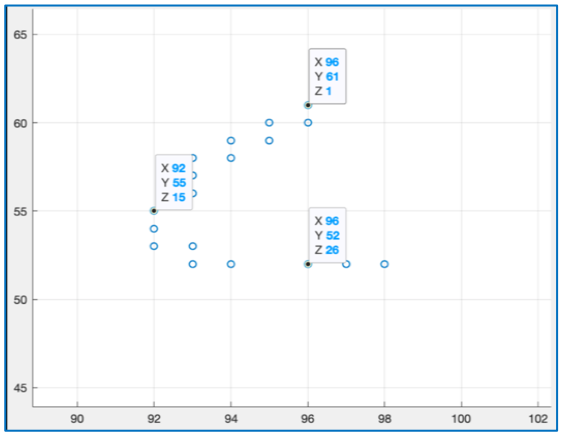


Fig. 4. Inner Right Plot.

VI. PathGazePIN Testing

For connecting the eye tracking system with Android Interface, the MATLAB DRIVE is used; the recorded videos are uploaded to MATLAB drive and then execute them through eye tracking system in the laptop. Before the connection, there are some changes is applied on eye tracking system. First, for each phase it will takes four videos on each run instead of one video, moreover in PIN identification phase the change is applied to save the PIN as a number instead of its place on the circle based on its place and the video number.

We tested three important aspects of authentication systems which are usability, accuracy, and its security of shoulder surfing attacks. We perform three experiments: usability experiments, accuracy experiments and security experiments. We did a usability experiment for two modes interface, one with random numbers and other with sorted numbers. To measure the usability in terms of which of them takes less time and easy to remember, we made a database to measure the time for every run. The test is conducted on a number of people and divided them as attackers and victims. The victims choose the mode of interface either sorted or random, or can use both. A questionnaire was made after the test with some questions to take the users opinion and information. Furthermore, we applied an experiment on two interface modes to analyze the security of the application if it is exposed for shoulder surfing attacks (side, iterative) attacks. We conducted a number of complete logins with different modes, which are random or sorted. In full session, there can be two attackers one of them is side attack and other is iterative attack. To check if this approach is exposed to iterative attack, victim will have to login more than one times with same password. Also, we measure the time if victim takes a long time to enter his/her password that will be easy for attacker to detect the password or not.

VII. Evaluation Results

A. Security

The most important aim of our application is shoulder surfing prevention. The results appear percentage of success login for attackers in both interfaces. In sorted interface, there are 30 logins trying for users that means there are 30 trying from the attackers to guessing the numbers. The success of iterative attackers for sorted interface are 7 times from 30 means 23.3% the attacker was guessed one to two numbers but not in the same order. For side attack there are 4 times from 30 attackers which means 13.3% guess one to two numbers also not in the same order. The success of attackers for the random interface are 2 times for iterative attacks from 15 trying 13.3%, and for side attacks was guess one number 2 times from 15 trying 13.3% also not in the same order as in Fig. 5.

We requested from users to enter their password fourth times, and attackers are different for each user to measure how many numbers attackers can guess in more than one login for user. The first login for 18 users represents 100% for first trying to attackers to guessing. There are four attackers for different users was guess just one number 22.2% and there two attackers were guessing two numbers for each user 11.1%. In the second logging for 12 users, there are four attackers three of them were guess one number which by 33.3% and the other was guess two numbers 8.3%. In the third login, one attacker was guessed one number 12.5% from 8 tries, and for the fourth login, there were 5 users try to log in for the fourth times and just two attackers were guessed two numbers 40% as shown in Fig. 6. At the end, all attackers that guessed the numbers, they did not guess it in the same order that users put at their password. They trying to guess but no one of them was guess it at same order like user chosen.

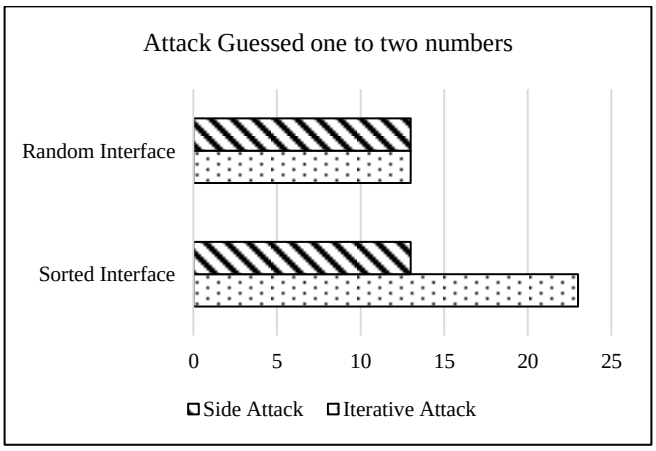


Fig. 5. Attacker Guessed One to Two Numbers.

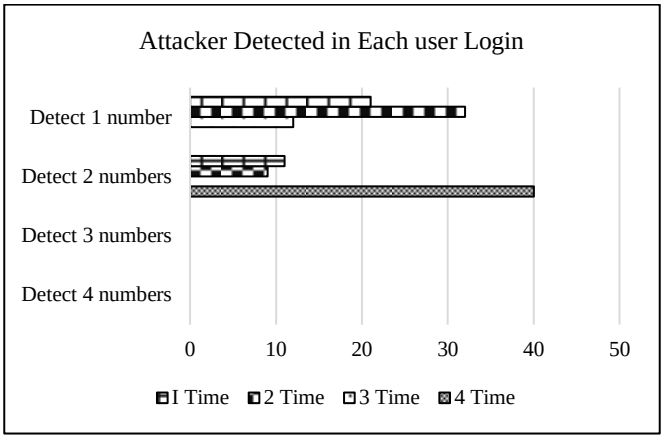


Fig. 6. Attacker Detection on Each user Login.

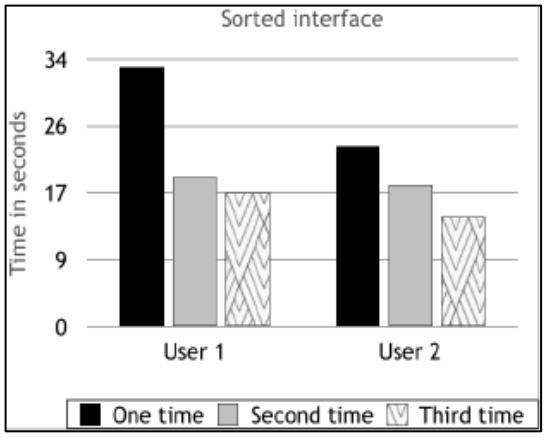


Fig. 9. Maximum, Minimum and Average Login Time on Sorted Interface.

B. Usability

Fig. 7 and 8 show four persons use the same mode with the same password multiple times. In the sorted interface, the time required from the user to complete the login was decreased. While, in the random interface the time required from the user to complete the login did not change.

Fig. 9 and 10 show the average, maximum and minimum login time for all the login in the sorted interface and the random interface. In the sorted interface, the average login time was 23s and the maximum login time was 37s and the minimum login time was 14s, while the average login time in the random interface was 24s and the maximum login time was 39s and the minimum login time was 18s.

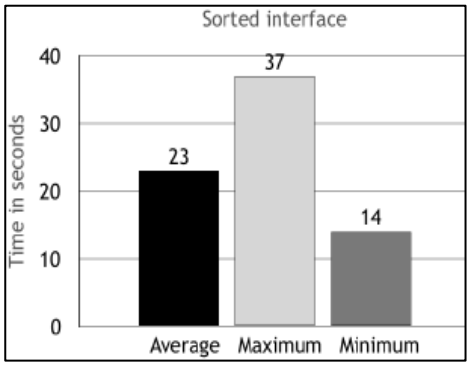


Fig. 7. Login Time for Three users on Sorted Interface.

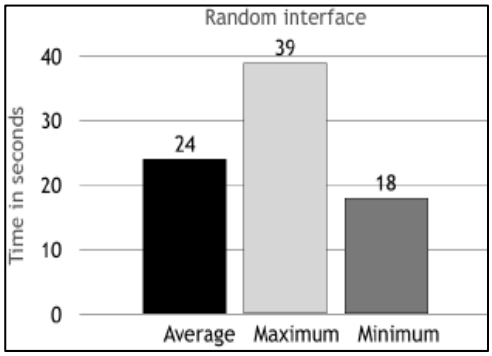


Fig. 8. Login Time for Two users Multiple Time on Random Interface.

C. Accuracy

To measure the accuracy of our eye tracking system the data was collected from users based on different situations to analyze factors that affect the accuracy, also with different numbers to analyze the accuracy of each path. In this experiment, the sorted interface was used and ten logins were performed with seven different passwords, the total entered numbers were 40 number.

In this experiment, the successful login was 40 %, as illustrated in Fig. 11, the means by successful login the eye tracking system detects all four number successfully. On the other hand, the unsuccessful login was 60%, the unsuccessful mean that the eye tracking system does not detect all four number successfully, it may be detected 0, 1, 2, or 3 numbers successfully.

For all the entered numbers, Fig. 12 shows the percentage of true detection of numbers and the percentage of wrong detection of numbers. The percentage of true detection of numbers was 75% which means 30 numbers out of 40 are detected, and the percentage of the wrong detection of numbers was 25% which means 10 numbers out of 40 are not detected truly.

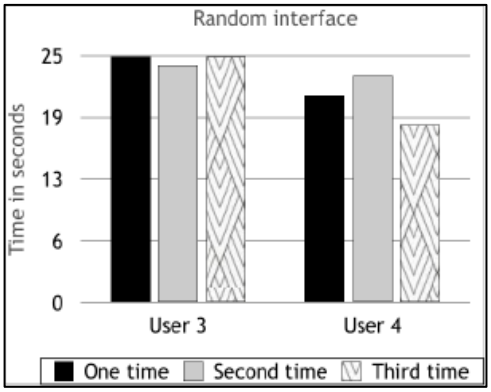


Fig. 10. Maximum, Minimum and Average Login Time on Random Interface.

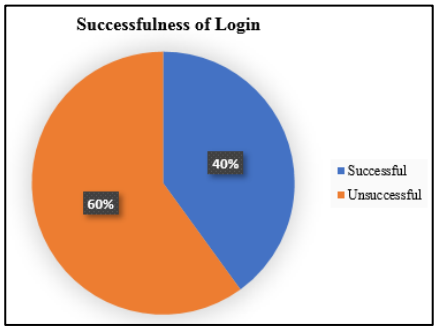


Fig. 11. Successfulness of Login.

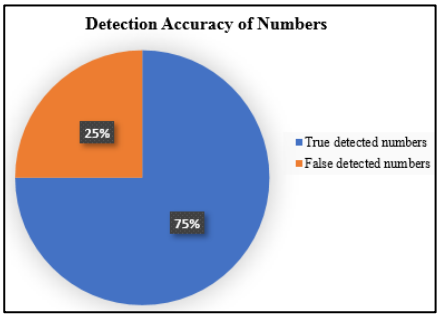


Fig. 12. The Detection Accuracy of Numbers.

VIII. Analysis and Discussion

A. Security-based Analysis

The first and important aim of our project is security. After we measured the security for each interface and how many percent for each trying was success, we found that both interfaces are secure from shoulder surfing attack. There is no successful complete login for attacks which guess the completely four numbers and with same order, and the attacker who guess one number or two numbers did not guess these numbers with same place of the number. For 45 user's login there are attackers by 28% who were guess one or two numbers only and not with the same order.

B. Usability-based Analysis

We have concluded that the difficulty of the password is not affected by login time. Also, if the user used sorted interface more than once it will reduce time login. While, if used random mode it will change at each run. The interface is user friendly and did not faced attenuation but the sorted interface is easier than random interface.

C. Accuracy-based Analysis

The third aim of our project was achieving high accuracy on the login. Unfortunately, this aim was not achieved because of the Eye Tracking System has just 40% successful detection of logins, and the unsuccessful detection was 60% of logins. From the 60% unsuccessful logins, 30% of them detected three number successfully, 20% of them detected two number successfully and 10% of them detected just one number successfully. This is being because a front camera of the smartphone is used, also there was no use special software for analyzing gaze data, in all previous studies that were studied in this project there was use a special tracking camera within its software, except one [3] which was used the device's front camera but it was tracking only the horizontal gaze direction.

IX. Conclusions and Future Work

In this paper, we presented a Shoulder Surfing Attacks Prevention scheme by using Enhanced gaze PIN Entry Method which we called PathGazePIN system, we applied it on Android device and for the tracking, the front camera was used. We implemented two authentication interfaces: Random interface and Sorted interface, and it was evaluated to measure the usability and security for each of them. Through system evaluations, for the usability, we found that the Sorted Interface has less time login if used more than once, while, the time in the Random Interface did not change, for the user perspective, we were found that the interface is user-friendly and did not require more attention from the user. For the security, we found that the sorted interface was less secure compared with the Random interface, however, it is still secure against shoulder surfing attack since there was no successful complete login for attacks in both Interfaces. For the accuracy, we found that our system has low accuracy 40% successful complete login, but with an individual number it detected 75% successfully, and this is being because we were used a front camera of the phone instead of special tracking camera within its software.

As a future work, we will improve this system to be more accurate by using eye-tracking tools that help analyze the results. In addition, we will try to increase accuracy and usability by improving the paths.

References

[1] Y. S, R.Sathishkumar, A. L, and A. V, "Counterfeit Shoulder Surfing Attack Using Random Pin," International Journal of Pure Appl. Math., vol. 118, no. 22, pp. 1757–1761, 2018.

[2] M. Khamis, M. Hassib, E. von Zezschwitz, and A. Bulling, "GazeTouchPIN: Protecting Sensitive Data on Mobile Devices using Secure Multimodal Authentication," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 446–450.

[3] Y. Kita and M. Park, "Proposal and its Evaluation of a Shoulder-Surfing Attack Resistant Authentication Method : Secret Tap with Double Shift," International Journal of Cyber-Security Digit. Forensics, vol. 2, no. 1, pp. 48–55, 2013.

[4] O. Kasat, U. Bhadade, N. Trivedi, "Study and Analysis of Shoulder-Surfing Methods ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN: 2456-3307, vol. 1, no. 6, pp. 256-261, November-December 2015.

[5] H. R. Chennamma and X. Yaun, "A Survey on Eye-Gaze Tracking Techniques," Indian Journal of Computer Science an Engineering (IJCSE), vol. 4, no.5, pp. 389-393, 2013.

[6] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget, "Design and evaluation of a shoulder-surfing resistant graphical password scheme," in Proceedings of the working conference on Advanced visual interfaces - AVI '06, 2006, pp. 177–184.

[7] H. Gao, X. Liu, S. Wang, H. Liu, and R. Dai, "Design and analysis of a graphical password scheme," 2009 4th International Conference Innov. Comput. Inf. Control. ICICIC 2009, pp. 675–678, 2009.

[8] P. Nand, P. K. Singh, J. Aneja, and Y. Dhingra, "Prevention of shoulder surfing attack using randomized square matrix virtual keyboard," 2015 International Conference Adv. Comput. Eng. Appl., pp. 916–920, 2015.

[9] A. Maiti, M. Jadliwala, and C. Weber, "Preventing shoulder surfing using randomized augmented reality keyboards," 2017 IEEE International Conference Pervasive Comput. Commun. Work. PerCom Work. 2017, pp. 630–635, 2017.

[10] A. Jesudoss, N. P. Subramaniam, "A Survey on Authentication attacks and countermeasures in A Distributed Environment," Indian Journal of Computer Science an Engineering (IJCSE), vol. 5, no.2, pp. 71-77, 2014.

[11] A.A. Gawande and G. Nathaney, "A Survey on Gaze Estimation Techniques in Smartphone," International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 4, pp. 2647-2650, 2017.

[12] M. Mehrube and N. Vuong, "Real-Time Eye Tracking for Password Authentication," International Conference on Consumer Electronics (ICCE), pp. 1-4, 2018.

[13] J. Weaver, K. Mock, and B. Hoanca, "Gaze-based password authentication through automatic clustering of gaze points," 2011 IEEE International Conference Syst. Man, Cybern., pp. 2749–2754, 2011.

[14] A. Bulling, F. Alt, and A. Schmidt, "Increasing the Security of Gaze-Based Cued-Recall Graphical Passwords Using Saliency Masks," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3011–3020, 2012.

[15] P. Sorate and G. J. Chhajed, "Survey Paper on Eye Gaze Tracking Methods and Techniques," International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 6, pp. 5612-5616, 2017.

[16] D. M. Ibrahim and S. Ambreen, "Gaze Touch Cross PIN: Secure Multimodal Authentication using Gaze and Touch PIN," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 1, pp. 777-781, 2019. doi: 10.35940/ijeat.A1381.109119.

[17] S. Schenk, P. Tiefenbacher, and G. Rigoll, "SPOCK: A smooth pursuit oculomotor control kit," in Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 2681–2687.

[18] V. Rajanna, S. Polsley, P. Taele, and T. Hammond, "A Gaze Gesture-Based User Authentication System to Counter Shoulder-Surfing Attacks," in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017.

[19] V. Rajanna, A. H. Malla, R. A. Bhagat, and T. Hammond, "DyGazePass: A gaze gesture-based dynamic authentication system to counter shoulder surfing and video analysis attacks," 2018 IEEE 4th International Conference Identity, Secur. Behav. Anal. ISBA 2018, vol. 2018, pp. 1–8, 2018.

Controlling High PAPR in Vehicular OFDM-MIMO using Downlink Optimization Model under DCT Transform

Ahmed Ali[1]
Department of Computer Science
College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University
Al-kharj 11942, Saudi Arabia
Higher Future Institute for Specialized Technological Studies, Cairo, Egypt

Esraa Eldesouky[2]
Department of Computer Science
College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-kharj 11942
Saudi Arabia
Department of Computer Science, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

***Abstract*—The persisting challenges of the radio channel in vehicular networks entail the use of multi-antennas which is known as Multiple-Input Multiple-Output (MIMO). In order to obtain an efficient multi-user MIMO system, the power of the radio frequency (RF) components should be optimized. Necessarily practical solutions are essential to lower the vehicular nodes' complexity and support a robust Orthogonal Frequency Division Multiplexing (OFDM) discipline with a simplified equalization at the receiver. In this paper, the pre-coding Zadoff-Chu Sequence (ZCS) is employed along with the Discrete Cosine Transform (DCT) to control and optimize the high peak power. It intends the transmission over multi-user MIMO downlink vehicular channels. At last, the convex optimization is utilized to guarantee the peak-to-average power ratio (PAPR) minimization. Simulation results have shown that the proposed model can lessen the high PAPR compared to the least-square pre-coding. At the same time, it proved its effectiveness and accuracy as it enhances the transmission quality over multi-user MIMO-OFDM downlink vehicular channel.**

***Keywords*—Zadoff-Chu Sequence; Discrete Cosine Transform (DCT); Peak-to-Average Power Ratio (PAPR); Vehicular Networks**

I. Introduction

In vehicular networks, the Dedicated Short Range Communication (DSRC) protocol supports the vehicle-to-roadside and vehicle-to-vehicle communications which enables numerous Intelligent Transportation Systems (ITS) applications [1, 2]. However, the highly vehicular mobility affects the vehicular channel due to the frequency-selective multi-path propagation [3-7]. Thus, the variety of the available vehicular applications requires the use of efficient technologies such as OFDM-MIMO. This technology can boost the spectral efficiency by deploying several antennas at both the transmitter and the receiver [8].

The MIMO wireless communication technology allows the synchronization of large number of antennas at both sides while serving various vehicular users concurrently [9]. Among the privileges of employing MIMO are the efficient utilization of the RF, enhancement of the operational power consumption, and achievement of low complexity multi-user schemes. Multi-user MIMO communications is required in the vehicular networks due to the presence of huge node populations [9-13].

The OFDM vehicular channels usually suffer from the frequency selective fading, scattering, and high PAPR. As a result, the use of the linear power amplifiers is necessary to avoid out-of-band radiation and signal distortion [14, 15]. The emergence of curtailment of the high PAPR in the multi-user OFDM-MIMO can achieve better signal qualities and guarantee adequate data transmission.

Few research works have focus on the vehicular networks and their challenges [16-19]. Accordingly, the gap between MIMO and vehicular networks should be narrowed by offering an efficient channel modeling and estimation along with timely signal processing. In this paper, a novel vehicular model is proposed to control and optimize the PAPR resulted from applying the OFDM system on the transmitted signal. This model implements the DCT along with the ZCS pre-coding to enhance the signal quality in a vehicular downlink channel. The use of the ZCS combined with the DCT-OFDM can lower the complexity and cost of the high-power amplifier in a multi-user MIMO vehicular system.

This paper is organized as follows. Section 2 summarizes the literature review related to MIMO and vehicular networks. In Section 3, MIMO vehicular network is explained including the OFDM components and the ZCS pre-coding. However, Section 4 introduces the proposed IDCT MIMO-OFDM model. The channel downlink optimization is verified in Section 5. Section 6 introduces the experimental results and analysis for the proposed model. Finally, the paper is concluded in Section 7.

II. Related Works

Under the single-antenna wireless communication systems, several PAPR reduction solutions have been proposed [20-24]. Nevertheless, the PAPR reduction problem concerning the multi-user MIMO systems are described based on Tomlinson-Harashima pre-coding, lattice pre-coding and joint pre-coding [25-27].

Christoph and Erik [27] proposed a PAPR reduction process by exploiting the excess of the transmitted antennas at the base station. The OFDM downlink transmission scheme is investigated for large-scale multi-user MIMO wireless systems to avoid expensive and inefficient power RF components at the base station. Yet, this work enlarges the processing complexity which makes it incompatible for the vehicular IEEE 802.11p channels [28]. The other two schemes [25, 26] based on Tomlinson-Harashima pre-coding require specialized signal processing algorithms at both terminals of the wireless link. This drawback prevents their implementation in wireless MIMO-OFDM systems, such as IEEE802.11n and 3GPP LTE. Similarly, they are inappropriate to vehicular networks as they increase the signal delay and latency.

An efficient detection ordering scheme in which the MIMO transmission is based on Zadoff-Chu Transform (ZCT) OFDM along with the root-raised cosine pulse shaping is presented in [29]. In [30], a joint pre-coding scheme is proposed for indoor downlink multi-user MIMO systems in which the terminals are able to receive spatially multiplexed symbol streams. Another spatial multiplexing multi-user MIMO downlink is proposed by Melvin et. al [31]. At this point, each user can receive multiple data sub-channels simultaneously using the same time and spectral resources.

Multi-user OFDM-MIMO communications are necessary to satisfy the requirements of any wireless communication system (i.e., vehicular networks) such as throughput and Quality-of-Service (QoS) demands. Since, vehicular networks and MIMO systems have been studied in several researches independently. Integrating such technologies can result in various benefits including significant improvement in the range and reliability of the vehicular communication via beamforming. Furthermore, the network throughput can be enhanced when using MIMO technology due to the efficient management of the multiuser interference.

IEEE 802.11p standard does not come with any specification for a vehicular MIMO version. Accordingly, researchers pay more attention to capture the benefits of using multiple antennas in vehicular environment [32]. Thus, the vast vehicular applications require a joint optimization for the medium access control (MAC), network and physical layers.

Sending dependent signals through multi-paths is considered the main clue for MIMO technology. Though, several replicas of the data symbol can be obtained at the receiver leading to a reliable diversity gain. Therefore, multiple parallel subchannels can be obtained (i.e., spatial dimension). Consequently, the link capacity can be enhanced by transmitting independent replicas of data symbols in parallel manner through orthogonal subchannels.

III. MIMO Vehicular Networks

With the increasing number of vehicular activities and drivers' demands, more adequate propagating channels should be guaranteed. The key idea is to enhance transportation safety, efficiency and stillness through the well-use of the short connectivity time among vehicular nodes. The availability of efficient vehicular channels can ensure the arrival of the warning alerts and safety messages between vehicles and roadside infrastructures. Consequently, traffic congestion, fatalities or any accidental injuries can be avoided.

Sending dependent signals through multi-paths is considered the main clue for MIMO technology. Though, several replicas of the data symbol can be obtained at the receiver leading to a reliable diversity gain. Therefore, multiple parallel sub-channels can be obtained (i.e., spatial dimension). Consequently, the link capacity can be enhanced by transmitting independent replicas of data symbols in parallel manner through orthogonal sub-channels.

A. Peak-to-Average Power Ratio in OFDM Model

The OFDM signal can suffer from high PAPR which should be reduced in order to improve the transmission quality and lower the system complexity. The resulted high PAPR can lead to nonlinear impacts on the wireless OFDM signal as well as distortion in the electrical equipment. Thus, an efficient OFDM channel with controlled PAPR should be suggested to cope with the fast-changeable vehicular channel.

According to [33], PAPR is the relationship between the maximum power of a transmitted OFDM sample and its average power as shown in equation (1).

$$PAPR = 10\log_{10}\frac{Power_{peak}}{Power_{average}}\ (dB) \tag{1}$$

Where $Power_{peak}$ and $Power_{average}$ are the peak and average power of a transmitted OFDM symbol, respectively. They can be estimated using equations (2) and (3).

$$Power_{peak} = \max_{0\le t\le LT_s}\left|x(t)\right|^2, \tag{2}$$

$$Power_{average} = \frac{1}{LT_s}\int_0^{T_s}\left|x(t)\right|^2, \tag{3}$$

Where L is the number of the modulated subcarriers, T_s is the sampling period, and $x(t)$ is the OFDM input signal over time t.

Hence, a high PAPR appears when the subcarriers of a given OFDM symbol are out of phase with each other (i.e., the use of a large number of subcarriers). Actually, it is one of the main challenges in any multi-carrier system using the OFDM MIMO.

B. Zadoff-Chu Sequence Pre-coding

It is necessary to reduce and control the PAPR of the transmit signal in the proposed vehicular OFDM-MIMO system through a pre-coding techniques. In the proposed model, the ZCS pre-coding technique is deployed to reshape the IDCT matrix row and reform the constellation symbols before the channel propagation step [34]. The main advantages of ZCS is its ability to provide constant amplitude for the output signal with optimum correlation properties as well as reduce the cost and complexity of output signal [35, 36]. Thus, the peak power of the transmitted signal can be under control. A ZCS z of length L_s can be mathematically defined as:

$$z(n)=\begin{cases} e^{\frac{j2r\pi}{L_s}\left(\frac{k^2}{2}+qk\right)} & for \quad L_s \in even, \\ e^{\frac{j2r\pi}{L_s}\left(\frac{k(k+1)}{2}+qk\right)} & for \quad L_s \in odd \end{cases} \tag{4}$$

Where $k=1,2,...,L_s$, q is any integer, r is any integer relatively prime to L_s ,and j is the imaginary number $\sqrt{-1}$. Thus, the kernel of the ZCS acts as a rowwise pre-coding of matrix Z of size W for $W=L_s^2$, as shown in (5) and it is applied to each symbol to reduce PAPR.

$$Z_{m,n}=\begin{bmatrix} z_{00} & z_{01} & \cdots & z_{0(L_s-1)} \\ z_{10} & z_{11} & \cdots & z_{1(L_s-1)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{(L_s-1)0} & z_{(L_s-1)1} & \cdots & z_{(L_s-1)(L_s-1)} \end{bmatrix} \tag{5}$$

Where, $z_{m,n}$ is the matrix element in the m^{th} row and n^{th} column.

Fig. 1 illustrates the transmitted signal over the vehicular channel in which, the input OFDM signal x is passed through the ZCS pre-coding block. Then, the ZCS pre-coding matrix is applied to obtain $A^{'}=\lfloor A_0^{'}, A_1^{'},..., A_{L_s-1}^{'} \rfloor$ as in equation (6):

$$A^{'}=ZA, \tag{6}$$

Given that, Z is the pre-coding matrix of size W whereas the new vector $A_m^{'}$ is obtained by using the following equation:

$$A_m^{'}=\sum_{n=0}^{L_s-1} Z_{m,n}A_n \qquad for \quad m=0,1,...,L_{s-1} \tag{7}$$

The ZCS-OFDM complex baseband signal with L_s subcarriers is defined as follows:

$$A_m=\frac{1}{\sqrt{L_s}}\sum_{m=0}^{L_s-1} A_n^{'}.e^{j2\pi\frac{m}{L_s}n} \quad for \quad m=0,1,...,L_{s-1} \tag{8}$$

Every transmitted data block is then propagated over the vehicular channel in its way to reach the receiver side.

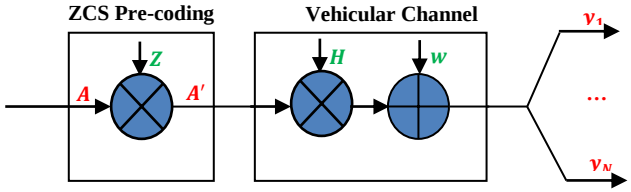


Fig. 1. Transmitted Signal over the Vehicular Channel.

IV. Proposed Inverse Discrete Cosine Transform OFDM-MIMO Model

MIMO technique was firstly introduced by Goldsmith et al. in [37], which addresses the multi-antennas IEEE 802.11 wireless networks. Mainly, it supports the multi-user case using similar capacity. In other words, a MIMO system with M transmit antennas and N receive antennas can have a linear capacity of $\min(M,N)$. In general, the multi-user MIMO downlink technology is applicable to several wireless networks with multi-antenna infrastructures such as base stations or access points. In the proposed model, a DCT transformation mechanism is considered based on a multi-user OFDM-MIMO using the ZCS pre-coding over vehicular downlink channels in order to decrease the resulted PAPR and enhance the network reliability.

The proposed model considers a roadside unit infrastructure with a large number of transmit antennas M that serve multi-user receivers N where $M \geq N$. The receiver antennas are treated as independent terminals (i.e., vehicular nodes) whereas each terminal has its own antenna. The vehicular MIMO channel has the form of matrix H with $N \times M$ dimension. $[H]_{i,j}$ is the channel gain (i.e., transfer function) from the j^{th} transmitter antenna to the i^{th} receiver antenna. The transmitted signal x is defined to be a complex vector that collects all the transmitted MIMO-OFDM symbols such that $x=[x_0, x_1, ..., x_{M-1}]^T$. Though, x_j is the transmitted symbol from the j^{th} transmit antenna; where $j=0,1,...,M-1$. The input-output relationship of the channel matrix H is computed as in equation 9.

$$y=Hx+w, \tag{9}$$

Where $w=[w_1, w_2, ..., w_N]^T$ is an one-dimensional vector of Gaussian additive noise; and $y=[y_1, y_2, ..., y_N]^T$ is the vector of the received data. The elements of the noise vector follow complex zero mean Gaussian distribution with variance σ^2. Since, this model deals with multi-user vehicular nodes, the N receivers are assigned to different users. Thus, each vehicular user has access to only one element of vector y.

This model considers a case that the roadside unit attempts to transmit replicas of the digital signal to multiple vehicular nodes (i.e., spatial diversity mode). For example, broadcasting a safety message, traffic status, or even an advertisement. According to Fig. 2, a roadside unit (RS) commences a broadcast action whereas a transmitted digital signal x is ready for propagation. Firstly, the replicas of the transmitted signal are created using the diversity processor block. Then, they are converted from serial to parallel form using the serial-to-parallel (S/P) converter in order to be sent through the roadside antennas. At this moment, the resulted sequence is defined as $A=[A_0, A_1, ..., A_{L-1}]$ based on the one-to-one mapping and the number of the received antennas N.

The DCT/IDCT transform is employed in this model [22, 36, 38]. Similar to other transformation methods, the DCT relies on a sum of cosine function that is fluctuating at different frequencies to define a finite sequence of data points. The cosine functions can particularly express the choice of the boundary conditions. Moreover, DCT transform allows a digital signal to be propagated efficiently in the frequency domain. Each OFDM-MIMO signal (i.e., A) is passed through the inverse DCT (IDCT) block to generate the complex baseband OFDM signal with subcarriers according to Equations (10) and (11).

$$a(k)=\begin{cases} s(k)\sum_{i=0}^{L-1} A(n)\cos\left[\frac{\pi(2l+1)k}{2L}\right] & for \quad 0\le k\le L, \\ 0 & Otherwise \end{cases} \tag{10}$$

$$s(k)=\begin{cases} \frac{1}{\sqrt{L}}\ for & for \quad k=0, \\ \frac{2}{\sqrt{L}} & for \quad k\neq 0 \end{cases} \tag{11}$$

Where L is the length of the OFDM symbols to get the time domain samples $a=[a_1,a_2,...,a_N]^T$.

After The IDCT transform, the Cyclic Prefix (CP) inserted to the samples. Then, the baseband is converted to passband signal using the Up-Frequency Conversion (UFC). Afterwards, the signal is pre-coded with the ZCS transform and passed over the IEEE 802.11p vehicular channel [28].

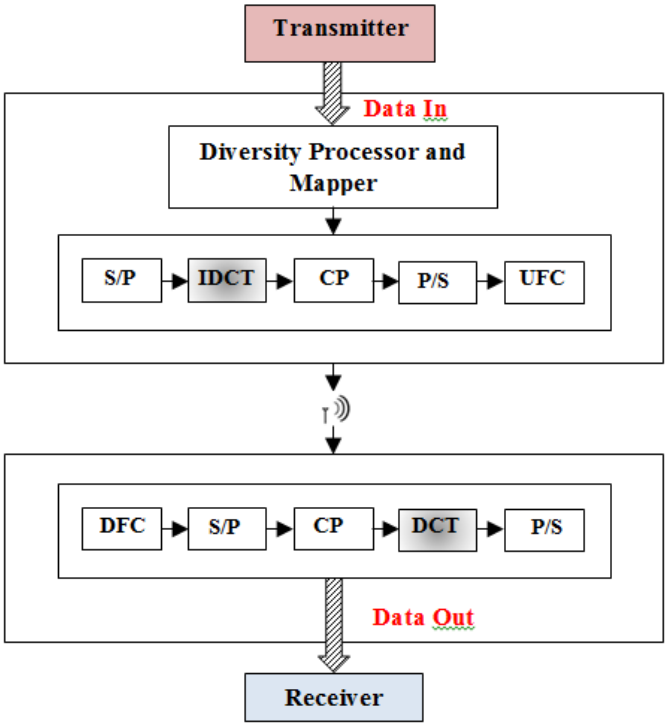


Fig. 2. The Proposed OFDM-MIMO Downlink Model.

At the receiver antennas, the passband signal is reconverted to baseband signal using the Down-Frequency Conversion (DFC).The DCT transform is applied to the received OFDM symbols. Finally, the signal is received through the vehicular users' antennas. The basic main steps of the proposed model can be summarized as follows:

Step1: An RSU with antennas transmits a digital signal to multi-vehicular users whereas each with single independent antenna.

Step2: The sent signal is converted to parallel subcarriers, transformed using the IDCT block, and cyclic prefix is added to avoid the intersymbol interference

Step3: The transmitted signal is then transformed to passband using the UFC.

Step4: ZCS pre-coding is applied to the transmitted signal at the roadside unit in order to control the PAPR.

Step5: The signal is propagated over the vehicular channel, and exposed to some noise.

Step6: At the receiver side, the baseband signal is obtained by using the DFC followed by the removal of the cyclic prefix and the signal is re-converted to serial form as to be received by the users' antennas.

This combinational model aims to employ the ZCS pre-coding and the DCT/IDCT transform to reduce the resulted PAPR in MIMO-OFDM downlink model.

V. Channel Downlink Optimization

As aforementioned, the proposed model aims to control the dynamic range while transmitting the input signal A. This signal is exposed to a channel impulse response H, as illustrated in equation (9). Given a constraint that the number of the transmit antenna is larger than or equal to the receive antenna. The ZCS pre-coding is employed to the input signal A in an effort to reduce the PAPR and get A'. Hence, to avoid the multi-user interference, A' must satisfy the pre-coding constraint $A'=HA$, which ensures $y=A+w$. This means that the pre-coding must be applied at the transmitter side (i.e., roadside unit).

In general, the primary objective is to find an appropriate A' which ensures a small dynamic range and optimized PAPR. This can be obtained by employing the following objective function:

$$(P-Dyn-A') \quad \min \quad \alpha - PAPR_{A'}$$

$$\text{Subject to } A'=HA \tag{12}$$

$$\text{And} \quad PAPR_{A'} \le |A'_m| \le \alpha \qquad m=1,...,B$$

Where $(P-Dyn-A')$ is the optimization objective that aims to control the PAPR values of A', $PAPR_{A'}$ is the actual PAPR value of the OFDM symbol, B is the number of the transmitted OFDM symbols and α is the desired PAPR

threshold. This means that $(P-Dyn-A^{'})$ exhibits optimized lower PAPR than the transmit vector $A^{'}$ resulted from the ZCS pre-coding block. However, satisfying the inequality $PAPR_{A^{'}} \le |A_m^{'}| \le \alpha$ seems to be difficult. Then, $(P-Dyn-A^{'})$ needs to be relaxed by only considering the upper bound constraint $|A_m^{'}| \le \alpha$.

Thus, the convex optimization problem can be reformulated as follows:

$$(P-Dyn-R) \quad \underline{\textbf{\textit{minimize}}} \ \max \left\| A_m^{'} \right\|_B \quad for \quad m=1,...,B, \tag{13}$$

Subject to

$$A^{'} = HA$$

At this moment, the $(P-Dyn-R)$ is expected to minimize the magnitude of the largest entry in $A^{'}$ which consequently ensures lower PAPR values. This single optimization problem is solved using the convex optimization which mainly studies the problem of minimizing the convex function over convex sets. Thus, the convex optimization function can be re-defined as follows:

$$(P-Dyn-a^{'}) \quad \min_{a_1^{'},a_2^{'},...,a_B^{'}} \quad \max\left\{\left\|a_1^{'}\right\|,\left\|a_2^{'}\right\|,...,\left\|a_B^{'}\right\|\right\}, \tag{14}$$

Subject to

$$a_b^{'} = H_b a_b \quad for \quad b=1,2,...,B$$

Where, the vector $a_b^{'}$ corresponds to the pre-coded OFDM symbols that are transmitted from each antenna to be minimized.

Thereafter, the result PAPR values of the optimized symbols $(P-Dyn-R)$ and the PAPR of the ZCS pre-coding should satisfy the following equality:

$$(P-Dyn-R) = \frac{N\left\|OP_A\right\|_B^2}{\left\|OP_A\right\|_2^2} \le \frac{N\left\|A^{'}\right\|_B^2}{\left\|A^{'}\right\|_2^2} = PAPR_{ZCS} \tag{15}$$

This implies that the PAPR of the convex optimization cannot exceed the PAPR values of the ZCS pre-coding technique for all the transmitted OFDM symbols.

According to proposition 1 in [27], which states that:

"If H has the full column rank and $1<M<N$ -given that M , N are integers- then, the solution $OP_{A^{'}}$ of the $(P-Dyn-R)$ can have $N-M+1$ entries with a magnitude equal to $\left\|OP_{A^{'}}\right\|_B$."

The upper bound of the $PAPR_{(P-Dyn-R)}$ is given by:

$$(P-Dyn-R) = \frac{N\left\|OP_A\right\|_B^2}{\left\|OP_A\right\|_2^2} \le \frac{N}{N-M+1} \tag{16}$$

Such that:

$$\left\|OP_A\right\|_2^2 = \sum_{\mu}\left\|OP_A\right\|_B^2 + \sum_{i\in\mu}\left|\left[OP_{A^{'}}\right]_i\right|^2$$

$$\ge \sum_{\mu}\left\|OP_{A^{'}}\right\|_B^2 = (N-M+1)\left\|OP_{A^{'}}\right\|_B^2$$

Where μ is the indices for the $N-M+1$ entries of $a^{'}$ for which $\left|\left[OP_{A^{'}}\right]_i\right| = \left\|OP_{A^{'}}\right\|_B$. As a result, the $PAPR_{(P-Dyn-R)}$ can approach to 1 for a predefined number of antennas M and N. This means that the $PAPR_{(P-Dyn-R)}$ can eliminate the multi-user interference while keeping constant envelope signal. As a result, these transmitting vectors can succeed to have low PAPR.

VI. Simulation Results and Analysis

The effectiveness and accuracy of the proposed model is demonstrated via simulation using MATLAB. As previously mentioned, this model considers the PAPR reduction problem in OFDM-MIMO vehicular networks. It is a downlink scheme that mimics the transformation of a digital signal from a multi-antennas roadside unit to independent multi-users. The DCT transformation is used instead of the Fast Fourier Transform (FFT) along with a ZCS pre-coding in order to reduce the PAPR and system complexity.

A. Simulation Setup

According to Studer and Larsson [27], the proposed model assumes a roadside unit with 100 antennas installed and 10 vehicular receivers. Nevertheless, the number of receivers is increased to reach 30 end-users. It is known that the IEEE802.11p amendments are basically based on the OFDM scheme [28]. Thus, an OFDM system with 64 subcarriers is employed [39], a bandwidth of 10 MHz, and a transmission data rate of 27 Mbps as specified in the IEEE 802.11p. The cyclic prefix is of the length 16 along with 480 m propagation distance whereas the excess delay is given by 6 µs.

In the proposed model, each user is assigned to a coded transmission with 216 information bits. These encoded bits are mapped using 16-QAM constellation as well as a convolutional code with coding rate of 1/2 and constraint length of 7 [40].

The vehicular channel is considered as a frequency-selective multipath channel. Hence, the pre-coded signal resulted from the ZCS pre-coding block is transmitted in the form of a tap-delay line with 4 taps. At the end, the transmitted signal is detected at the receiver side (i.e., independent vehicular users) which performs a demodulation process and applies the DCT transform. The proposed model is applied on the transmitted OFDM-MIMO-symbols in order to reduce the PAPR ratio. Table I lists the parameters values used in the conducted simulations.

TABLE. I. SIMULATION PARAMETERS AND VALUES

Items	Values
Number of Subcarriers	64
Data Subcarriers	52
Cyclic Prefix	16
Modulation	16-QAM
Number of taps	4
Transmit antennas	100
Receive antennas	10, 20, 30
Transmission Data Rate	27 Mbps
Excess Delay	6 µs
Coding rate	½
Constraint Length	7

To evaluate the performance of the proposed model, the Complementary Cumulative Distribution Function (CCDF) is used. The CCDF metric computes the probability of the OFDM signal exceeding the predefined PAPR threshold within the OFDM symbols and it is given by:

$$CCDF[PAPR_i] = Prob[PAPR_i > PAPR_0] \quad (17)$$

Where, $PAPR_0$ is the PAPR threshold and $PAPR_i$ is the PAPR value of the i^{th} OFDM symbol in the transmitted signal.

In addition, the average symbol-error rate is measured in order to estimate the performance bits error rate across the vehicular users. A symbol is considered to have an error if the received symbols are different from the transmitted symbols. As the result, the transmitted information bits are changed. Then, the average symbol-error rate (AVG_{SER}) is given below:

$$AVG_{SER} = \frac{\sum_{i=1}^{V_N} SER_i}{V_N} \quad (18)$$

Where, SER_i is the symbol-error rate for the receiver overall the vehicular receivers V_N .

B. Results Analysis

In this simulation, the proposed model is compared with the Least-Square pre-coding technique [41] and the classical OFDM system without using pre-coding techniques. The experiments are conducted in a downlink scheme whereas various receiver populations are used.

Fig. 3, 4 and 5 show the CCDF of the PAPR distribution where 16-QAM is applied to this model.

In Fig. 3, the PAPR values illustrate the use of various pre-coding techniques using 10 independent receiving antennas (i.e., vehicular users). The target PAPR of the proposed model is obviously decreased compared to the other two techniques due to applying of the ZCS pre-coding technique along with the DCT transform. The maximum reached PAPR value for the proposed model is 6.22 dB. However, the conventional OFDM (i.e., without a PAPR reduction mechanism) and the Least-Square pre-coding [42] are 10.35 dB and 12.51 dB, respectively.

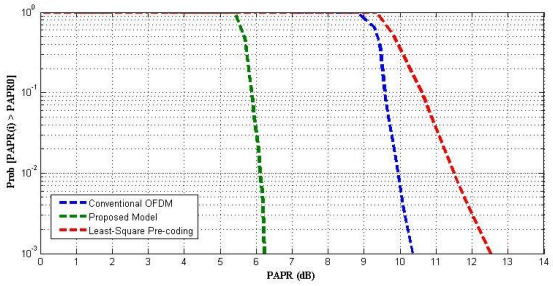
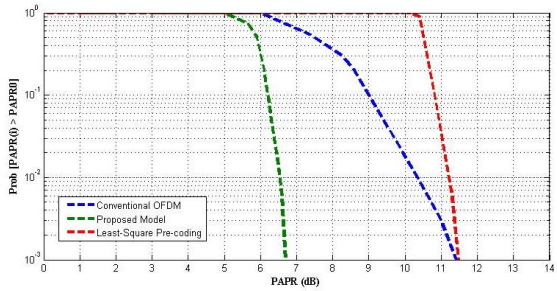
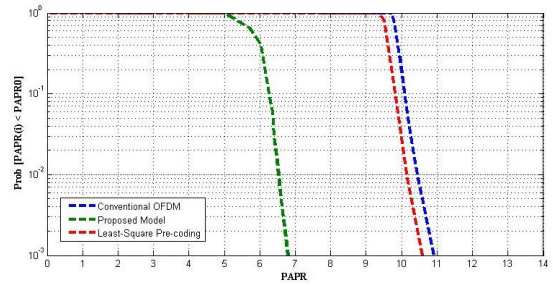


Fig. 3. PAPR Performance using DCT Transform in Vehicular Networks.



(a)



(b)

Fig. 4. PAPR CCDF using the ZCT Pre-Coding in a MIMO Downlink Vehicular Channel.

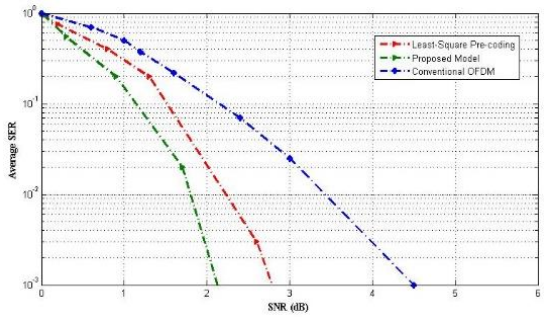


Fig. 5. Average SER using 16-QAM and 10 Receive Vehicular Antennas.

Nevertheless, the PAPR of the tested vehicular channel is effectively decreased when the number of receivers is increased to reach 20 and 30 vehicular users as respectively shown in Fig. 4(a) and (b). The Least-Square pre-coding is affected by the antennas population so that if the number of the network nodes increases, the associated PAPR increased either. The maximum PAPR when using the Least-Square Pre-coding are 11.48 and 10.59 dB, respectively. Nevertheless, the PAPR for the proposed model is reduced by 5 dB when compared to the other two schemes.

According to the simulation results, the proposed model (i.e., ZCS pre-coding plus DCT transform) succeeds to reduce the high PAPR resulted due to the use of OFDM- MIMO system in the vehicular networks. The received time-domain signal has a significant smaller PAPR than of the Least-Square (i.e., up to 6 dB).

The Symbol-Error-Rate (SER) is the probability of receiving a symbol in error. It is usually calculated by comparing the ratio of symbols in error to the total number of bits. Hence, the SER performance is studied for the proposed pre-coding downlink model, Least-Square pre-coding and the conventional OFDM. In this paper, the SER has been estimated with respect to different Signal-to-Noise Ratio (SNR) to determine the information bits which are affected during transmission (i.e., symbols that are not correctly received).

Fig. 5 shows the SER performance of the three models. The horizontal axis indicates the SNR in dB and the vertical axis represents the average SER. It demonstrates that the increase in the SNR values lead to lower average SER for the three schemes using 16-QAM modulation and 10 receiver antennas. Moreover, it has been observed that the proposed model provides the best SER compared to the Least-Square pre-coding.

According to the simulation results, the proposed model that relies on ZCS pre-coding and DCT transform in an OFDM-MIMO vehicular channel succeeds to significantly reduce both the PAPR and the SNR for the transmitted symbol.

VII. Conclusions

Since the IEEE 802.11p standard is mainly designed for low complexity receivers in vehicular networks, a degraded performance can occur due to the high PAPR resulted from using the OFDM system. Hence, a low packet error rate is crucial for specifically the safety-related scenarios. The use of multiple antennas at both the transmitter and the receiver may improve the network throughput but may result in a multi-user interference problem. This can be resolved using a multi-user OFDM -MIMO system. In this paper, a joint multi-user OFDM-MIMO model based on the ZCS pre-coding combined with DCT/IDCT transform is proposed. This model aims to control and reduce the high PAPR over the vehicular channel in a downlink vehicle-to-roadside environment. According to the conducted simulations, the results show that the proposed model improves the vehicular channel performance by decreasing the PAPR. This model achieves observable lower PAPR compared to the Least-Square pre-coding and the conventional OFDM. Using 10, 20 and 30 receive independent antennas; the PAPR for the proposed model is 6.22 dB, 6.69 dB, and 6.79 dB, respectively. Accordingly, this model improves the transmission quality over multi-user OFDM-MIMO downlink vehicular channel.

Although, vehicular networks acquire numerous privileges of the presence and utilization of multiple antennas, challenges regarding PAPR in MIMO vehicular networks still exist. A future pathway is the proper analysis of the imperfect Channel State Information (CSI) effect using the ZCS pre-coding for vehicular communication. Efficient metaheuristic optimization algorithms can be implemented in order to powerfully minimize the PAPR values for the input signal. Finally, the nonstationary nature of vehicular topology requires analyzing the channel estimation technique recursively in order to track the impulse response of the MIMO channel.

References

[1] P. Mutalik, and V. C. Patil, "A survey on vehicular ad-hoc network [VANET's] protocols for improving safety in urban cities." International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bangalore, pp. 840-845, August 2017.

[2] E. Eldesouky, W. Elashmawi, L. Renfa, and W. Dong, “A Hybrid Cooperative Model for Bandwidth Utilization in Vehicular Ad hoc Networks Based on Game Theory,” International Journal of Control Automation, vol. 7, no. 12, pp. 177-188, 2014.

[3] B. Bloessl, F. Klingler, F. Missbrenner, and C. Sommer, "A systematic study on the impact of noise and OFDM interference on IEEE 802.11p.", IEEE Vehicular Networking Conference (VNC), Torino, pp. 287-290, Novemeber 2017.

[4] A. Ali, W. Dong, L. Renfa et al., “ISI and PAPR Immune IEEE 802.11 p Channels Based on Single-Carrier Frequency Domain Equalizer,” KSII Transactions on Internet Information Systems, vol. 10, no. 11, pp. 5513-5529, 2016.

[5] U.A. Khan, and S.S. Lee, " Multi-Layer Problems and Solutions in VANETs: A Review," Electronics, vol. 8, no. 2, pp. 1-19, 2019.

[6] L. Liang, H. Peng, G.Y. Li, and X. Shen, " Vehicular communications: A physical layer perspective,", IEEE Transactions on Vehicular Technology, vol. 66, no.12, pp. 10647-10659, 2017.

[7] L. Cheng, W. Viriyasitavat, M. Boban, and H.M. Tsai, “Comparison of radio frequency and visible light propagation channels for vehicular communications,” IEEE Access, vol. 6, pp. 2634-2644, 2018.

[8] N. Adhikari, "Antenna selection and MIMO capacity estimation for vehicular communication systems" The University of North Dakota, Theses and Dissertations, 2014.

[9] S. T. Abraha, D. F. Castellana, X. Liang, A. Ng'oma, and A. Kobyakov, "Experimental Study of Distributed Massive MIMO (DM-MIMO) in In-building Fiber-Wireless Networks.", Optical Fiber Communications Conference and Exposition (OFC), San Diego, pp. 1-3, March 2018.

[10] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser MIMO systems,” IEEE Transactions on Communications, vol. 61, no. 4, pp. 1436-1449, 2013.

[11] F. Rusek, D. Persson, B. K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” IEEE Signal Processing Magazine, vol. 30, no. 1, pp. 40-60, 2013.

[12] J. Zhang, Y. Huang, J. Wang et al., “Per-antenna constant envelope precoding and antenna subset selection: A geometric approach,” IEEE Transactions on Signal Processing, vol. 64, no. 23, pp. 6089-6104, 2016.

[13] A. El-Keyi, T. ElBatt, F. Bai, and C. Saraydar, "MIMO VANETs: Research challenges and opportunities.", International Conference on Computing, Networking and Communications (ICNC), Maui, pp. 670-676, January 2012.

[14] A. Chadha, N. Satam, and B. Ballal, “Orthogonal frequency division multiplexing and its applications,” International Journal of Science and Research (IJSR), vol. 2, no. 1, 2013.

[15] S. Bachir, B. Koussa, C. Perrine, C. Duvanaud, and R. Vauzelle, "Peak power reduction for OFDM systems in vehicular wireless communications context.", Proceedings of the International Conference on ELECTRONICS, COMPUTERS and ARTIFICIAL INTELLIGENCE - ECAI-2013, Pitesti, pp. 1-9, June 2013.

[16] T. Abbas, J. Nuckelt, T. Kürner, T. Zemen, C.F. Mecklenbräuker, and F. Tufvesson, " Simulation and measurement-based vehicle-to-vehicle channel characterization: Accuracy and constraint analysis," IEEE Transactions on Antennas and Propagation, vol. 63, no. 7, pp. 3208-3218, 2015.

[17] A. Attia, A. A. ElMoslimany, A. El-Keyi, T. ElBatt, F. Bai, and C. Saraydar, “MIMO vehicular networks: Research challenges and opportunities,” Journal of communications, vol. 7, no. 7, pp. 500-513, 2012.

[18] S. Fikar, W. Walzik, and A. L. Scholtz, "Vehicular multi/broadband MIMO antenna for terrestrial communication." IEEE Antennas and Propagation Society International Symposium , Toronto, pp. 1-4, July 2010.

[19] F. Hu, Vehicle-to-vehicle and Vehicle-to-infrastructure Communications: A Technical Approach: CRC Press, 2018.

[20] R. J. Baxley, and G. T. Zhou, "Comparing selected mapping and partial transmit sequence for PAR reduction," IEEE Transactions on Broadcasting, vol. 53, no. 4, pp. 797-803, 2007.

[21] N. I. Miridakis, and D. D. Vergados, "A survey on the successive interference cancellation performance for single-antenna and multiple-antenna OFDM systems," IEEE Communications Surveys Tutorials, vol. 15, no. 1, pp. 312-335, 2013.

[22] S. Abouty, L. Renfa, Z. Fanzi, and F. Mangone, "A novel iterative clipping and filtering technique for PAPR reduction of OFDM signals: system using DCT/IDCT transform," International Journal of Future Generation Communication, vol. 6, no. 1, pp. 1-8, 2013.

[23] A. Gangwar, and M. Bhardwaj, "An overview: Peak to average power ratio in OFDM system & its effect," International Journal of Communication Computer Technologies, vol. 1, no. 2, pp. 22-25, 2012.

[24] G. Wunder, R. F. Fischer, H. Boche et al., "The PAPR problem in OFDM transmission: New directions for a long-lasting problem," IEEE Signal Processing Magazine, vol. 30, no. 6, pp. 130-144, 2013.

[25] S. Umeda, S. Suyama, H. Suzuki, K. Fukawa, "PAPR reduction method for block diagonalization in multiuser MIMO-OFDM systems." IEEE 71st Vehicular Technology Conference, pp. 1-5, May 2010.

[26] C. Siegl, and R. F. Fischer, "Selected basis for PAR reduction in multi-user downlink scenarios using lattice-reduction-aided precoding," EURASIP Journal on advances in signal processing, vol. 2011, no. 1, pp. 17, 2011.

[27] C. Studer, and E. G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," IEEE Journal on Selected Areas in Communications vol. 31, no. 2, pp. 303-313, 2013.

[28] D. Jiang, and L. Delgrossi, "IEEE 802.11 p: Towards an international standard for wireless access in vehicular environments.", VTC Spring 2008 - IEEE Vehicular Technology Conference, Singapore, pp. 2036-2040, May 2008.

[29] V. Sreevani, and K. Suma, "Reduction of PAPR and Efficient detection ordering scheme for MIMO Transmission Systems Using Power Control," International Journal of Engineering Research Applications, vol. 5, no. 12, pp. 31-35, 2015.

[30] Q. Zhao, Y. Fan, and B. Kang, "A joint precoding scheme for indoor downlink multi-user MIMO VLC systems," Optics Communications, vol. 403, pp. 341-346, 2017.

[31] M. C. Lim, M. Ghogho, and D. C. McLernon, "Spatial multiplexing in the multi-user MIMO downlink based on signal-to-leakage ratios.", IEEE Global Telecommunications Conference, Washintgon, pp. 3634-3638, November 2007.

[32] S. Moser, L. Behrendt, and F. Slomka, "MIMO-enabling PHY layer enhancement for vehicular ad-hoc networks," IEEE Wireless Communications and Networking Conference Workshops (WCNCW), New Orleans, pp. 142-147, March 2015.

[33] T. Jiang, and Y. Wu, "An overview: Peak-to-average power ratio reduction techniques for OFDM signals," IEEE Transactions on broadcasting, vol. 54, no. 2, pp. 257-268, 2008.

[34] I. Baig, and V. Jeoti, "On the PAPR reduction in OFDM systems: a novel ZCT precoding based SLM technique," Journal of Engineering Science Technology, vol. 6, no. 3, pp. 357-368, 2011.

[35] I. Baig, and V. Jeoti, "Zadoff–Chu matrix transform precoding-based orthogonal frequency division multiple access uplink systems: a peak-to-average power ratio performance," Arabian Journal for Science Engineering, vol. 38, no. 3, pp. 613-620, 2013.

[36] A. Ali, W. Dong, and L. Renfa, "A Novel ZCS-Clipping Precoding Technique for PAPR Reduction using DCT-OFDM in Vehicular Channels," International Journal of Hybrid Information Technology, vol. 8, no. 5, pp. 11-22, 2015.

[37] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," IEEE Journal on selected areas in Communications, vol. 21, no. 5, pp. 684-702, 2003.

[38] F. Mangone, J. He, J. Tang, J. Xiao, M. Chen, F. Li, and L. Chen, "A PAPR reduction technique using Hadamard transform combined with clipping and filtering based on DCT/IDCT for IM/DD optical OFDM systems," Optical Fiber Technology, vol. 20, no. 4, pp. 384-390, 2014.

[39] C. F. Mecklenbrauker, A. F. Molisch, J. Karedal, F. Tufvesson, A. Paier, L. Bernadó, T. Zemen, O. Klemp, and N. Czink, "Vehicular channel characterization and its implications for wireless system design and performance," Proceedings of the IEEE, vol. 99, no. 7, pp. 1189-1212, 2011.

[40] I. Baig, and V. Jeoti, "PAPR reduction in OFDM systems: Zadoff-Chu matrix transform based pre/post-coding techniques.", 2nd International Conference on Computational Intelligence, Communication Systems and Networks, Liverpool, pp. 373-377, July 2010.

[41] R. Ganesh, and J. J. Kumari, "A survey on channel estimation techniques in mimo-ofdm mobile communication systems," International Journal of Scientific Engineering Research vol. 4, no. 5, pp. 1850-1855, 2013.

[42] D. C. Araujo, T. Maksymyuk, A. L. de Almeida, T. Maciel, J.C. Mota, and M. Jo, "Massive MIMO: survey and future research topics," Iet Communications, vol. 10, no. 15, pp. 1938-1946, 2016.

The Correction of the Grammatical Case Endings Errors in Arabic Language

Chouaib MOUKRIM[1], Abderrahim TRAGHA[2], El Habib BENLAHMER[3]
Faculty of Sciences Ben M'sik, Hassan II University
Laboratory of Information Technologies and Modelling
Casablanca, Morocco

***Abstract*—Syntax plays a key role in natural language processing, but it does not always occupy an important position in applications. The main objective of this article is to solve the problem of the grammatical case ending errors produced by Arabic learners or certain common errors. Arabic can be considered more complex than English or French. He does not have vowels; diacritic signs (vowels) are placed above or below the letters. These diacritic signs are abandoned in most Arabic texts. This induces both grammatical and lexical ambiguities in Arabic. The present paper describes an automatic correction of this type of errors using "Stanford Parser" with an ontology containing the rules of the Arabic language. We segment the text into sentences, then we extract the annotations of each word with the syntactic relations coming from our parser, then we treat the relations obtained with our ontology. Finally, we compare the original sentence with the corrected one in order to detect the error. The implemented system achieved a total detection of about 94%. It is concluded that the approach is clearly promising by observing the results as compared to the limited number of available Arabic grammar checkers.**

***Keywords*—Automatic correction; ontology; syntactic errors; case endings; natural language processing; Arabic**

I. Introduction

Natural language processing (NLP) can appear as a homogenous and unified domain. Its goal is to design a software capable of automatically processing data expressed in a so-called "natural" language. These linguistic data may depend on the case, and are of different types (written texts, written or oral dialogues, etc.). It proposes a set of systems, allowing among others: to extract the information automatically, to summarize automatically long texts, to make an automatic translation of a text, and it also allows the generation of text, the parsing, etc.

Some NLP applications require syntactic analysis of texts apprehended in terms of grammatical relationships, and encoded as functional annotations in syntagmatic trees or dependency trees, which seems to be useful for many tasks. However, these parsers make errors in syntactic relationships and annotations; we can give an example with the following erroneous sentence: "دخل المعلمون المجتهدين"

دخل/VBD

المعلمون/DTNNS

المجتهدين/DTJJ

Root (Root, دخل)

dobj (دخل , المعلمون)

dep (المعلمون, المجتهدين)

(S (VP (VBD دخل) (NP (DTNNS المعلمون) (DTJJ المجتهدين))))

Such as (see Table I)

We noticed that it analyzed the sentence although it is incorrect. The relations it established between the words are also not correct.

The adjective is a noun that qualifies another noun. In the English language, the adjectives come before the noun being qualified. However, in the Arabic language the adjectives come after the noun being qualified. The adjective in Arabic is termed as "نعت" or "صِفَةٌ" (property/quality). The noun being qualified is termed as "مَنعوتٌ".

The following rules apply to the formation of the Arabic qualifier adjective:

- The adjective "النعت" follows in gender (feminine / masculine), in state (defined, indefinite) and in number (singular, dual, and plural) the qualified "المنعوت" to which it refers.
- The adjective follows the qualified in its case endings.

If we want to apply the rules of the adjective, we must first correct the error when the parser starts to annotate the tokens, certainly the correct sentence is:

"دخل المعلمون المجتهدون"

Such as:

المجتهدون /DTJJS

It should be noted that the DTJJS annotation does not exist in Stanford Parser, where DTJJS is an adjective, and plural with the determiner "Al_ال".

The previous error is called an error of grammatical case endings, which typically appears on the last letter of the stem that indicates its syntactic role.

TABLE. I. SYNTAX ANNOTATION OF THE SENTENCE: " دخل المعلمون المجتهدين "

DOBJ	Direct Object
DEP	Unclassified dependent
VBD	Verb, past tense
DTNNS	noun, plural with the determiner "Al" (ال)
DTJJ	adjective with the determiner "Al" (ال)

Despite the improvements made on parsers, there remains the correction of syntactic errors among its main objectives, also for all NLP systems, and especially for the Arabic language, which is the language for which harmful phrase trees are often syntactically specified. The automatic processing of the error is hindered at several levels [1] because of the complexity, the richness of this language, the absence of short vowel (حركات) in most texts, the irregularity of the word order in the structure of sentences, the agglutination, and the problems of morphological parsing.

Many studies focus on natural language processing at various levels. Such as the morphology [2, 3], the syntax [4, 5] and the semantics [6]. Many programs (as, for example, machine translation, question answering, information retrieval, text summarization, etc.) can exploit them. Nevertheless, the necessity of these programs manifests in the relationship between the words that can be even ungrammatical and can, therefore, lead to incorrect results. This absolutely lacks an effective correction system.

The purpose of our work is to use the results of the Stanford Parser and try to improve them in order to correct syntactic case ending errors based on a logical description of the grammatical relations in the ontology.

The rest of this article is structured as follows. We begin with a brief presentation of the state of the art on parsing and related work in Section 2. In Section 3, we define and describe the grammatical case endings in Arabic language. Section 4 explains the syntactic approach that we adopted as well as the chosen parser. Then, in Section 5, we present in detail our methodology; next, we illustrate our work by an example in Section 6. Section 7 presents the evaluation of our approach. Finally, in Section 8 we draw our conclusion.

II. A BRIEF OVERVIEW ON PARSERS AND RELATED WORK

Parsing has been the subject of several research studies in NLP, so the detection of grammatical errors usually requires parsing the sentence. Checking grammar techniques are classified into three main categories: syntax-based, statistics-based and rule-based analysis.

A. The Syntax-based Approach

This approach usually requires a complete grammar, as well as morphological analysis and syntactic parser. It represents the structure of a sentence according to a universal syntactical scheme. This approach can be classified as a deep syntactical analysis. Among the most famous software that uses it, we find Bokmål (Johannessen et al., 2002) [7]. This is a system used in the Microsoft Word Office XP package released in 2001. The disadvantage of this approach is that it only recognizes if the sentence is incorrect, it will not be able to tell the user where the error is. Thus, if a rule or a constraint of several rules is incorrectly used, the identification of errors will be difficult to apply considering the complexity of natural languages.

B. The Statistics-based Approach

The existence of a large amount of text (corpus) motivated researchers to innovate statistical models to extract linguistic knowledge [8]. Among the linguistic-statistical tools, we find Part Of Speech Taggers (POS) and statistical analyzers. Some parsers use statistical tools to implement various tasks to detect grammar errors, this approach can be used with rules-based techniques such as the Granska parser [9].

These parsers need to be developed in a tagged text to infer a grammar that matches the sentence structure. However, they usually do not lead to reusable grammars and are not always easily interfaced with high-level processing. In addition, the cost of annotation of the learning corpus is far from negligible.

C. The Rule-based Approach

In this latter approach, there is a set of rules developed manually and which corresponds to a text that has been labeled by POS, GramCheck [10] is among the software of this approach. It is similar to the statistics-based approach.

For example, after a determinant, the next word is a noun in 50% of cases in English. This is a kind of interesting rule to add in a system that can predict this situation. Other cases can be solved in a more complex way to incorporate these rules.

However, many special cases are not managed and researchers quickly realized that it was not possible to manage the situation. It is very difficult to capture from simple rules all the complexity of natural languages and all the rules are hand-written, which made the development of the constraint grammar a time-consuming, especially in Arabic language.

D. Related Work

Most Arabic error-correction research is devoted to spelling errors, without worrying about syntactic error correction issues. Indeed, the number of searches (that we know so far) on the correction of syntactic errors in Arabic is very limited.

The Arabic GramCheck [11] is implemented in Prolog (SICStus) for some common grammatical errors. The initial purpose of this tool is to detect the errors and possibly to offer suggestions for improvement. The system is based on deep syntactic analysis and relies on a feature relaxation approach for the detection of ill-formed Arabic sentences. However, there has been no new version since its appearance in 2005.

Recently, a proposed project under-development of a Web-based tool [12] can detect Arabic grammar errors using the Deep Learning, Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM) and bidirectional LSTM. Each deep learning model is formed of a set of training data, which will contain labelled sentences. However, the central element of this project is the Arabic corpus, which should not only be annotated for linguistic errors but also indicate types of errors.

Unfortunately, the available corpus does not identify the types of errors, which is a major challenge for this project.

We have also contributed to this field by a new approach based on the automatic generation of correct sentences [13]. This is a very interesting study because we are devoting an important area of Arabic grammar, namely the grammatical case endings.

III. Grammatical Case Endings in Arabic Language

The Arabic language is written from right to left. Its words are generally classified into three principal categories: noun, verb, and particle.

There are two kinds of Arabic sentences [14], nominal and verbal sentence:

- The nominal sentence, where the sentence's first word is a noun (e.g. "الرجل مغربي _ al-rajol maghribi" _ the man is Moroccan).
- The verbal sentence, where the sentence's first word is a verb (e.g. "ولد الرجل في المغرب _ wulida al-rajol fi al-maghrib" _ the man was born in Morocco).

The grammatical case endings "الاعراب" is the change that occurs in the ending of words because of the various factors involved in its grammatical functions.

The Arabic language has three cases endings: the nominative case "المرفوع", the accusative case "المنصوب" and the genitive case "المجرور" [15].

The following is all the possible situations that indicates when to use these categories:

A. The Nominative Case

Nominative case is used in several situations:

- The subject of a verbal sentence.
- The subject and predicate of a nominal sentence.
- The vocative (addressing someone directly).
- The nominative case is also the default for words that are on their own.

B. The Accusative Case

Accusative case is used in the following cases:

- The object of a transitive verb.
- Adverbial expressions of time, place, and manner, indicating the circumstances under which an action takes place.
- The internal object or cognate accusative structure. What does that mean? It is just a way of intensifying an action by following the verb with its corresponding verbal noun and an adjective modifying it.
- The circumstantial accusative. This is a way to describe a condition/action going on at the same time as the main action.
- Shows the purpose of an action.
- The accusative of specification.
- The predicate of "*kāna* كان" 'be' and its sisters (there are 13 of these light verbs).

C. The Genitive Case

We can use the genitive case in these cases

- The object of a preposition
- the object of a locative adverb
- the second term of a genitive construct "إضافة"

Grammatical case endings cause the noun to change in one of the following three ways:

- The final letter's short vowel "حركة" is changed i.e. "مدرسٍ", "المدرسُ", "مدرسٌ" or "المدرسِ".
- A whole letter at the end of the noun is changed i.e. "المدرسون" or "المدرسين"
- Sometime a different form of the noun is used; it is rarely used and only applies to pronouns i.e. "كَ" or "أنتَ".

In the Arabic language, consonants are always written and short vowels are optional. As a result, written Arabic can be fully, partially, or entirely vowelized. In general, Arabic texts are unvowelized except religious texts, texts used in the education of children and poems. In modern Arabic, some vowels are indicated to help readers remove the ambiguity of certain words.

For this reason, we have agreed to treat unvowelized texts and to deal the apparent errors. At a more general level, errors that can be related to grammatical case endings, we quote:

1) The sound masculine plurals "جمع المذكر السالم"

- The final letter is "ن" i.e. "مدرسون" "teachers (male)".
- The second last letter: When the word is nominative, it is the letter "و" i.e. "مدرسون". When it is accusative or genitive then it becomes the letter "ي" i.e. "مدرسين".

2) The dual nouns

- The final letter in the dual is the letter "ن" i.e. "رجلان" "two men".
- The second last letter: When the word is nominative, it is the letter "ا" i.e. "مدرسان". When it is accusative or genitive then it becomes the letter "ي" i.e. "مدرسين".

3) The six nouns

- In Arabic are six singular masculine nouns that take various forms: "أَخٌ" brother, "أَبٌ" father, "حَمٌ" father-in-law, "فَمٌ" mouth, "هَنٌ" thing, "ذُوْ" possessor of.

IV. The Syntactic Approach Adopted and the Chosen Parser

There are different formalisms to create the parsing of a text. Besides, nearly all of the research deals with two syntactic representation structures (i.e. constituency and dependency structure).

Parsing by constituents "studied by linguists like Noam Chomsky. (Covington, 2001)" [16], First divide the sentence into several groups of words called syntagmas. A syntagma is an intermediary between the global set, which is the sentence, and the unitary division composed of the words. The principle of this concept is to be able to subdivide the sentence logically into smaller and smaller groups. The syntagma is actually a set of words, or smallest phrases, having a common linguistic role in the sentence (Fig. 1 represents an example).

On the other hand, as we can see, it is difficult to predict the number of phrases that will represent the sentence. It is even possible that different parsers produce syntagmatic representations of different sizes.

In dependency parsing, each word is syntactically dependent on another word in the sentence, except the main verb of the sentence that will be designated as the root of the sentence. Each word, except for the root, has exactly one and one head of which it is dependent. The tree structure can still be used to visually represent dependency parsing.

The words in the sentence represent the nodes of the tree and the dependency relationships represent the arcs connecting these nodes. In Fig. 2, we visually represent the dependency structure of the same sentence shown above, "The student writes the lesson".

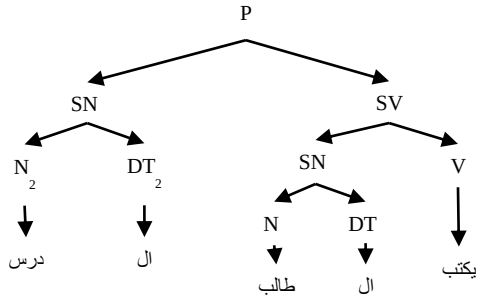


Fig. 1. Syntax of the Sentence: يكتب الطالب الدرس " the Student Writes the Lesson" Represented by Constituents.

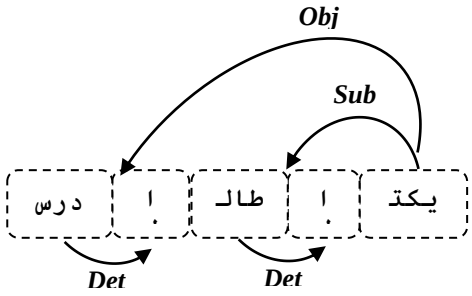


Fig. 2. Syntax of the Sentence: يكتب الطالب الدرس " the Student Writes the Lesson" Represented by Dependency.

We have noted that dependency parsing provides easy processing, promoting supervised machine learning and the use of classical algorithms [17]. Certainly, dependency trees draw a hierarchical design of structuring information wherever each word is associated with a headword of which it is dependent.

As opposed to constituency parsing, where the number of syntagmas defining the sentence cannot be predicted in advance, each generated parsing contains a fixed number of tree elements. Consequently, recognizing that all words have a single head, the dependency parsing includes precisely one component of representation for each word.

Moreover, since parsing of the Arabic language is often followed by semantic analysis, we seek to facilitate it with more appropriate representation. Now the representation of dependency links is much closer to the semantic links than the syntagmatic representation, this is necessary for the Arabic grammar, indeed there is a strong relationship between the syntactic and semantic aspect of the Arabic language.

Since our work is based on a logical description of the syntactic rules of the Arabic language, our choice fell on the parser: Arabic Stanford Parser [4] in order to set up our methodology. It is based on stochastic non-contextual grammars to solve parsing. It was trained on the Penn Arabic Treebank.

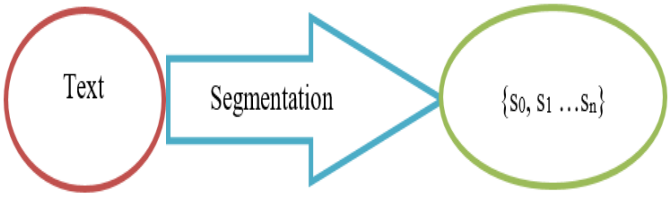
V. Methodology

The syntax correction method adopted is based on the annotation of the Stanford Parser and its syntactical relationships. We can divide this method in 4 steps, which we will detail later:

- Segmentation of text into sentences
- The processing of sentences with the Stanford Parser to obtain the annotation of each word with the syntactic relations between the extracted words.
- The development of correct linguistic relations.
- The detecting and correcting step of grammatical case endings errors.

A. The Segmentation of the Text into Sentences

The segmentation problem of the text into sentences is difficult in the Arabic language. Certainly, it uses neither capital letters nor conventional punctuation that makes the traditional techniques of segmentation adopted in other languages not appropriate to this language. Besides, the agglutination of words is another characteristic of the Arabic that gives segmentation even more challenging to realise (Fig. 3 summarizes this step).



Such as: s_0, s_1, ... are Sentences

Fig. 3. The Segmentation of the Text into Sentences.

B. The Parsing Step

The goal of parsing is summarized in two tasks: creating usable syntactic annotations for the next step the enrichment of linguistic resources for use with our ontology that contains the rules of Arabic syntax.

At this point, we insert the sentences into the Stanford Parser and we draw the label for all language components of the sentence.

The order of linguistic elements

This parser arranges the annotated words using the two fields "index" and "leafNumber" (see Fig. 4), which allows us to know the order of each linguistic element in the sentence, its role appears when there is an ambiguity in the relationships identified by Stanford Parsrer.

```
<S index="0">
     <VP index="1">
           <VBD index="2" value="دخل"
leafNumber="1">
           <NP index="3">
                 <DTNNS index="4"
value="المعلمون" leafNumber="2">
                 <DTJJ index="5"
value="المجتهدون" leafNumber="3">
           </NP>
     </VP>
</S>
```

Fig. 4. The Order of Linguistic Elements.

C. The Development of Correct Linguistic Relations

This step consists of developing the correct linguistic relations by applying the results of the previous step in order to correct the bad relations by using the Arabic ontology [21], which contains a logical description on the set of rules of the Arabic language (we are going to present later the conception of our ontology).

We adopted the syntactic approach of dependency grammar (DG) founded by Lucien Tesnière. Based on the logic of the predicate we will propose a descriptive formalism of traditional Arabic grammar that will finally take a domain ontology [18].

Grammatical relations fall into two categories:

1) A category for grammatical case: This category shows casuals relations between grammatical components such as subject, object, etc. (Fig. 5 shows an example).

Example:

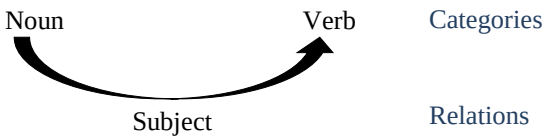


Fig. 5. Categories and Relationships.

Fig. 6. An Example of a Matrix of Values.

2) A category that represents the characteristics of words: The characteristics of the time, the morphology and the genre of the words such as the relation of gender, which assigns to the words the values of masculinization, feminization, etc. This category will be designed using a matrix of values (see an example in Fig. 6).

D. The Detecting and Correcting Step of Grammatical Case Indings Errors

This step consists of comparing the correct sentence and the original sentence, by calculating a distance of "Levenshtein" [19].

Levenshtein distance:

The Levenshtein distance measures the degree of similarity between two strings (in this case words). It is equal to the minimum number of characters that must be deleted, inserted or replaced to move one string into the other. It is (in the mathematical sense of the term) a positive or zero number.

In our case, we have two possibilities:

- If the system has found that, the original sentence and the corrected sentence with our system are the same, then it moves to the next sentence because it considers that the sentence is correct.
- If the two sentences are not the same, then the system suggests the correct sentence to the user.

VI. Example

The purpose of this section is to illustrate the link that must be made between parsing and correcting detected errors.

We shall consider the following incorrect sentence example:

"جاء المعلمين المدعوين" _ the invited teachers came.

To explain the working of the system as a whole, we will apply our approach in order to correct the errors of grammatical case endings in this sentence:

A. The Segmentation of the Text into Sentences

In this example, we only have one sentence, so the text segmentation is equal to the sentence itself.

The result obtained is:

Segmentation (جاء المعلمين المدعوين) = Seg1 (جاء المعلمين المدعوين)

B. The Parsing Step

After the segmentation, we move to this step, which aims to obtain the annotation of each word as well as the syntactic relations:

جاء/VBD

المعلمين/DTNNS

المدعوين/DTJJ

Root (Root, جاء)

dobj (جاء , المعلمين)

dep (المعلمين , المدعوين)

(S (VP (VBD جاء) (NP (DTNNS المعلمين) (DTJJ المدعوين))))

C. The Development of Correct Linguistic Relations

After the parsing, we have to correct the linguistic elements of these relations:

- The first relation dobj(جاء , المعلمين) is a subject relation that exists in our ontology whose relational space must be a nominative noun (here plural) and the extent of this relation must be a verb, however "المعلمين" is a noun accusative plural noun, with the ending "ين" then we have to change it by "ون" and this piece becomes: "المعلمون"
- The second relation dep(المعلمين , المدعوين) is an adjective relation between a qualifying "المدعوين" and a qualified "المعلمين". As we explained in the introduction, in our ontology qualifying it follows the qualified in its case endings, since we have changed in the first relation the ending of "المعلمون" we must change it also for "المدعوين", then it becomes "المدعوون".

We obtain a result of this step a new correct sentence using our ontology: "جاء المعلمون المدعوون".

D. The Detection and Correction of Case Endings Errors

Finally, we compare the obtained sentence and the original sentence using Levenshtein's distance, which detects the errors in an indirect way, so he proposes the correct sentence.

VII. Evaluation and Discussions

In this section, we present the results of the evaluations performed on Arabic sentences.

There are two main approaches useful to NLP evaluation: black-box and glass-box [20]. In black-box evaluation, the test data are chosen only according to the specified relations between input and output, without considering the internal structure of the tested system. In glass-box evaluation, the evaluator has access to the various workings of the system.

In our work, we have chosen the black-box evaluation approach due to it keeps its original meaning, i.e. evaluation of a system's result "correct sentences" for a given input "incorrect sentences". In this context, the evaluation may not be able to locate the source of the error since it detects them in an indirect way, but it will provide an indication of the good or the bad functioning of the system.

In order to validate our approach, we need to evaluate a set of 100 Arabic sentences, which regroups several types of syntactic endings errors, namely:

- 5 errors of the six nouns.
- 30 errors of disagreement between the circumstantial accusative and the subject.
- 35 errors of disagreement between the adjective and the noun.
- 30 errors of disagreement between the permutative and the antecedent "المبدل و المبدل عنه".

For the reader to understand them easily, and so that linguists will find them simple to evaluate, the majority of these sentences were short and simple.

The average sentence length was three words and the longest sentence was 10 words long. Our system includes approximately 200 grammar rules.

A summary of the evaluation results is presented in Tables II and III. The first column shows the various case endings errors in input sentences.

TABLE. II. Results of Detection and Correction of Case Endings Errors of this Work

Case endings errors	Detected	No detected	Total
Disagreement between the adjective and the noun	34	1	35
Disagreement between the circumstantial accusative and the subject	28	2	30
Disagreement between the permutative and the antecedent	26	4	30
The six nouns	5	0	5
Total (in percentage)	94.28%	5.72%	100

TABLE. III. RESULTS OF DETECTION AND CORRECTION OF CASE ENDINGS ERRORS FROM ANOTHER WORK OF THE SAME RESEARCH TEAM

Case endings errors	Detected	No detected	Total
Disagreement between the adjective and the noun	32	3	35
Disagreement between the circumstantial accusative and the subject	23	7	30
Disagreement between the permutative and the antecedent	26	4	30
The six nouns	4	1	5
Total (in percentage)	85%	15%	100

These results are shown in columns 2 and 3. The output is considered correct if the system gives a syntactically correct sentence.

The results of our approach given in the bottom row show a total detection of 94.28% compared with 85% of the available corpus from another work of the same research team, which is a good level for this type of task. Particularly noteworthy is the high level of precision, which characterizes a very significant level of reliability.

This property of detection is important in this case as correcting sentences because if the system finds that the sentence contains an error, then it will automatically move to the next step to generate the correct sentence.

It would be interesting to compare this work with others. However, there are no available systems on the correction of the grammatical case ending errors in Arabic except for the research study [13] of the same team that developed the present work. Furthermore, it is not practicable to make a comparison for other languages.

To conclude, the correction result is overall very positive. It allows correcting the average errors with 94.28% accuracy.

VIII. CONCLUSION AND PERSPECTIVES

We have presented in this article, a methodology of the detection and automatic correction of syntactic errors and precisely the errors of the case endings in Arabic texts. This work is based on Stanford Parser; the processing of his results is done with the rules and constraints obtained through a logical description of Arabic grammar by an ontology.

It is hoped that the results presented will be useful for the development of Arabic syntax checkers for all errors. We wish also to propose an improvement to Stanford Parser in order to analyze and possibly correct syntactic errors.

ACKNOWLEDGMENT

I would like to express the very thanks to my PhD supervisor, Professor Abderrahim TRAGHA from Hassan II University, who gave me the opportunity to do such research, as well as all the members of the laboratory of Information Technologies and Modelling (LTIM).

REFERENCES

[1] A. Farghaly, K. Shaalan, “Arabic natural language processing: challenges and solutions”, ACM Transaction on Asian Language Information Processing, 8(4), pp.1–22, 2009.

[2] M. Boudchiche, A. Mazroui, Ould Abdallahi Ould Bebah, M.Lakhouaja, A. Boudlal AlKhalil Morpho Sys 2: a robust Arabic morpho-syntactic analyzer J. King Saud Univ. – Comput. Inf. Sci., 29, pp. 141-146, 10.1016/j.jksuci.2016.05.002, 2017.

[3] Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania.LDC Catalog No.:LDC2002L49, 2002.

[4] Chen and Manning, Danqi Chen and Christopher Manning. A Fast and Accurate Dependency Parser using Neural Networks. pages 740–750. Association for Computational Linguistics, 2014.

[5] Bikel, D. Intricacies of Collins‟ Parsing Model. Computational Linguistics, 30:4 (479-511), 2004.

[6] C. B. O. Zribi, M. B. Ahmed, "Detection of semantic errors in Arabic texts", Artif. Intell., vol. 195, pp. 249-264, 2013.

[7] J. Johannessen, K. Hagen, P. Lane, “The performance of a grammar checker with deviant language input”, Proceedings of the International Conference on Computational Linguistics (COLING), Taiwan, 2002.

[8] E. Atwell, S. Elliott, “Dealing with Ill-formed English Text in the Computational Analysis of English”, Longman, 1987.

[9] J. Birn, “Detecting grammar errors with Lingsoft’s Swedish grammar checker”, The 12th Nordic Conference in Computational Linguistics, NODALIDA’99, Nordgard T (ed.), Department of Linguistics, Norwegian University of Science and Technology, Trondheim, 2000.

[10] F. Bustamante, F. Leon, “GramCheck: A grammar and style checker”, Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, 5–9, pp. 175–181, 1996.

[11] K. F. Shaalan, “Arabic GramCheck: a grammar checker for Arabic,” Softw. Pract. Exp., vol. 35, no. 7, pp. 643–665, 2005.

[12] Nora Madi et al. A Proposed Arabic Grammatical Error Detection Tool Based on Deep Learning. ACLING, 352–355 DOI: 10.1016/j.procs.2018.10.482, 2018.

[13] C. Moukrim, A. Tragha, H. Benlahmer, T. Almalki, “An Innovative Approach to Autocorrecting Grammatical Errors in Arabic Texts”, Journal of King Saud University–Computer and Information Sciences, doi.org/10.1016/j.jksuci.2019.02.005, 2019.

[14] K. Ryding, “Reference Grammar of Modern Standard Arabic”, Cambridge University Press, Cambridge, UK, 2005.

[15] H. Al-Muhtaseb, C. Mellish, “Some Differences Between Arabic and English: A Step Towards an Arabic Upper Model”, The 6th International Conference on Multilingual Computing, Cambridge, UK, 1998.

[16] M. Covington, “A fundamental algorithm for dependency parsing”, Proceedings of the 39th annual ACM southeast conference. Citeseer, pp. 95–102, 2001.

[17] S. Kubler, R. Mcdonald and J. Nivre, “Dependency parsing. Synthesis Lectures on Human Language Technologies”, vol. 1, pp. 1–127, 2009.

[18] D. Klein and D. Christopher, “Fast exact inference with a factored model for natural language parsing”, in Suzanna Becker, vol. 15, pp. 3–10. MIT Press, 2003.

[19] G. V. Bard, “Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric”, in Proceedings of the fth Australasian symposium on ACSW frontiers, vol. 68, pp. 117-124, Darlinghurst Australia, 2007.

[20] V. Illingworth, “Dictionary of computing”, London, Oxford University Press, 1990.

[21] T. Almalki, “ أنطولوجيا النحو العربي نحو توصيف منطقي لساني للنحو العربي القديم”, دار النابغة للنشر والتوزيع, Tanta. Egypt, 2015. http://arabicontology.org/.

Partition Ciphering System: A Difficult Problem Based Encryption Scheme

Ziani Fatima Ezzahra[1], Omary Fouzia[2]
Computer Science Department
University Mohammed V
Rabat, Morocco

***Abstract*—In this article, a new encryption scheme called Partition Ciphering System is proposed to adapt and process the message according to the partition problem. The objective of this system, that can be applied as a standalone system or as a building block in a bigger system, is to achieve confidentiality, and maintain a balance between ones and zeros in the output so that attacks like frequency cryptanalysis is avoided and good entropy is met. At first, the authors describe the partition problem together with an adapted version. Secondly, the encryption and the decryption processes are provided. Next, a comparison, in terms of the statistical properties using the DIEHARDER battery, security analysis and performance, with other encryption schemes is presented. From the results, the proposed cryptosystem is resistant to frequency analysis and shows good entropy in the output. Moreover, compared to the Advanced Encryption Standard, it has a random behavior and good confusion and diffusion(Avalanche effect). Also, it displays better performance and resistance to brute force attack on the key.**

Keywords*—Encryption scheme; partition problem; frequency analysis; avalanche effect; confusion; diffusion; statistical properties

I. Introduction

One of the principal concerns in cryptography is confidentiality which includes two concepts [1]: the first one is data confidentiality that would not make personal or confidential data accessible or revealed to unauthorized parties. The second one is privacy that helps concerned entities to control and check who can collect or store their data and who is not allowed to have accessibility to these data in any way.

This paper is concerned by the first concept. There are two types of cryptographic algorithms: the first category includes symmetric systems requiring the same key for encryption and decryption. The second category comprises asymmetric systems demanding two different keys for encryption and decryption [2].

This article's contribution lies in the context of symmetric encryption. It was supposed to be combined with the Symmetrical Evolutionist-based Ciphering (SEC) scheme [3] to make it robust against frequency analysis. In brief, SEC [3] is a symmetric encryption system that substitutes the plaintext's characters and consequently changes their appearance frequency using an evolutionary algorithm. The principal purpose is to make them appear at the same rate. As a result, frequency analysis does not reveal any information. Therefore, the partition problem is the source of inspiration to design the proposed algorithm with this constraint. The authors studied the partition problem and defined an adapted version called the Card-Partition problem to accomplish the goal stated before. They have concluded that the scheme can be considered as a standalone system or as a part of a bigger system. Therefore, they decided to present it alone in this paper.

The rest of this article is organized as follow: Section 2 presents related works. Section 3 defines the partition problem. Next, Section 4 describes the Card-Partition problem and the proposed scheme in detail. In Section 5, some symmetric encryption systems follow. Finally, in Section 6, the results are displayed and discussed.

II. Related work

The partition problem, also called Equal Piles Problem, was defined and first studied by Jones and Beltramo [4]. Different genetic algorithms were proposed to resolve this problem, as Jones and Beltramo's genetic algorithm, Falkenauer's one [5], and by 2000 William A.Green proposed a better one named Eager Breeder Greene [6]. Later, genetic and evolutionist algorithms were used in the design of cryptographic algorithms to reach better security like in [7-11]. In [3], [12], [13], and [14], Omary proposed an evolutionist-based encryption scheme and extended versions that aim to substitute the plaintext's characters to change their occurrence frequency. Later, in [15], Trichni proposed an improved version that comprises a new mutation operator based on the partition problem to provide resistance to frequency analysis and brute force attacks. Afterward, in [16], Bougrine proposed a new encryption scheme inspired by SEC [3] based on the same problem to achieve the same objectives, moreover, in [17] and [18] Kaddouri proposed a revised version of SEC[3]. These works produce a variation of the appearance frequency. But in this article, the purpose is not only to change it, but it is also to produce a balanced output. The authors were inspired by the SEC and the Equal Piles Problem to achieve their goal.

III. Background on the Partition Problem

The authors studied SEC [3] and the partition problem, which is also termed The Equal Piles Problem [4], to design this encryption scheme. The idea is to represent the message as a partition that will be processed by the algorithm. This section defines the partition problem.

A. Partition Problem (Equal Piles Problem)

The partition problem (Equal Piles Problem) purpose is to partition a set into subsets(piles) evenly. Formally, it is defined as follow:

Definition

Given a set S of integer numbers, and an integer k. Divide the set S into k subsets such that:

$$\sum_{e\in S_1} e = \sum_{e\in S_2} e = \ldots = \sum_{e\in S_k} e$$

and$S_1 \cup S_2 \cup \ldots \cup S_k = SandS_1 \cap S_2 \cap \ldots \cap S_k = \emptyset$

where e are elements of the subset S_i, and S_i is the i^{th} subset.

The problem is known to be hard. Also, this problem represents the motivation of this article [4]. The authors described a revised version, that leads to the concerned objective, in the next section.

IV. Partition Ciphering System (PCS) Description

In this section, the authors proposed a revised version of the partition problem called the Card-Partition Problem. Moreover, they presented a detailed description of the proposed scheme.

A. Card-Partition Problem

Given a set S of integer numbers, and an integer k. Divide the set S into k subsets such that:

$$Card(S_1) = Card(S_2) = \ldots = Card(S_k)$$

and$S_1 \cup S_2 \cup \ldots \cup S_k = SandS_1 \cap S_2 \cap \ldots \cap S_k = \emptyset$

where Card(S_i) is the cardinal of the subset S_i.

This definition is the main idea to achieve the objective of the article. Therefore, the authors represented the plaintext by a partition. Furthermore, a new partition, that satisfies the constraint of the subsets cardinalities' equality, is constructed (an instance of the card-partition problem). Thus, the resulting ciphertext is resistant to frequency analysis.

B. PCS (Partition Ciphering System) Encryption

The objective of this system is to get a partition in which all the subsets have the same cardinality. The encryption scheme consists of 3 steps:

- At first, the authors define the plaintext partition.
- Secondly, the authors compute the ideal cardinality.
- Finally, the authors add or delete some blocks to construct the ciphertext partition depending on the ideal cardinality value.

The secret key is constructed during the algorithm process.

K={{k},{NumberOfAddedBlocks}, ListOfDeletes, Permutation}

Where, k is the size of the blocks and NumberOfAddedBlocks is the number of added blocks. The ListOfDeletes represents the deleted blocks and their corresponding positions of deletes. Finally, the permutation is the transformation mapping between the plaintext partition and the ciphertext partition.

1) Step 1: The plaintext partition construction: Let the binary message M be the input. The plaintext partition is formed as follow:

At first, an integer $k \geq 2$ is randomly chosen, then the message is split into blocks of size k (B_0, B_1, … , B_{m-1}). Thereafter, to each block B_i, a list of occurrence Li is associated with.

Let n be the number of blocks in M. In other words, n is the size of M. The L_is form a partition of {0, 1, ..., n − 1} such that $0 \leq i < m$ (m is the number of L_is). This partition subsets do not have the same cardinality. Next, the resulting partition is constructed in the next steps.

To reach the scheme's aim, the ideal cardinality, representing the occurrence number of each block in the ciphertext, is specified.

2) Step *2: Ideal Cardinality (IC) definition:* Let $c = \frac{n}{m}$, where n is the number of blocks in M, and m is the number of distinct blocks in M. If c is an integer, then IC=c. Otherwise, $IC = \lceil c \rceil$.

3) Step 3: The ciphertext partition construction: The ciphertext partition is a partition of the set {0, ..., n'- 1}, where n' is the number of blocks in the ciphertext.

Let n'=IC × m, and NumberOfAddedBlocks = 0 at first.

According to the ideal cardinality value, the authors conclude the next step for each list Li.

For each block B_i where ($0 \leq i < m$), the cardinality of the L_i Card(L_i) is compared with the ideal cardinality.

a) Case 1: If Card(L_i) < IC, then the corresponding block B_i is appended to the message M, and the NumberOfAddedBlocks is incremented. Also, the plaintext partition is updated by the position where the block B_i is inserted in L_i.

b) Case 2: If Card(L_i) > IC, then the corresponding block B_i is removed from a randomly chosen position from L_i. And the ListOfDeletes is updated as follows: at first the index i of B_i is inserted in the ListOfDeletes (if i does not exist in the ListOfDeletes), next the position from where it was deleted is inserted. The key is updated during the encryption progressively together with the plaintext partition to produce the ciphertext partition (the deleted block B_i must be removed also from the L_i in the partition).

At the end, the final partition representing the ciphertext is reached.

The secret key is denoted by:

K={{k},{NumberOfAddedBlocks},{IndexOf(B_i)→{PositionsOf(B_i)},...,IndexOf(B_l)→{PositionsOf(B_l)}},Permutation}

Algorithm 1 provides the pseudocode of the PCS encryption process. Where, the ListOfDeletes and the permutation are initially empty.

Algorithm 1: PCS encryption

Input : The message M
Output: The ciphertext C and the secret key K

Begin
M'← EncodeToBinary (M)
k← randomNumber()
NumberOfAddedBlocks ← 0
K← {{k},{NumberOfAddedBlocks},ListOfDeletes, Permutation}
M'← DivideIntoBlocks(M',k)
n← sizeOf(M')
m← NbOfDiffBlocks(M')
PlaintextPartition← ToPartition(M')
ListOfBlocks← DiffBlocks(M')
IC← ComputeIdealCardinality(n,m)
For I from 0 to m-1 **do**
While Card(L_i)<IC **do**
M'← add(B_i, M')
NumberOfAddedBlocks← NumberOfAddedBlocks+1
K← Update(K, NumberOfAddedBlocks)
PlaintextPartition← updatePartition(PlaintextPartition)
EndWhile
While Card(L_i)>IC **do**
M'← Delete(B_i, randomPosition(L_i), M')
K← Update(K, ListOfDeletes)
PlaintextPartition← updatePartition(PlaintextPartition)
EndWhile
EndFor
C← M'
CiphertextPartition← PlaintextPartition
Permutation←GeneratePermutation(PlaintextPartition, CiphertextPartition)
K← Update(K, Permutation)
End

Fig. 1 summarizes the encryption process detailed before.

C. PCS (Partition Ciphering System) Decryption Algorithm

The decryption algorithm consists of two steps: the ciphertext is first split, and then the inverse actions of the encryption process are done.

1) Step 1: Given the ciphertext C and the secret key *K = {{k}, {NumberOfAddedBlocks}, ListOfDeletes, Permutation}*, C is split into blocks of size k. The *ListOfDifferentBlocks* is defined to be the list of different blocks in the ciphertext.

2) Step 2: At first, each of the inserted blocks is removed from the last position in the message and the *NumberOfAddedBlocks* is decreased by 1 each time.
Next, the *ListOfDeletes, ListOfDifferentBlocks, and Permutation* are used to insert each of the deleted blocks in the position it was removed from.

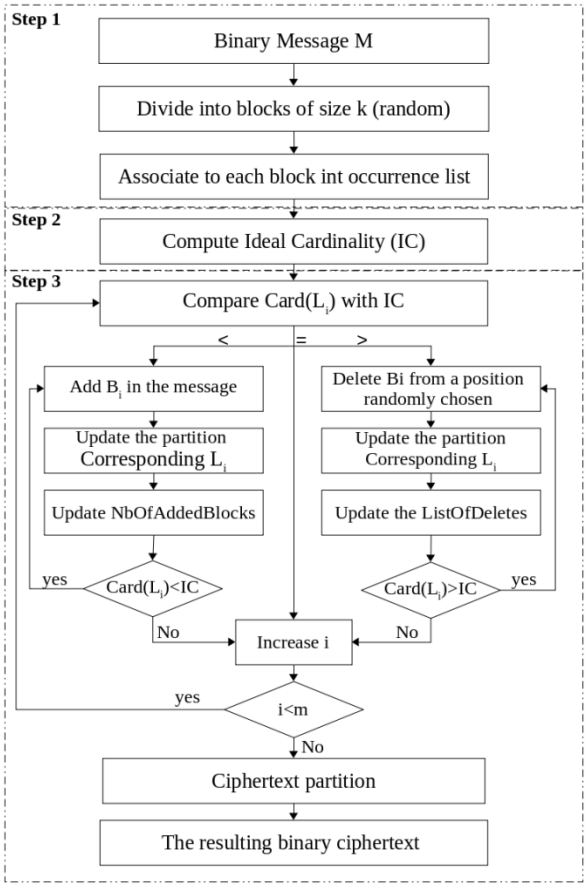


Fig. 1. PCS Encryption Process.

Algorithm 2 describes the decryption algorithm.

Algorithm 2: PCS decryption

Input : The ciphertext C and the secret key K
Output: The message M

Begin
C←DivideIntoBlocks(C,k)
ListOfDiffBlocks← DifferentBlocks(C)
While NbOfAddedBlocks > 0 **do**
C← DeleteFromLast(C)
NbOfAddedBlocks← NbOfAddedBlocks-1
EndWhile
For i from 0 to sizeOf(ListOfDeletes) **do**
Add(Permutation, ListOfDiffBlocks,ListOfDeletes, C)
EndFor
M← C
End

Fig. 2 displays the PCS decryption process.

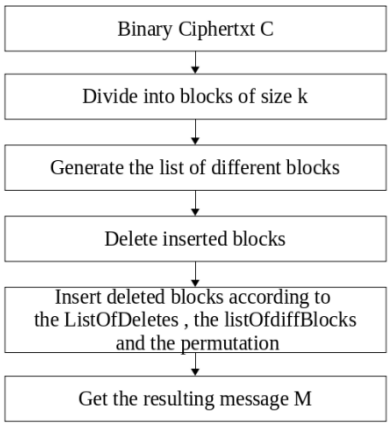


Fig. 2. PCS Decryption Process.

V. Some Symmetric Encryption Schemes

The most famous symmetric encryption schemes are DES (Data Encryption Standard), 3DES (3-Data Encryption Standard) and AES (Advanced Encryption Standard). This section describes them briefly.

A. DES (Data Encryption Standard)

The Data Encryption Standard is a symmetric block cipher that encrypts 64-bit data blocks by using a 56-bit key. It consists of 16 Feistel iterations surrounded by two permutations, IP at the input, and its inverse IP^{-1} at the output. The 48-bit sub keys are formed from the 56-bit key using a permutation [19].

B. 3DES (Triple Data Encryption Standard)

The 3DES is a symmetric block cipher that encrypts 64-bits data block. Using three keys of size 56-bits. It is an enhancement of DES, which consists of 48 Feistel rounds. It is vulnerable to differential cryptanalysis [20].

C. AES (Advanced Encryption Standard)

The Advanced Encryption Standard is a symmetric block cipher that encrypts 128-bit data blocks. It uses symmetric 128-bit, 192-bit or 256-bit keys. It consists of 10, 12 and 14 rounds depending on the key size. Brute force attack is the only effective attack known against this algorithm. AES encryption is fast and flexible [21]. Table I presents the characteristics of DES, 3DES and AES.

TABLE. I. Characteristics of DES, 3DES and AES

Parameters	**Encryption systems**		
	DES	***3DES***	***AES***
Key length	56 bits	168 bits	128,192, or 256 bits
Block size	64 bits	64 bits	128 bits
Developed	1977	1978	2000
Cryptanalysis resistance	Vulnerable to differential and linear cry-ptanalysis	Vulnerable to differential cryptanalysis	Strong against differential, tr-uncated differe-ntial, linear, int-erpolation and square attack

VI. Results and Security Analysis

This section presents the statistical tests and the test of confusion and diffusion properties. Also, PCS and some symmetric encryption schemes are compared.

A. Dieharder Test

Dieharder battery was developed to test the behavior of the pseudo-random number generators and other cryptographic features like encryption schemes and hash functions. This battery consists of 32 tests [22]. A file that contains a sequence of 10 Mb is generated using the PCS algorithm. The algorithms of the battery compute the p-values. The significance level for Dieharder is 0.005, and if the p-values are on the range [0.005, 0.995], then the results are good enough[21]. Fig. 3 shows that PCS passed all the tests of Dieharder battery, as 0.2<p-values(PCS)<0.9. Also, the AES outputs p-values 0.05<p-values(AES)<1. All the p-values of PCS are good enough to conclude that the behavior of the scheme is random. Also, the AES has a random behavior even if some p-values are near to 0.995 and 0.005. Fig. 3 shows that PCS results are better than the AES results.

B. Confusion and Diffusion Properties

In this part, the confusion and diffusion properties are tested for the PCS scheme and compared to the AES. From Shannon's view point, to decide if an encryption scheme is secure against statistical analysis, it is required to satisfy the confusion and diffusion properties [23]. AES is known to have good confusion and diffusion properties. Confusion represents the relation of the ciphertext with the key that must be complex. Moreover, diffusion represents the relation between the plaintext and the ciphertext (changing one character /bit in the ciphertext/plaintext should influence a significant number of the plaintext/ciphertext characters).

The avalanche effect is the best tool to check these properties. The diagram in Fig. 4 illustrates the confusion property for PCS. From Fig. 4, the portion of the changed bit in the ciphertext is approximately 50 % for PCS(Avalanche effect) and AES. These values mean that this scheme satisfies the confusion property.

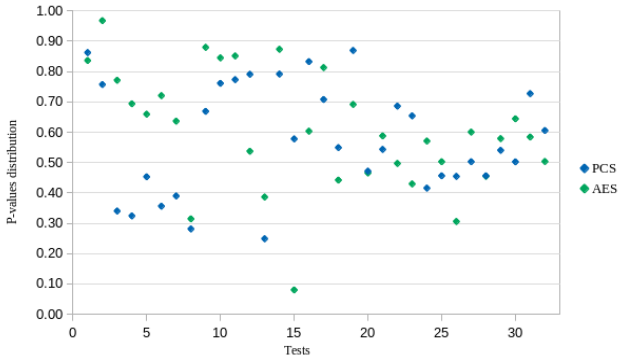


Fig. 3. The Dieharder Results of PCS Encryption Algorithm and AES.

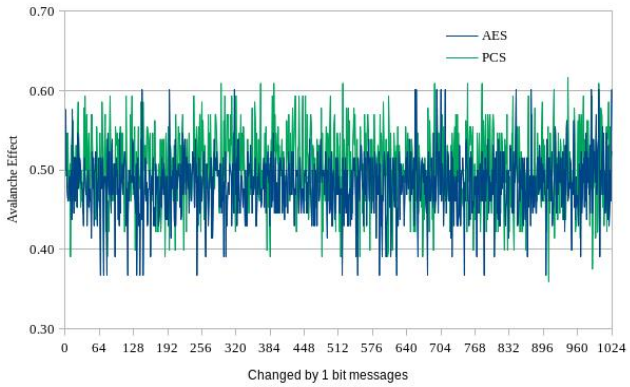


Fig. 4. The Confusion Property of PCS Encryption Algorithm Compared to AES (Avalanche Effect).

Table II illustrates the diffusion property of PCS. The average percentage of changed bit in the ciphertext is around 50%. Furthermore, since PCS has a session key, in case of a compromised state, it will not affect other pairs of plaintexts and ciphertexts. It is related to its encryption process randomness. Also, the same message is differently encrypted each time. To conclude, from the results of the avalanche effect and the statistical tests provided, the PCS has good confusion and diffusion properties.

C. Comparison of PCS with AES, DES and 3DES

1) Encryption and decryption time: In this part, the authors compared the encryption and decryption time of the proposed scheme PCS with DES, 3DES, and AES (see Fig. 5).

The authors noticed that:

- DES has a higher encryption and decryption time compared to the AES and PCS. Parallel computing made breaking DES quite simple. 3DES is not as vulnerable as DES, but it is too slow compared to the other schemes.
- AES has a shorter encryption time compared to DES and 3DES encryption schemes. And it is equivalent to the PSC. Even if the structure simplicity might be inconvenient, AES is faster, more flexible, and stronger than DES and 3DES from the Table I.
- PCS has an encryption time equivalent to the AES and better than DES and 3DES. It has a shorter decryption time comparatively with the others.

2) Security comparison: The keys length is a useful metric when it comes to the cryptographic strength. Because if a longer key is used to encrypt a text, it is hard to decrypt without the appropriate key.

TABLE. II. DIFFUSION PROPERTY OF THE PCS ENCRYPTION ALGORITHM

	Ciphertext pairs								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
% of change	51	47	56	59	41	49	52	60	56

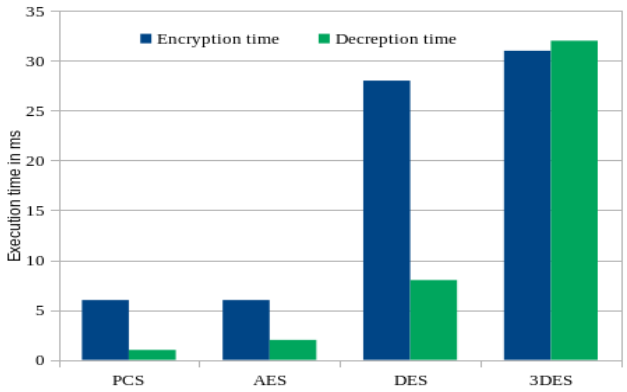


Fig. 5. Encryption and Decryption Time of PCS, AES, DES and 3DES.

Table III shows the strength of the PCS compared to 3DES, DES, and AES, depending on the key length. Moreover, PCS encryption is resistant to linear and differential cryptanalysis. Further, it is robust against frequency analysis. Additionally, the key used in the PCS is a session key, which is not the case of AES, DES, and 3DES.

D. Frequency Analysis

In this subsection, the authors performed the frequency analysis before and after the encryption algorithm. Fig. 6 presents the results of the frequency analysis. As stated before, the purpose of PCS is to have a balanced output. In other words, each block appears with the same frequency. The articles [15],[16],[17], and [18] did not achieve this objective. In PCS, the encryption did not use an evolutionary algorithm to solve the problem.

TABLE. III. KEY LENGTH OF THE 3DES, DES, AES AND PCS FOR M_1

	Encryption Schemes			
	DES	***3DES***	***AES***	***PCS***
Key length	56 bits	168 bits	128 bits	952 bits
# possiblle keys	2^{56}	2^{168}	2^{128}	2^{952}

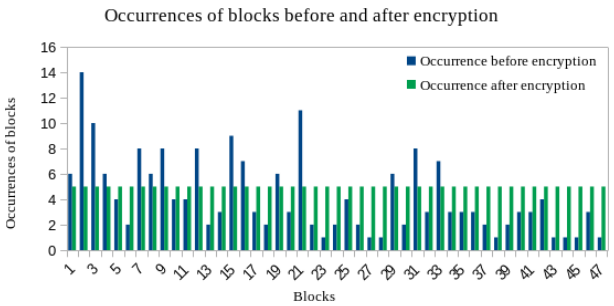


Fig. 6. Occurrences of Blocks before and after Encryption using PCS.

VII. CONCLUSION

In this article, the Partition Ciphering System (PCS), which is inspired by a previously designed system SEC and the partition problem, is proposed. The purpose of PCS is to generate a balanced ciphertext to avoid frequency analysis. Yet, the standard definition of the partition problem did not

provide the objective of PCS. Consequently, the authors proposed a revised partition problem definition called the Card-Partition problem, in which the subsets have the same frequency. The PCS encryption algorithm is a possible way to solve the problem. They performed a comparison of the PCS with AES, DES, and 3-DES. Thus, according to the Dieharder test battery, the results display the random behavior of PCS. Compared to the AES, the PCS provided better results. The results of the brute force attack and frequency analysis are promising. Moreover, the confusion and diffusion properties are satisfied in PCS.

In further work, a new version of the PCS scheme is going to be done to make it suitable for the security of the wireless body area networks and IoT devices in general.

REFERENCES

[1] Stallings, W. (2017). Cryptography and Network Security. 7th ed. Harlow, United Kingdom: Pearson Education Limited.

[2] Douglas ,R. Stinson , Cryptography theory and practice,Chapman & Hall/CRC, U.S .2006. 19-20.

[3] Omary, F., Mouloudi, A. , Tragha, A., Bellaachia, A., A New Ciphering Method Associated with Evolutionary Algorithm,Lecture Notes in Computer Science – Publisher: Springer Berlin / Heidelberg -ISSN: 0302-9743 -Subject: Computer Science-Volume 3984 ,pp 346-352(2006).

[4] Jones,D.R., Beltramo,M.A.," Solving Partitionning Problems with Genetic Algorithms", in Belew, K. R. & Booker,L. B. (Eds.),Proceedings of the Fourth International Conference on Genetic Algorithms;Maurgan Kaufmann Publ.,San Francisco(1991).

[5] Emannuel, F., Solving Equal Piles with the Grouping Genetic Algorithm, in Eshelman, L. J. (Ed.), Proceedings of the Sixth Intenational Conference on Genetic Algorithms; Morgan Kaufmann Publ., San Francisco (1995).

[6] Greene, W.A., Partitioning Sets with Genetic Algorithms, in J. Etheredge and B. Manaris(eds.), Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society (FLAIRS) Conference, May 22-24 , 102-106. (2000).

[7] Kumar, A. and Chatterjee, K. (2016). An efficient stream cipher using Genetic Algorithm. 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).

[8] Putera, A., Siahaan, U. and Rahim, R. (2016). Dynamic Key Matrix of Hill Cipher Using Genetic Algorithm. International Journal of Security and Its Applications, 10(8), pp.173-180.

[9] Pareek, N. and Patidar, V. (2014). Medical image protection using genetic algorithm operations. Soft Computing, 20(2), pp.763-772.

[10] Khare, M. and Yadav, C. (2017). Secure data transmission in cloud environment using visual cryptography and genetic algorithm: A review. 2017 International Conference on Innovations in Control, Communication and Information Systems (ICICCI).

[11] Sethi, P. and Kapoor, V. (2016). A Proposed Novel Architecture for Information Hiding in Image Steganography by Using Genetic Algorithm and Cryptography. Procedia Computer Science, 87, pp.61-66.

[12] F.Omary, A.Tragha, A.Lbekkouri, A.Bellaachia, A.Mouloudi, "An Evolutionist Algorithm to Cryptography"-Brill Academic Publishers-Lecture Series And Computational Sciences, Volume 4, 2005, pp.1749-1752.

[13] A.Mouloudi, F.Omary, A.Tragha, A.Bellaachia, "An Extension of evolutionary Ciphering System". 2006 International Conference on Hibrid Information Tchnology,Novembre 9th – 11th,2006.

[14] F.omary, A.Tragha, A.Bellaachia, A.Mouloudi. « Design and Evaluation of Two Symmetrical Evolutionist-Based Ciphering Algorithms ». International Journal of Computer Science and Network Security (IJCSNS) February 28, 2007 pp 181-190.

[15] S.Trichni, F.Omary, B.Boulahiat, M.Bougrine,"A new approach of mutation's operator applied to the ciphering system SEC", 6th ICCIT: International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2011), Jeju,Corea,pages:680-685,Novembre 2011.http://www. aicit.org/iccit "

[16] M. Bougrine, F. Omaiy, S. Trichni and B. Boulahiat, "New evolutionary tools for a new ciphering system SEC version," 2012 IEEE International Carnahan Conference on Security Technology (ICCST), Boston, MA, 2012, pp. 140-146. doi: 10.1109/CCST.2012.6393549.

[17] Kaddouri Z, Omary F, Abouchouar A, Daari. M. Balancing Process to the Ciphering System Sec. Journal of Theoretical and Applied Information Technology 2013; 52: 092-093.

[18] Kaddouri Z, Omary F, Abouchouar A. Binary Fusion Process to the Ciphering System "Sec Extension to Binary Blocks". Journal of Theoretical and Applied Information Technology 2013; 48: 067-075.

[19] Biryukov, A., De Cannière, C., Encyclopedia of Cryptography and security, 295-300. Springer, Heidelberg (2011).

[20] Hamdan,O.A., Zaidan, B.B., Zaidan, A.A., Hamid, A.J., Shabbir, M., and Al-Nabhani, Y. , New Comparative Study Between DES, 3DES and AES within Nine Factors,Journal Of Computing, Volume 2, Issue 3, March (2010).

[21] G. Brown, R. (2019). Robert G. Brown's General Tools Page. [online] Webhome.phy.duke.edu. Available at: http://webhome.phy.duke.edu/~rgb/General/dieharder.php.

[22] Daemen, J., Rijmen, V., Encyclopedia of Cryptography and security,pp 1046-1049. Springer, Heidelberg (2011).

[23] Shannon,C. E.. A mathematical theory of communication. The Bell System Technical Journal, 27(3): 379–423 ,(1948).

The Implementation of Business Intelligence and Analytics Integration for Organizational Performance Management: A Case Study in Public Sector

Jamaiah Yahaya[1]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Aziz Deraman[3]
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu, Kuala Terengganu
Malaysia

Nur Hani Zulkifli Abai[2]
UUM Information Technology
Universiti Utara Malaysia
Sintok, Kedah, Malaysia

Yusmadi Yah Jusoh[4]
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia (UPM)
Serdang Selangor, Malaysia

***Abstract*—Literature study shows that several works have been conducted on the implementation of BI in performance management, but the analytical aspects were not being considered. Business analytics is an activity of applying analytics to strengthen strategic and operational business activities. While performance management is important to determine organisational success and in public sector, it has become more challenging due to generality of public sector objectives and different level of stakeholders involved. Existing frameworks were built separately and this limits the implementation of Business Intelligence and Analytics as an integrated component, and could not meet the current performance management needs and expectations. The objective of this study is to establish a framework that integrates elements of business intelligence, analytics and performance management for the comprehensive implementation in public sector. This study identifies four main components of this integrated framework: Process, People, Governance and Ability. Each component consists of several key elements and sub-elements. The proposed framework is validated and implemented by real case study conducted in one organisation in Malaysia. The implementation demonstrates the suitability and practicality of this framework to be implemented in real environment.**

***Keywords*—Business intelligence; business analytics; organisational performance management; framework; case study**

I. INTRODUCTION

The implementation of business intelligence (or BI) in managing organisation's performance in public sector had drawn people and government attention and interest. The essential of this integration is due to the great impact to the nation and people in general. People are now wiser to evaluate the transparency of public sector administration in managing national resources. Therefore, managing and measuring organisational performance [1] had been critical agenda in public sector transformation process. BI has been identified as an effective technology in strategically managing performance. It enables users to gather, integrate, access and analyse data to assist efficient decision making in the organisations.

Performance management is a process to facilitate in managing resource and measuring outcomes of the organisation [2]. It analyses organisational goals and divides them into specific benchmarks to ensure the goals are measurable. Performance management is important to determine organisational success. In public sector, organisational performance management (OPM) become more challenging due to generality of public sector objectives involving different level of people. OPM also consists of multi-level hierarchy that causes complexity in decision making process and dissemination of information to the target group. As we know, public sector includes strict rules and procedures throughout their management processes thus make it more challenging and complicated. BI implementation enables to manage and coordinate information within organisation effectively. However, current BI implementation in managing performance does not effective enough to achieve organisations' competitiveness in business. This is due to large volume of information that beyond the ability of decision makers to conduct analysis for best actions in decision making without proper, integrated and systematics mechanism and tool. In addition, scattered piles of data that led to the provision of information for analysis takes longer time.

BI implementation today should consider and focus on analytic aspects to meet current performance management needs. Analytic generally means skills in applying data analysis, especially in thinking or reasoning process. In the context of this study, analytic refers to the process of develop an understanding of action through defining problem and use of statistical models on existing data. The integration of business intelligence and analytics (BIA) improves the sustainability of organisation in their business environment and stays competitive. The increasing in complexity and competitive in current business environment had urged

This work is supported in part by the Malaysia Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/2015/ICT04/UKM/02/1) and PP-FTSM-2019.

managers to use analysis, trends and forecasting in their business operations. Base on this, BIA is highly demanded to drive actionable insight for better decision making. Nowadays, BI implementation has shifted to new perspective that require advanced analytic adoption. Both BI and analytic should be implemented parallel to maximise the impact in organizational performance management.

II. Background Works

The theoretical study of this research focuses on four main areas which are business intelligence, analytics, organizational performance management (OPM) and Business Intelligence and Analytics (BIA) models.

A. Business Intelligence

Business Intelligence (or BI) can be referred to as the technologies, applications and practices that is usually employed to give support to decision making in business. With the use of BI, managers are able to transform and manipulate information from data to valuable knowledge or insight [3]. BI has been earlier introduced as early as in year 1959, where the definition was vague [4] but was strengthen by Howard Dresner in 1989 [5]. Data warehouse introduction in BI is the beginning of transformation in decision support landscape towards data-driven. Data analysis was focused from using previous data and moved to real-time information. Additionally, BI provides enhanced visualisation features that are user friendly, appealing and easy to comprehend, such as scorecard and dashboard to carry out decision making in strategic actions [6] that will improve organisation's performance and operation.

However, the effectiveness of BI implementation depends on how it can be implemented. Kimball Lifecycle methodology to develop BI outlined phases of BI implementation [7] which can be implemented in series of phases: project planning, identify business requirements, design, development, installation and enhancement. This lifecycle has been used as a baseline in many other BI implementation studies and [8] [9] are some of the mentioned studies. In our study, Kimball Lifecycle is used as a baseline of holistic BI implementation as proposed in this research.

BI implementation is a complex process that requires a thorough understanding. Success factors of the implementation play the important role in ensuring the realisation of BI implementation. These factors demonstrate the ability of the organisation to implement and gain value added from its implementation. Critical literature review has revealed and classified the factors into four categories [10] which are the enabler, process, governance and technology. Currently, BI has been applied in various sectors such as medical, education, retails, banking, manufacturing [39] and Enterprise Resource Planning (ERP) [40]. Previous studies has revealed that BI enabled to improve the effectiveness of strategies formulation, operational planning and strengthen relationships with customers [8][11][12]. It improves business processes and collaboration between departments, thus increase organisational performance [13]. BI technology is still growing rapidly with the demand from users to get future forecast information embedded. Therefore, its implementation needs to be integrated with advanced analytics to be used in an increasingly challenging business environment.

B. Business Analytics

Business analytics (BA) is an activity of applying analytics to strengthen strategic and operational business activities. It is defined as "delivering the right decision support to the right people at the right time" [14]. As revealed by [15], BA comprises of several activities: the process of data collection, data analysis and data transformation. BA implementation represents a combination of a few data analysis procedure for gathering unstated information directed to practical insight. In which each procedure shall combine various analytical methods and techniques for effective BA implementation strategy. The problem to be solved will determine the appropriate analytical method to be used. Hence, BA effectiveness depends on the accuracy of the identification of problems.

Business analytics comprises of six levels of processes that begin with understanding the needs to perform analytics in organization. Down to the task of establishing data to be analysed through the usage of data mining methods to provide solution for identified problem. In identification of data, three levels are involved, these are: (i) data preparation, (ii) data quality, and (iii) format transformation. The process continues with the development of model that consists of designing sequence of data analysis techniques. Thereafter, the developed model needs to be evaluated using selected testing method. And the final process is to apply obtained information in decision-making.

In summary, there are 11 activities that have been identified [10]: Identify problem/opportunity, understand data, collect data, transform data, analyse data, develop model, evaluate model, use, translate output, measure impact and maintenance. Other business analytics implementation process such as measuring impact of model usage and model maintenance got less attention from most researchers. However, in terms of business analytics implementation, it is believed that it is able to ensure sustainability. The success of implementation in business analytics also depends on the skill and knowledge of data analyst and ability of appropriate software used [3].

C. Organisational Performance Management

Managing performance is a critical task in organisation to ensure every sources are organised to improve overall performance of an organisation. One the main activities in managing performance is analysing business process and its related metrics for ensuring the optimisation of general organisation achievement [16]. This research was conducted to enquire into the implementation of BI for Organisational Performance Management (OPM). It focuses on strategizing organisational goals by taking into accounts metrics and processes that impact the organisational performance.

OPM is conducted in organisation to interpret objectives into actions. This includes determining strategic plan, monitoring its implementation, and distributing performance achievement [16] [17]. Concurrently, the performance's implementation shall increase to maximum impact by

continuously assess and enhance the operational processes [18]. This will also support proactive environment within organization to produce strategy [14] that is aligned with organisational objectives. Previous study found [15] that OPM implementation is more challenging in public sector rather than in private sector due to different target in objective between tangible and intangible objectives. Therefore, public sector's performance management requires business intelligence and analytics technologies to support decision making process and organisation's objectives alignment.

D. Current Works of Business Intelligence, Analytics and Organisational Performance Management Integration

Previously, several studies have been carried out on BI, BA and OPM implementation. The Gartner Business Analytics Framework (GBAF) [19] has developed an integrated framework of these three fields. Nevertheless, this framework has limitation in term of implementation from public sector performance management perspective and the overall integration.

In addition, several studies also considered identifying BI, BA and OPM elements and enterprise resource planning [20]. Other studies proposed the implementation framework of BI by focusing on the application of data mining techniques [21][22] [23]. While, others studies [24][25][26] proposed the framework of integration BI and business analytics only. Thus, we can conclude that the existing frameworks are not comprehensive in the perspective of BIA implementation that integrates in managing organisational performance.

Rayner and Schlegel revealed characteristics needed for the implementation of BI, BA and OPM [27] and shall be integrated and comprehensive throughout the organisation, particularly in IT section, management and financial [28]. The organisation should have a complete, effective and efficient life cycle process [27][29] for optimum performance. With proper and systematic life cycle process, any changes in the strategies will follow by appropriate actions from parties involving IT, management and data analysis. Hence, it creates the new innovative environment using intelligence and implied information [30].

In matured BI implementation, organisation should be ready with the integrated technology architecture [27][29] which includes software and hardware to support BI activity and analytics. The secure and flexible architecture [31] is required for easy customisation according to the dynamic needs of the organisation and also to ensure the quality and reliability of information obtained for supporting the analytic implementation and decision making. Furthermore, matured analytic implementation applies prescriptive analysis that requires real time data.

III. METHOD

An empirical study was conducted in Malaysia to understand issues and problems in current implementation of business intelligence and analytic for performance management (or BIAPM) specifically in Malaysian public sector. In this survey, 18 individuals were interviewed as the informants with various positions in their organisations. They were divided into three different categories of backgrounds which were business intelligence (BI), business analytics (BA) and organisation performance management (OPM). The results reveal 20 important elements for BIAPM implementation and they are classified into four main components: Process, Governance, People and Ability. The elements are then broken down into 64 sub elements. The 20 elements are: Top Management, Performance Manager, BI implementer, Data Scientist, Domain Expert, Skill, Work Culture, Technology, Strategic Planning, Requirement Analysis, Decision Maker, Performance Evaluation, BI implementation, Software, Finance, Data and Change Management, Documentation, Analytics and Visualisation [10]. Furthermore, in order to obtain the detail implementation of each element in BIAPM, the relationship analysis between sub elements was also been conducted.

The result of this study has disclosed the essential factors and sub factors for the BIA and performance management implementation. The integration factors between business intelligence, analytics and performance management in public sector organisations was conducted and the relationships and relative association of each sub factors were performed and identified based on the significance of the sub factors in related field. The analysis concluded that there were four main factors integrated between these three fields, which are skill, documentation, visualisation and work culture as discussed in our previous paper [10]. The findings verify that in order to integrate BIAPM implementation, every parties involved should have appropriate skills to drive each fields, supported with skilled staff with positive working culture. Furthermore, every staff involved in this implementation should capable to work in a team and highly motivated to ensure that projects and tasks run efficiently. In relation, each team should prepare sufficient documentation to support the project. Another important element in the integrated BIAPM implementation is visualisation of data. This is due to the importance of portraying data to be understood by every party even though they come from different backgrounds.

Fig. 1 illustrates the integration between these factors and the fields. It also illustrates that each factor is interrelated with each other's within these three main fields as defined in this research.

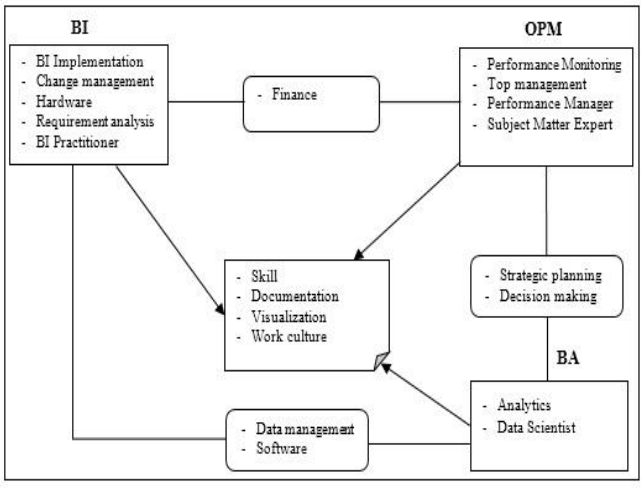


Fig. 1. BA, BI and OPM Integration.

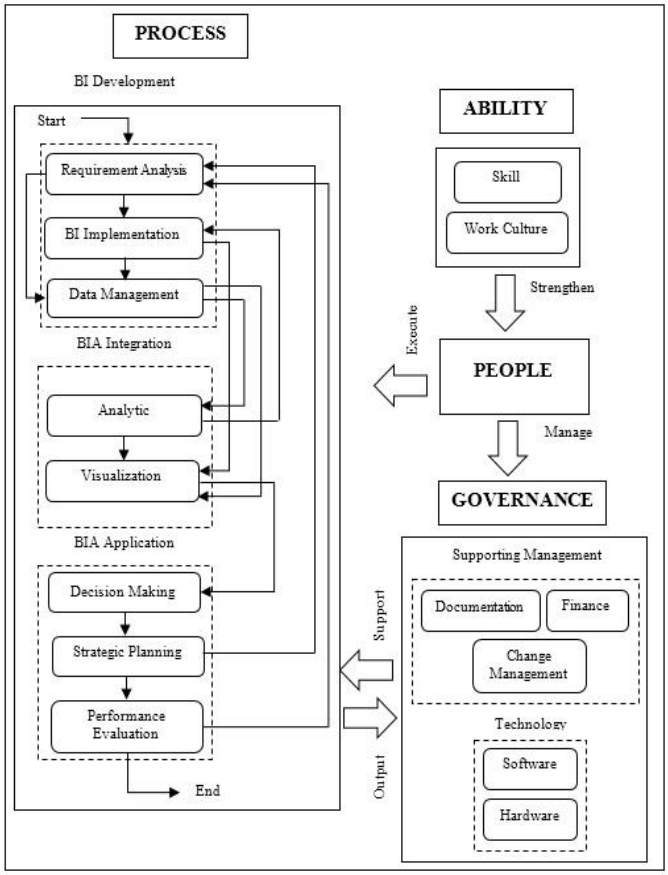


Fig. 2. BIAPM Framework.

Based on this finding, the BIAPM integration framework was developed as shown in Fig. 2. The integrated BIAPM implementation framework consists of four main components: Process, Governance, People and Ability and their relationships between components and elements. The detailed description of BIAPM integrated components and BIAPM framework can also be referred in our previous publications [10][32].

IV. The Case Study

In order to successfully integrate the business intelligence and analytic with organisation performance management, the flow of the main components and their relationships must be linked and followed systematically. The implementation and application are conducted to verify and validate the proposed framework. It is conducted in a real case study implementation. This method is considered suitable for verification and validation of framework [33] through internal investigation on specific phenomenon and real context data. The case study implementation and application are carried out in three steps as suggested by [33][34]. The activities include case selection, application of the framework, and validation of framework suitability.

A. Case Selection

The organisation selection is based on purposive sampling [35] and the criteria used for selection are: 1) practicing business intelligence technology in organisation performance; and 2) applies to only organisation in public sector. Several public sector organisations that fulfil these criteria were invited to join this study but only two agreed to participate in depth and comprehensively.

The case study was carried out with two public organisations in Malaysia, Case A and Case B. Case A is a statutory body agency which is also a public university in Malaysia. It has 17 faculties and supported by 28 departments, centres and institutes. In 2009, this public sector organisation has experienced in business intelligence technology, launched a project associated with it and still operating with the staff involved from information technology section, statistic, researchers and management.

B. Application of the Framework

Application of the framework is conducted in three main phases as shown in Table I.

C. Framework Validation

For validation of the framework for suitability, the results and outcomes from the implementation and validation were presented to the stakeholders of the organisation. Recommendation for improvements were being proposed to assist the organisation in implementing integration of business intelligence and analytics for performance management.

V. BIAPM Implementation

The implementation and application of the framework in Case A was carried out based on the steps and activities defined in Table I.

A. Phase 1: Pre Assessment Phase

This phase aims to plan in detail the assessment process of the BIAPM implementation of organisation in Case A. A discussion was carried out to achieve the following aims:

a) Develop commitment with organisation management

b) Select assessment team which comprises of head of IT operational, external assessor and representative, internal assessor and representative.

c) Plan assessment activities to be carried out, resources needed and assessment time duration.

d) Prepare for assessment material.

TABLE. I. BIAPM Framework Implementation and Application

Phase	Activity
Phase 1 : Pre Assessment	1) Form the task force team 2) Appoint the assessment team and participants 3) Plan for the assessment 4) Preparation for assessment
Phase 2: Current Implementation Assessment	5) Identify and preparation for assessment participants 6) Interview participants 7) Document review 8) Observation 9) Record information
Phase 3: BIAPM Implementation & Analysis	10) Assessment analysis 11) Assessment result presentation 12) Review result and discussion 13) Report preparation

B. Phase 2: Current Implementation Assessment

a) Identify and ensure commitment from participants.

b) Obtain current implementation status and information.

c) Identify documents status.

d) Observe working environment.

e) Gather assessment form.

C. Phase 3: Implementation and Analysis

a) Assess BIAPM implementation level.

b) Prepare assessment report.

c) Present findings to stakeholder.

Phase 3 starts with analysis of data to identify the BIAPM implementation level. There are four score ratings that are for sub elements, element, component and overall implementation. The computational method is adapted and configured from [36] [37]. The scores are computed using formulas as shown in (1), (2), (3) and (4).

Formula (1) is used to compute scores for sub elements.

$$S = \frac{1}{6} * \sum_{i=1}^{n} X_i \quad (1)$$

Where, S= average score for each sub element

n= number of items

X_i = value for sub element

Formula (2) is to compute scores for each elements.

$$E = \frac{1}{n} * \sum_{i=1}^{n} S_i \quad (2)$$

Where

E= average score for each element

N= number of items

S_i = average score value for sub elements

In the third step, formula (3) is used to compute scores for each components in the proposed framework.

$$K = \frac{1}{n} * \sum_{i=1}^{n} E_i \quad (3)$$

Where

K= average score for each component

n= number of items

E_i= average score value for element

In the fourth step, formula (4) is used to compute the overall score for the implementation of BIAPM as proposed in this framework.

$$F = \frac{1}{n} * \sum_{i=1}^{n} K_i \quad (4)$$

Where

F= average score for BIAPM implementation

n= number of items

K_i= average sore for each sub elements and elements

TABLE. II. IMPLEMENTATION LEVEL FOR SUB ELEMENT, ELEMENT AND COMPONENT

Level	Implementation Level	Description
5	Excellent	The implementation is complete and organisation.
4	Good	Organisation has a clear definition on practices and has overall implementation.
3	Moderate	Organisation has become part of management practice but seem not considered as important to be implemented.
2	Need Improvement	Organisation recognise practices but the implementation has not be focused.
1	Not Implement	No information on the implementation practices.

Scores obtained in this exercise which comprises of scores by sub elements, elements and components are then mapped into implementation assessment scale for public organisational performance assessment as shown in Table II.

VI. RESULTS AND DISCUSSION

Organization A is selected as a suitable case study to apply and implement BIAPM framework. It is a public university in Malaysia. The application and case study implementation are described and detailed in three stages of implementation: The application of BIAPM framework, findings and results of the BIAPM application, and recommendation for improvement. The following sections describe these in detail.

A. The Application of BIAPM Framework

The application and implementation exercise were taken placed during the month of June 2015. It was conducted according to the processes explained in previous section. Five person were identified to be the valid participants which included Senior IT Officer, IT Officer, Assistant IT Officers and Social Science Research Officer. The current implementation assessment was carried out through interview, observation and document review.

B. Findings and Results of the BIAPM Application

The collected data was analysed and assessed based on computational method and formulas discussed in previous section. The results of the implementation are mapped into the implementation level as shown in Table IV.

TABLE. III. IMPLEMENTATION LEVEL FOR BI, BA AND OPM INTEGRATION

Level	Implementation Level	Description
1	Not implemented	No information on the implementation practices.
2	Initial	Organisation recognise practices but the implementation has not be focused.
3	Moderate	Organisation has become part of management practice but seem not considered as important to be implemented.
4	Nearly comprehensive	Organisation has a clear definition on practices and has overall implementation.
5	Strategic	Organisation has used and applied BIAPM components for strengthen the implementation strategic for organisation's future growth.

TABLE. IV. IMPLEMENTATION LEVEL FOR EACH ELEMENT AND SUB ELEMENT (CASE A)

Element	Sub Element	Average score	Score	Implementation Level
Process				
Requirement Analysis	• Requirement acquisition	5.00	5	Excellent
	• Data Security	2.50	3	Moderate
	• Software Identification	5.00	5	Excellent
	• Data Resource	5.00	5	Excellent
	• Story Board Requirement	5.00	5	Excellent
BI Implementation	• Design	5.00	5	Excellent
	• Development of BI	5.00	5	Excellent
	• Data Transfer	5.00	5	Excellent
	• Testing	5.00	5	Excellent
Data Management	• Data Source	5.00	5	Excellent
	• Criteria	5.00	5	Excellent
	• Standardisation	5.00	5	Excellent
	• Quality	4.25	4	Good
Decision Making	• Report Analysis	5.00	5	Excellent
	• Make Decision	2.75	3	Moderate
	• Implied Information	1.00	1	Not Implemented
Analytic	• Problem Analysis	1.00	1	Not Implemented
	• Analysis Design	1.00	1	Not Implemented
	• Data gathering	1.00	1	Not Implemented
	• Data Analysis	1.00	1	Not Implemented
	• Model development	1.00	1	Not Implemented
	• Quality Analysis	1.00	1	Not Implemented
Visualisation	• Display	4.75	5	Excellent
	• Story	1.00	1	Not Implemented
	• Self-Service BI	2.50	3	Moderate
Strategic Planning	• Vision & Mission	5.00	5	Excellent
	• Objective	5.00	5	Excellent
	• Success Factor analysis	1.00	1	Not Implemented
	• Strategy	5.00	5	Excellent
	• Action Plan	5.00	5	Excellent
	• KPI	5.00	5	Excellent
	• Target	5.00	5	Excellent
	• Distribute information	5.00	5	Excellent
Performance Evaluation	• Measurement	5.00	5	Excellent
	• Control	5.00	5	Excellent
Governance				
Documentation	• Business Definition	1.80	2	Need Improvement
	• User Manual	3.00	3	Moderate
	• Metadata documentation	3.00	3	Moderate
	• Operational Documentation	2.67	3	Moderate
	• Analysis Data Documentation	1.00	1	Not Implemented
Finance	• Source	4.20	4	Good
	• Budget	4.20	4	Good
	• Management	3.80	4	Good
Change Management	• Time management	5.00	5	Excellent
	• Data Scope & Addition	5.00	5	Excellent
	• BI improvement	5.00	5	Excellent
Software	• Database	5.00	5	Excellent
	• ETL	5.00	5	Excellent
	• BI tools	5.00	5	Excellent
	• Analysis Tools	1.25	1	Not Implemented
Hardware	• Server	5.00	5	Excellent
	• Network	5.00	5	Excellent
Practitioner				
Top Management		5.00	5	Excellent
BI Implementers		5.00	5	Excellent
Performance Manager		5.00	5	Excellent
Data Scientist		1.20	1	Not Implemented
Domain Expert		5.00	5	Excellent
Ability				
Skill	Training	3.00	3	Moderate
	Knowledge	3.40	3	Moderate
	Experience	3.20	3	Moderate
Work Culture	Motivation	4.80	5	Excellent
	Collaboration	5.00	5	Excellent
	Adaptability	4.80	5	Excellent
	Positive Attitude	5.00	5	Excellent

TABLE. V. AVERAGE SCORE OBTAINS BY EACH COMPONENTS (CASE A)

Component	**Average Score**	**Score**	**Implementation Level**
Process	3.81	4	Good
Governance	4.08	4	Good
Practitioner	4.24	4	Good
Ability	4.05	4	Good

Furthermore, the average score values for each components are computed using formula (3) and the results are shown in Table V.

C. Recommendation for Improvement

Based on results in Table V shows that Case A attains good level for the four components which are process, governance, people and ability. These components are referred to the components inclusive in our proposed BIAPM framework. The highest score is obtained by component people where it shows that this organisation has almost complete and essential implementers to support the BIAPM implementation. The lowest score is obtained by the process component. This is due to more works need to be done in term of processes for implementing the BIAPM integration in this organisation.

In summary, the overall score achieved by Case A in this exercise is 4.04. This score is computed using formula (4). The result shows that BIAPM implementation in organisation A fall in level 4 and this score value is mapped into Table III to gain the appropriate implementation level. In this case, it

obtains Nearly Comprehensive which shows that this organisation has a clear definition on practices and has the overall implementation plan. At the same time several elements need to be improved and put in place in integrated and strategically way.

VII. CONCLUSIONS

This paper has presented our proposed integrated framework of business intelligence and data analytics for organisational performance management or BIAPM. BIAPM comprises of 4 main components which are process, people, governance and ability. Each components are broken down into several elements and sub elements as discussed in our previous papers. This paper focuses on the implementation of BIAPM framework through real case study conducted in Malaysia. It was implemented collaboratively with a semi-government organisation which in this case was a public university. The implementation and application were carried-out according to the proposed framework and the analysis and results of this exercise are presented in this paper. As discussed in this paper, organisation of the case study obtained overall score as Nearly Comprehensive in term of their BIAPM implementation. Further improvement are suggested to the organisation. The case study implementation has revealed the suitability and practicality of the BIAPM framework.

Comparing with similar and available frameworks from literature, discovers that even though several studies have been done in the similar domain of BI, BA and OPM implementation but only a few frameworks that integrate the three fields as being focused in this research. One of them is Gartner's Business Analytics Framework (GBAF) [19]. This framework is aimed to be the reference for business intelligence, analytics and performance management implementation. However, Gartner's framework requires enhancement particularly in the perspective of managing performance in public sector and comprehensive BIAPM implementation. Besides GBAF, there are few more studies which relate to BIAPM and they can be identified as Pourshahid et al. [38], Cosic et al. [25], Martin et al. [22] and Wu [23]. Martin et al. and Wu considered two fields integrated which are BI and BA without integrating with OPM, while Pourshahid et al and Cosic considered BIA but without comprehensive implementation with OPM.

ACKNOWLEDGMENT

This research is funded in part by Malaysia Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/2015/ICT04/UKM/02/1) and PP-FTSM-2019.

REFERENCES

[1] N. Mirsepasi, A. Faghihi, and M. R. Babaei, "Design a System Model for Performance Management in the public sector," Arab. J. Bus. Manag. Rev., vol. 1, no. 4, pp. 23–32, 2013.

[2] D. Isaev, "Development of Performance Management Systems," in 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 168–171, 2011.

[3] M. Obeidat, M. North, R. Richardson, V. Rattanak, and S. North, "Business Intelligence Technology , Applications , and Trends," Int. Manag. Rev., vol. 11, no. 2, pp. 47–56, 2015.

[4] V. L. Sauter, Decision Support Systems for Business Intelligence. New Jersey: Wiley, 2010.

[5] C. Elena, "Business intelligence," J. Knowl. Manag. Econ. Inf. Technol., vol. 1, pp. 101, 2011.

[6] W. H. Inmon, Building the Data Warehouse, 5th Editio. John Wiley & Sons, 2005.

[7] S. Jou and R. Ng, "Introduction and the Changing Landscape of Business Intelligence," in Perspectives on Business Intelligence, Morgan & Claypool Publishers, pp. 1–3, 2013.

[8] L. Serbanescu, "Necessity to Implement a Business Intelligence Solution for the Management Optimization of a Company," USV Ann. Econ. Public Adm., vol. 12, no. 2, pp. 114–123, 2012.

[9] B. Gupta, M. Goul, and B. Dinter, "Business intelligence and big data in higher education: Status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs," Commun. Assoc. Inf. Syst., 2015.

[10] N. H. Z. Abai, J. Yahaya, A. Deraman, A. R. Hamdan, Z. Mansor, and Y. Y. Jusoh, "Integrating Business Intelligence and Analytics in Managing Public Sector Performance : An Empirical Study," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 9, no. 1, pp. 172–180, 2019.

[11] S. Moro, P. Cortez, and P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation," Expert Syst. Appl., 2015.

[12] M. Zamani, M. Maeen, and M. Haghparast, "Implementation of Business Intelligence to Increase the Effectiveness of Decision Making Process of Managers in Companies Providing Payment Services," J. Internet Bank. Commer. Ottawa, 2017.

[13] C. M. Olszak, "Business Intelligence and Analytics in Organizations," in Advances in ICT for Business, Industry and Public Sector, vol. 579, 2015.

[14] M. Bronzo, P. T. V. de Resende, M. P. V. de Oliveira, K. P. McCormack, P. R. de Sousa, and R. L. Ferreira, "Improving performance aligning business analytics with process orientation," Int. J. Inf. Manage., vol. 33, no. 2, pp. 300–307, Apr. 2013.

[15] R. S. Kaplan, "Strategic performance measurement and management in nonprofit organizations," Nonprofit Manag. Leadersh., vol. 11, no. 3, pp. 353–370, 2001.

[16] The BPM Standard Group, "Business Performance Managenent : Industry Framework Document," 2005.

[17] M. Tresnadi, I. Primiana, and D. Wibisono, "Analyzing Mainstreams In Current Performance Management Studies And Its Relationship With HR And Other Practices : A Literature Review," Sains Humanika, vol. 2, pp. 41–51, 2016.

[18] P. I. Bogdana, A. Felicia, and B. Delia, "The Role of Business Intelligence in Business Performance Management," Ann. Fac. Econ., vol. 4, no. 1, pp. 1025–1029, 2009.

[19] N. Chandler, B. Hostmann, N. Rayner, and G. Herschel, "Gartner ' s Business Analytics Framework," Gartner, pp. 1-18, 2011.

[20] M. I. Nofal and Z. M. Yusof, "Taxonomy framework of erp success usage in smes in middle east region," J. Theor. Appl. Inf. Technol., vol. 86, no. 3, pp. 420-431, 2016.

[21] R. Fitriana, J. Saragih, and N. Luthfiana, "Model business intelligence system design of quality products by using data mining in R Bakery Company," in IOP Conference Series: Materials Science and Engineering, 2017.

[22] A. Martin, T. Miranda Lakshmi, and V. P. Venkatesan, "A business intelligence framework for business performance using data mining techniques," 2012 Int. Conf. Emerg. Trends Sci. Eng. Technol., pp. 373–380, Dec. 2012.

[23] J.-Y. Wu, "Computational Intelligence-Based Intelligent Business Intelligence System: Concept and Framework," 2010 Second Int. Conf. Comput. Netw. Technol., pp. 334–338, 2010.

[24] G. Richards, W. Yeoh, A. Y. L. Chong, and A. Popovič, "Business Intelligence Effectiveness and Corporate Performance Management: An Empirical Analysis," Journal of Computer Information Systems , vol. 59, no. 2, pp. 188-196, 2019.

[25] R. Cosic, G. Shanks, and S. Maynard, "A business analytics capability framework," Aust. J. Inf. Syst., vol. 19, pp. 5–19, 2015.

[26] C. Holsapple, A. Lee-Post, and R. Pakath, “A unified foundation for business analytics,” Decis. Support Syst., vol. 64, pp. 130–141, Aug. 2014.

[27] N. Rayner and K. Schlegel, “Maturity Model Overview for Business Intelligence and Performance Management,” Gartner, Stamford, 2008.

[28] P. P. Dooley, Y. Levy, R. A. Hackney, and J. L. Parrish, “Critical Value Factors in Business Intelligence Systems Implementations,” In book: Analytics and Data Science, 2018.

[29] C. Sacu and M. Spruit, “BIDM : The Business Intelligence Development Model,” Technical Report 2010-010, Institute of Information and Computing Sciences, Utrecht University, pp.1-14, 2010.

[30] S. LaValle, “Breaking away with business analytics and optimization,” Executive Report, IBM Global Business Services, New York, pp. 1-13, 2009.

[31] P. Brooks, O. El-Gayar, and S. Sarnikar, “A framework for developing a domain specific business intelligence maturity model: Application to healthcare,” Int. J. Inf. Manage., vol. 35, no. 3, pp. 337–345, 2015.

[32] J. H. Yahaya, N. H. Zulkifli Abai, A. Deraman, A. R. Hamdan, Z. Mansor and Y. Y. Jusoh, “The integrated framework of business intelligence and data analytics for organizational performance management in public sector,” in Lecture Notes in Engineering and Computer Science, vol. 2235, 2018.

[33] E. Bonollo, S. Lazzini and M. Z. Merli, “Business Intelligence in Public Sector Organizations : A Case Study,” Lect. Notes Inf. Syst. Organ. Strength. Inf. Control Syst., vol. 14, pp. 199–216, 2016.

[34] V. B. Vukšić, M. P. Bach, and A. Popovič, “Supporting performance management with business process management and business intelligence: A case analysis of integration and orchestration,” Int. J. Inf. Manage., vol. 33, pp. 613–619, 2013.

[35] J. W. Creswell, Educational Research Planning, Conducting, and Evaluating Quantitative and Qualitative Research, 4th edition, Allison Wesley, 2011.

[36] Y. H. Kwak and C. William, “Assessing Project Management Maturity,” Proj. Manag. J., vol. 31, no. 1, pp. 32–43, 2000.

[37] M. Niazi, D. Wilson, and D. Zowghi, “A maturity model for the implementation of software process improvement: An empirical study,” J. Syst. Softw., vol. 74, no. 2 SPEC. ISS., pp. 155–172, 2005.

[38] A. Pourshahid, G. Richards and D. Amyot, "Toward a goal-oriented, business intelligence decision-making framework," E-Technologies: Transformation in a Connected World, pp. 100–115, 2011.

[39] E. M. Mohd Yusof, M.S. Othman, L. M. Yusuf, S. Raja Kumaran and A. R. Mohd Yusof, "A model of acceptance factors for business intelligence in manufacturing using theoretical models," Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no.3, pp. 1544-1551, 2019.

[40] S. Aldossari and U. A. Mokhtar, "Enterprise resource planning and business intelligence to enhance organizational performance in private sector of KSA: A preliminary review," The Third International Conference on Reliable Information and Communication Technology, pp. 343-352, 2018.

A Survey on Distributed Greenhouse Gases Monitoring Systems

Adela Puscasiu[1], Alexandra Fanca[2], Dan-Ioan Gota[3], Silviu Folea[4], Honoriu Valean[5]
Department of Automation, Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—Monitoring of air quality represents a major task, due to the direct impact of pollution on human health. Pollution has been further aggravated by the progresses that have taken place in the last decades: traffic growth, traffic noise in cities and growth of urban areas, rising cities, increased energy consumption, industrialization, and economic development. Global warming and acid rain are the results of these factors; thus, air quality is essential to be monitored. This survey presents a set of researches and applications related to air quality monitoring, aimed to detect, measure, collect and process data aggregated from sensors, such as gas sensors for sensing concentration of gases such as CO_2 (Carbon dioxide), relative humidity, temperature, TVOC (Total Volatile Organic Compounds), PM (particulate matter) and noise level. Based on some processes, users will be able to see the polluted areas on the map. The paper presents a state of the art of air monitoring systems, noise monitoring systems, air pollution systems. Also, the paper proposes a distributed greenhouse monitoring system for pollution measurement and control.**

Keywords*—*Air quality; air monitoring systems; noise monitoring systems; air pollution systems

I. Introduction

Indoor and outdoor air quality represents one of the major problems in the world. Indoor air quality is a very important issue in our life because we spend over 80% of time inside. Obviously, outdoor air quality is even important. So, this is the reason to understand that we inspire about 12000 liters of air every day, and we are exposed to dust, viruses, smoke, pollen, bacteria and other compounds present in the air. It is important to pay attention to the place and time we spend in the locations where symptoms occur.

The results of more than 154 studies carried out in 37 countries by the World Health Organization [1] show that indoor air is eight times more polluted than outside.

Nowadays, it can be seen as a continuous evolution of technology, intense traffic, rapid urbanization that have a significant impact on society as it contributes to the increase in the amount of polluting emissions. Thus, air quality must be monitored permanently, and any unusual values should be corrected with the optimal solution.

In the last decades, a set of organizations, such as EEA [2], EPA [3] in the USA, and NAQMN [4] in Romania, has been created with the aim to monitor and protect the environment. These organizations have set up their own network of stations in order to monitor the characteristics of the air (concentrations of the main air pollutants: nitrogen oxides (NO_x), sulfur dioxide (SO_2), tropospheric ozone, carbon monoxide (CO), particulate matter (PM_{10} and $PM_{2.5}$), polycyclic aromatic hydrocarbons (PAHs), heavy metals such as cadmium (Cd), lead (Pb), nickel (Ni), mercury (Hg) and arsenic(As)). Usually the results of monitoring processes are presented on online platform. According to the characteristic air measurements [5] carried out by these stations, the monitored parameters can be divided into two categories:

- Physical: temperature, humidity, air pressure, wind direction, and speed;
- Chemicals: air pollutants such as ozone (O3), NO_2, CO, SO_2, PM_{10}, $PM_{2.5,}$ and CO_2.

Chemical components are considered by almost all the environmental organizations to be the most harmful to health, due to their respiratory symptoms, decreased lung function, lung inflammation, possible lesions of the lungs and premature mortality.

Fine particles, or particles in suspension, are a mixture of solid particles and liquid droplets found in the air. Some particles can be seen with the naked eye: dust, dirt, smoke, soot, and others are so small that they can only be seen with the microscope. These particles are divided into two categories:

- $PM_{2.5}$, fine inhalable particles; their diameter does not exceed 2.5 micrometers;
- PM_{10}, inhalable particles; their diameter does not exceed 10 micrometers.

There are a lot of sources for PM. Some examples of such sources are presented below:

- Indoor: gas and oil heating, air conditioning – especially if the pipes are dirty, fireplaces – it should be well ventilated, furniture – pillows, beds, chairs, wood-burning stoves - toxins resulting from wood-burning;
- Outdoor: fuel emissions, near fires, changes in atmospheric conditions, the outside air, factories, and power plants.

Certain effects may occur shortly after exposure to a pollutant and may be treated (irritation of the eyes, nose, throat, fatigue, headache, dizziness) but unfortunately, other effects occur after longer periods of time and may be fatal (respiratory diseases, heart disease, cancer). Understanding and controlling common indoor pollutants can help reduce the risk of illness.

A major issue of using extensively vehicles with internal combustion engines is the emission of harmful gases. Polluted air is an important problem that affects human health, but it is usually ignored. The measure of contribution to CO_2 and greenhouse effect gasses emission (GHG) and implicitly to air pollution is a very important task. Excessive pollution as well as dust pollution compromise health and lead to serious respiratory and cardiovascular problems. Some elements that lead to excessive pollution of the atmosphere are:

- Intense and crowded traffic: insufficient measures to streamline traffic in cities and lacks in checking emissions and imposing sanctions.
- Industrial pollution: the absence of rigor in industry surveillance and slow upgrading in state-run units.
- Construction deficiency: prolonged infrastructure works without justification and non-application of environmental legislation to public and private yards.
- Deficiency of monitoring: despite some investments, insufficient or inadequate monitoring, display and warning systems.
- Organizational deficiency: the absence of interest, failure of legislation.
- Deficiency of transparency: issues of access to relevant information and communication deficiencies on behalf of the responsible authorities.

The importance of this survey was reflected in the improvement of the air quality and measures taken for reducing the risk of population illness and the efficiency of the urban traffic control in order to reduce the carbon footprint.

The rest of the paper is organized as follows: Section II presents air quality systems and advantages and disadvantages of them; Section III presents the proposed system and the conclusions are presented in Section IV.

II. Related Work

The paper presents a review of the literature based on greenhouse gas emissions. A set of measuring and monitoring air quality systems are taking into account. These systems can be divided into three categories:

A. Air Pollution Systems

IoT (Internet of Things) is a new concept, which interconnects via Internet all objects (mobile devices, portable devices, lighting systems). In fact, IoT is a dynamic global information network consisting of Internet-connected objects, such as radiofrequency identifications, actuators, sensors, as well as smart appliances that are becoming an integrated component of the future Internet. Thus, there are also many devices and applications that measure different pollutants which can be considered as parts of IoT, such as:

- Foobot [6] determines the air quality in real-time and uses the following sensors: $PM_{2.5}$, TVOC, Humidity, and Temperature. It measures VOC, PM2.5, and CO_2 (derived from VOC), temperature and humidity. Available for both iOS (8+) and Android (4+). It is compatible with 2.4GHz Wi-Fi.
- Air Mentor 6 in 1 [7] Indoor pollutant concentration monitor, can detect CO_2, VOC concentrations with CO, $PM_{2.5}$ and PM_{10}, Relative Humidity (RH) and temperature. Easy to connect via Bluetooth – and it is available for both iOS and Android devices. It is portable, and it can also be used in an outdoor environment.
- NETATMO [8] has the following sensors: humidity, air quality, noise, and temperature. Available for both iOS and Android devices.
- Speck [9] detects fine particles from indoor environments, provides information on PM concentration changes, Integrated Wi-Fi, continuously uploading data, users can view real-time data, as well as historical representation, using a computer or mobile device.

The paper presents a solution for collecting, displaying and distributing environmental data based on mobile phones and portable devices.

B. Noise Monitoring Systems

In [10] a CyberGIS (a new generation of Geographic Information System) designed to facilitate volunteer participation in monitoring urban pollution using mobile devices is described. Cybernetic infrastructure and toolkits facilitate the development of applications that require access to computing and distributed resources. This framework allows scalable data management, analysis and viewing for data collected from mobile devices. To demonstrate its functionality, it focuses on noise mapping. This framework integrates a MongoDB2 cluster for data storage, a MapReduce for extracting and aggregating noise data collected and loaded by mobile devices and a parallel kernel smoothing algorithm using the GPU (Graphics Processing Unit) to efficiently create maps noise pollution from the massive collection of recordings. This framework also implements a mobile application for capturing noise and time measurements as users move around in the urban environment.

The map is dynamically generated using the "kernel smoothing" method. This method is used to estimate a continuous surface (e.g. Noise Level) at the observation points. By processing this data, noise mapping is performed along the trajectory of an individual participant.

Author in [11] makes an approach to automatically creating input data for noise simulations and focuses on using 3D methods to present the results of noise simulation caused by road traffic and industrial activities in urban areas. Several standard noise models for industrial and road noise in CityGML have been deployed by expanding NoiseADE (Application Domain Extension) with new objects and attributes in order to standardize input and output data for noise studies and 3D data reconstruction.

The models allow the computing of the level for:

- the noise produced by road traffic: it is necessary to know certain information such as location, building height, traffic flow, percentage distribution of vehicles on the streets (light/heavy vehicles), speed limits, asphalt type, noise barrier information.
- industrial noise: we need to know the type of noise source, operating hours of cars, noise barriers, location of cars, etc.

Author in [12] presents a system for the protection of users' privacy, which is based on cryptographic techniques and cloud computing. There is a general architecture of the system including a cryptographic protocol based on a homographic encryption scheme for mapping data into maps. Noise Tube is a participatory framework for noise pollution, featuring experiments with real and artificial data sets and a demo on a heterogeneous set of commercial cloud providers. Noise Tube allows the creation of noise maps generated from data collected by multiple users without revealing their location.

Key features of NoiseTubePrime include:

- Correctness: Accurate aggregate statistics are calculated using each user's private measurement data while maintaining the confidentiality of the participating users - there is no location/time data that is revealed.
- Cloud services: The outsourcing of NoiseTubePrime to the cloud frees the users from trouble to run their own software agent and maintain permanent access to the Internet. The computing and networking requirements of each software agent are low and are (currently) provided at no cost to the various cloud service.
- Decentralization: The main task of NoiseTube is to decrypt the finalized map of encrypted noise which is the result of distributed cloud computing. Thus, the central workload is much smaller than a scenario in which the entire calculation is performed by a single central server. Therefore, the NoiseTube service calculation is independent of the number of users participating, making NoiseTubePrime a decentralized system theoretically that can be scaled to treat a large number of users.

NoiseTube [13] is a software platform that uses a mobile application and a website. The mobile application uses the combination of the microphone and the built-in GPS receiver to monitor the sound in different locations of a city. The site allows users to see the noise measurements sent by users.

Sound Meter [14] shows the decibel values by measuring the ambient noise and displaying their values in different forms. This application uses the phone's microphone to measure dB and to display the recorded values.

C. Air Monitoring Systems

AirCasting [15] is an open-source solution for collecting, displaying and sharing health and environment data using a smartphone. The platform is made up of portable sensors that detect environmental shifts, including an air quality monitor, called AirBeam, an Android application called AirCasting and portable LED accessories.

AirBeam uses a light dispersion method to measure fine particles or $PM_{2.5.}$ Air is pulled through a detection chamber where the light from an LED bulb shakes particles from the airflow. This light dispersion is recorded by a detector and transformed into a measurement that estimates the number of particles in the air. This is important because the US Environmental Protection Agency monitors and control six air pollutant criteria, one of which is $PM_{2.5.}$ EPA measurements indicate that $PM_{2.5}$ levels represent substantial health risks in cities across the country.

AirBeam measurements are sent once a second to AirCasting Android via Bluetooth, maps and graphs are being displayed in real-time on the smartphone. At the end of each AirCasting session, the collected data is sent to the AirCasting site, where data is collected from different AirCasters in order to generate thermal maps indicating where the $PM_{2.5}$ concentrations are highest and lowest.

AirCasting Luminescent devices connect to AirCasting via Bluetooth and are designed to illuminate the LEDs in response to AirCasting sensor measurements: green for low intensity, then yellow, then orange and high-intensity red. So far models for two luminescent accessories are released: AirCasting Luminescent Vest and LiteBeam. AirCasting Luminescence has been developed to communicate sensor measurements without normally referring to a screen interface and triggering the interaction between AirCasters and people in their immediate proximity.

AirBeam is just an instrument among dozens of people able to transmit data to the AirCasting platform. There is already a flourishing ecosystem of AirCasting compatible tools - some created by manufacturers, others by consumer electronics companies - that can connect to the AirCasting Android application to record health and environment data, including but not limited to:

- Sound levels recorded by the phone microphone;
- Temperature, humidity, concentrations of CO and NO_2 recorded by the air monitor;
- Cardiac rhythm, heart rate variability, R to R, breathing rate, activity level, peak acceleration and core temperatures recorded by Zephyr BioHarness 3;
- Measurements of heart rate recorded by Zephyr HxM.

AirVisual [16] provides real-time data on air pollution and weather meteorological data for more than 6,000 cities around the world, by live monitoring of six pollutants: real-time monitoring of concentrations of $PM_{2.5}$, PM_{10}, ozone, NO_2, $SO_{2,}$ and CO. The application allows users to view air quality in over 9,000 locations worldwide.

SmartAir PM2.5 developed by Thermo Fisher Scientific is the application for pDR-1500 (Personal Aerosol Monitoring). The pDR-1500 [17] is a very sensitive non-photometric (photometric) device whose light diffusion detection configuration has been optimized to measure the respirable dust fraction in air, smoke, and vapor. It measures $PM_{2.5}$ in

real-time and can compare the results with outdoor air. It also presents a chart of intuitive lines in 24 hours/week/month representation. Computer requirements: Compatible PC, processor 486 or higher; Windows ™ 95 or later.

D. Discussions

As a result of the study carried out, can be noticed that there are mobile devices and applications for measuring, monitoring and interpreting indoor pollutant data. A summary of the systems is presented in the following Table I.

From the energy consumption point of view, the optimal system is AirMentor, because it is the only one using Bluetooth Low Energy (BLE), which is a portable device and can be used indoor and outdoor, but a disadvantage is that it does not measure noise. From the measured parameters point of view, the best system is Speck, and in terms of reduced dimensions the best system can be considered to be Netatmo.

In conclusion, none of the presented indoor systems can cover all the parameters required (including the noise sensor). Also, none of the presented systems have small size and only AirMentor uses BLE.

Outdoor pollutant measurement applications are presented in the Table II.

TABLE. I. INDOOR SYSTEMS

	Summary of indoor systems			
	Foobot	***AirMentor***	***NETATMO***	***Speck***
Size [inch]	6.69x2.75	4.2x1.8x4.5	1.77x1.77x6.1	4.5x3.5x3.7
Noise			x	
Temperature	x	x	x	x
Relative humidity	x	x	x	x
$PM_{2.5}$	x	x		x
PM_{10}				x
VOC	x	x		
Power		5-1 A		5V
Wi-fi	x		x	x
Bluetooth		BLE		

TABLE. II. OUTDOOR SYSTEMS

		Summary of outdoor systems				
		Air Mentor	***AirVisual***	***Smart Air***	***Noise Tube***	***Air Beam***
Weather Forecast			x			
Noise					x	
Temperature		x	x			x
Relative humidity		x	x			x
CO		x	x			x
CO2		x	x			
SO2			x			
NO2						x
PM	2.5	x		x		x
	10	x				
Bluetooth		BLE				BLE

Among the outdoor systems, the better system is AirMentor, as it covers a large part of the required parameters and uses BLE. Compared with AirBeam, AirMentor measures PM_{10} and CO_2.

The system proposed in the last paragraph of the paper cover all the air quality parameters that use BLE and can be used in indoor and outdoor environments.

III. PROPOSED IDEA

Based on the research presented in Section II, the paper proposes a system which is divided into two subsystems:

A. Indoor

Compared with the previously presented systems, the proposed system uses the beacon shown in [18] and noise measurement sensors are added. The client-server application will allow notifying the user when measured values exceed the normal limits.

B. Outdoor

The main objective of the proposed system is to minimize the carbon footprint. Certain areas have increased pollution every day, at the same time interval (morning between 7:30 and 10:00 when people go to work and in the evening between 15:00 and 20:00 when people return home). Therefore, the GHG concentration depends on the number and speed of the cars traveling through the targeted areas. If the system achieves its goal, it will lead to reducing greenhouse gas emissions and will decrease the negative effects on the population health. An important asset of this system is a solution to decongest traffic in large cities, especially in peak hours when traffic is intense and, moreover, in areas where high-rise buildings are located because they prevent gas dispersion. Using this solution, decongesting traffic directly leads to the minimization of greenhouse gases. A description of the proposed system functionality can be seen in Fig. 1.

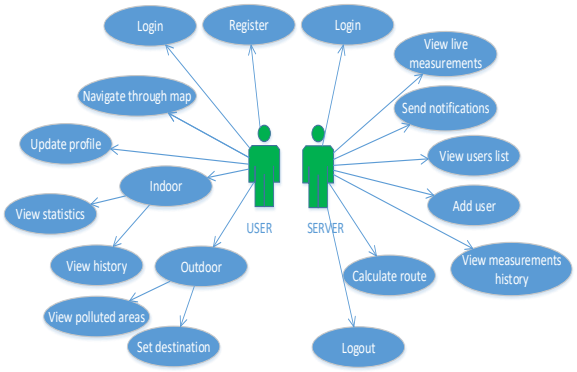


Fig. 1. Use Case Diagram.

The system has a client-server architecture. On the client-side, a mobile application allows the user to do the following actions: register to the application, login, navigate through map, update profile. The user can select indoor display, which shows the statistics and history of measured data or outdoor display, in order to see the polluted areas. Also, if the user wants to travel from the current position to another, the destination can be set. For the server-side, a web application with the following features is implemented: login, view live

measurements, send notification (the server will send a warning message to the user when certain measured parameter values exceed the threshold limits), view user list (all users are stored in the database), add user, view measurements history and calculate route (depending on the destination chosen by the user, the server will calculate an optimal route so as to avoid areas with instantaneous measurements of high pollution).

Thus, the traffic in a city or in a specified geographic area can be monitored in real-time, reported to the central server and controlled, in order to reduce the pollution if it increases over some accepted values. Control of traffic avoids the development of highly polluted areas and decreases the greenhouse gas footprint and the noise intensity to acceptable levels. That extends and makes more feasible and useful traffic control with respect to the pollution level compared to the existing systems which use fixed sensors buried in the tread, already implemented on some highways (such as IGL control in Austria).

IV. Conclusions

The main objective of this survey was to emphasize a state of the art in air quality domain and to propose a new system for monitoring a control air quality. Some indoor and outdoor monitoring systems were presented. Also, a comparison was made between the air quality systems, as can be seen from Table I and Table II.

Based on this research, the best-found system is AirMentor, but the paper proposes a better solution that uses smaller dimensions beacon compared with AirMentor. After adding the noise sensor, a complete system can be developed.

The vast technological developments in wireless communication technology have led to the emergence of many pollution monitoring sensors and wireless networks for monitoring and reporting pollution. This information could be used to take necessary action such as emergency warning messages and evacuation of people from problematic spaces.

These systems are designed to measure the parameters of pollutants that can affect people's health, especially in small spaces.

Due to the increasingly crowded traffic and the carbon dioxide emitted by vehicles that have internal combustion engines and emit harmful gases for our health, we need to take measures to streamline city traffic.

The second factor of pollution is traffic noise. The effects of exposure to motors sound are stress, insomnia, cardiovascular disease, and stroke.

References

[1] World Health Organization's studies (2015). Available: http://www.euro.who.int/en/data-and-evidence/databases.

[2] European Environmental Agency (2017). Available: http://www.eea.europa.eu/ .

[3] United States Environmental Protection Agency (2017). Available: https://www.epa.gov/.

[4] National Air Quality Monitoring Network (2014). Available: http://www.calitateaer.ro/public/home-page/?__locale=en.

[5] Environmental Protection Agency. Air Quality Index Reporting; Federal Register/Vol. 64, No. 149/Wednesday, August 4, 1999/Rules and Regulations, pag.42548.

[6] Sinan Sousana, Kirsten Koehlerb, Laura Halletta, Thomas M. Peters, "Evaluation of consumer monitors to measure particulate matter", Journal of Aerosol Science 107, February 2017.

[7] The Jen Reviews Team, "Best Air Quality Monitor" [online]. Available: https://www.jenreviews.com/best-air-quality-monitor/#comments.

[8] Dominic Worner, Thomas von Bomhard, Marc Roschlin, Felix Wortmann, "Look Twice: Uncover Hidden Information in Room Climate Sensor Data", IEEE Conference 2015, Cambridge, MA, USA, 05 February 2015.

[9] Ron Williams, Amanda Kaufman, Tim Hanley, Joann Rice, Sam Garvey, "Evaluation of Elm and Speck Sensors", EPA, United States Environmental Protection Agency, Report, EPA/600/R-15/314 November 2015.

[10] Junjun Yin, Yizhao Gao, and Shaowen Wang, "CyberGIS-Enabled Urban Sensing from Volunteered Citizen Participation Using Mobile Devices", Seeing Cities Through Big Data, pp.83-96, , October 2017, DOI 10.1007/978-3-319-40902-3_5.

[11] K. Kumar , H. Ledoux, T.J.F. Commandeur, J. E. Stoter, "Modelling Urban Noise in Citygml Ade: Case of the Netherlands", ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-4/W5, 2017 12th 3D Geoinfo Conference 2017, 26–27 October 2017, Melbourne, Australia.

[12] E.D'Hondt, M.Stevens, A.Jacobs, "Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring", Pervasive and Mobile Computing, Volume 9, 2013, pages. 681-694.

[13] George Drosatos, Pavlos S. Efraimidis, Ioannis N. Athanasiadis, Matthias Stevens, Ellie D'Hondt , " Privacy-preserving computation of participatory noise maps in the cloud", The Journal of Systems and Software 92, 2014, pages 170-183.

[14] Ibekwe T.S., Folorunsho D.O., Dahilo E.A, Gbujie I.O., Nwegbu M.M., Nwaorgu O.G., "Evaluation of moile smartphone app as a screening tool for environmental noise monitoring", Journal of Occupational and Environmental Hygiene, Volume 13, Issue 2, 2016, pages D31-D36.

[15] S.Gao, R.W.Hilts, M.S.Ross, S.A.Styler, "Particulate matters: student-led air quality research in the third-year environmental chemistry classroom and the field ", The Journal of Analytical and Bioanalytical Chemistry, Volume 410, Issue 14, 2018, pp 3223-3229.

[16] PRNewswire, AirVisual Creates Stunning 3D Air Pollution Map-Allows Developerws Access to the Word's Most Compreensive Air Quality Data Set (2016). Available: https://www. prnewswire.com/ news-releases/airvisual-creates-stunning-3d-air-pollution-map---allows-developers-access-to-the-worlds-most-comprehensive-air-quality-data-set-300384537.html.

[17] T.M. Kyeong, P. Lundrigan, K. Sward, S.C. Collingwood, N. Patwari. Smart Home Air Filtering System: A Randomized Controlled Trial for Performance Evaluation. Journal of Smart Health. 9-10. 10.1016/j.smhl.2018.07.009, 2018.

[18] A.Puscasiu, S.Folea, H.Valean, A.Fanca, T. Sanislav, "Monitoring the on-site contribution to the greenhouse effect by distributed measurement of carbon dioxide",International Carpathian Control Conference, 2017.

The Development of a Visual Output Approach for Programming via the Application of Cognitive Load Theory and Constructivism

Marini Abu Bakar[1], Muriati Mukhtar[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Fariza Khalid[3]
Faculty of Education
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—Programming is a skill of the future. However, decades of experience and research had indicated that the teaching and learning of programming are full of problems and challenges. As such educators and researchers are always on the look-out for suitable approaches and paradigms that can be adopted for the teaching and learning of programming. In this article, it is proposed that a visual output approach is suitable based on the current millennials affinities for graphics and visuals. The proposed VJava Module is developed via the application of two main learning theories, which are, the cognitive load theory and constructivism. There are two submodules which consist of eight chapters that cover the topics Introduction to Programming and Java, Object Using Turtle Graphics, Input and Output, Repetition Structure, Selection Structure, More Repetition Structures, Nested Loops and Arrays. To enable Java programs to produce graphical and animated outputs, the MJava library was developed and integrated into this module. The module is validated by three Java programming experts and an instructional design expert on the module content, design and usability aspects.**

***Keywords*—Introductory programming; CS1; novices; Java programming; learning; objects-first**

I. Introduction

Programming education research has been going on for over five decades. Teaching and learning of programming have continuously drawn the attention of researchers among academics. Many studies are conducted at respective institutions including studies on the programming languages used, curriculum aspects, teaching and learning approaches as well as supporting materials and software tools. Most researchers around the world agree that teaching and learning of programming are difficult for novice students as well as for teachers [1]–[5]. Hence, many institutions have taken measures to address this problem to motivate, enhance students' interest, skills and competitiveness in programming.

The main challenges that novice students face in learning programming are related to problem solving [6] , understanding programming concept [7], programming language syntax [8] and motivation [9], [10]. While the challenges that teachers face are the need for appropriate teaching methods and tools [10].

Starting with a review of existing approaches to teaching and learning of programming, this paper proposes a new approach that addresses three key issues. First is the current generation of students who prefer visual approaches; second is the nature of programming courses that cause students, especially novice students, to experience high cognitive load in writing programs [11]–[13]; and third is the nature of programming courses that require active participation of students in building their own knowledge based on existing knowledge [14]–[17]. Thus, this new approach, presented in the form of teaching modules, was built using visual elements and based on two main theories, cognitive load theory and constructivism.

This paper is organized as follows: Section II presents previous studies on teaching of programming approaches. Section III describes the learning theory applied in this study. Section IV details the result and discussion of the study. Section V describes the conclusion and further work.

II. Teaching of Programming Approaches

Many previous studies have discussed the approaches used in the teaching of programming. In addition to the traditional approaches, the commonly used approaches are visual programming, graphical and animation library and object-first.

A. Traditional Approach

Generally, the traditional approach to computer programming courses follows closely the order of the topics in most textbooks. The first section covers the topics of introduction, data types, assignment statements, arithmetic expressions, input/output followed by three basic programming structure namely sequence, selection and repetition structure. The second part of the course covers advanced topics such as arrays, strings, methods and classes. Students need to apply the basic concepts in the first section to solve the problems presented in the second section.

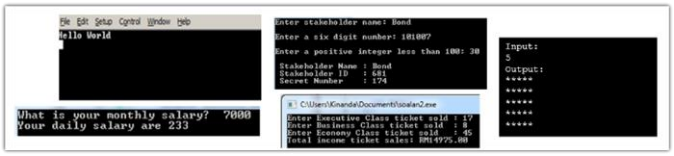


Fig. 1. Program Output in Text Mode.

In the traditional approach input and output are textual in which the students' program gets input from the keyboard and displays the output on the screen in text mode. Fig. 1 shows some examples of the output in text mode. Input and output in text mode were common in the early 1990s as most of the computers available in the programming lab were in text mode and the use of personal computers and laptops was quite limited. This traditional approach to output in text mode is less appealing to today's digital native students who are familiar with the latest computers and gadgets with graphical interfaces and touch screens.

Almost all institutions of higher learning in Malaysia adopt a traditional approach in teaching programming courses.

B. Visual Programming Approach

Among the popular visual programming language environment are Scratch and Alice which is widely used in primary and secondary schools [3], [18], [19]. In the visual programming language environments, programs are created by manipulating graphical components rather than writing textually. Creating programs is easier as there is no compilation errors and students are not required to know the syntax of a specific programming language.

Scratch was developed by Lifelong Kindergarten, MIT Media Lab in 2007 and is designed for students ages 8 and older. Author in [20] reports that students learning Scratch during primary school will easily learn advanced topics in high school. They don't have much trouble learning new topics and can reach a higher level of understanding for most basic concepts. As a result, some students choose to take programming courses in higher education. Students are also seen to have a high level of motivation and self-confidence.

Alice is a programming environment developed by the researchers in Carnegie Mellon. Alice provides an environment where students are able to drag-and-drop objects to create animations in three-dimensional. C++/Java programs are generated automatically. In higher institution Alice is usually used as a course in parallel with Computer Science 1 (CS1) or Computer Science 0 (CS0) courses [21]. CS1 generally refers to the first computer programming course in the computer science programme while CS0 is the programming related course at the pre-university level. From this study it was found that students taking Alice courses are better compared to those who do not take Alice courses in CS1. In this case, Alice is an additional course rather than being used extensively in the CS1 programming course.

C. Graphics and Animation Library Approach

Most graphics and animation library approaches are derived from LOGO. The LOGO programming language was introduced in 1967 by Seymour Papert with several researchers at the Massachusetts Institute of Technology [22]. LOGO is a language that teaches kids the basics of computer programming. In LOGO, the turtle is a cursor that can be controlled and operated according to the simple instructions given. Lines are drawn according to the movement of the turtle cursor. When it was first introduced in the 1960s, the LOGO language commands controlled turtle-shaped physical cursors. Later, with the technological advancements of the computer display screen, the turtle is represented as a cursor on the screen as in today's computer.

Graphics libraries have long been used for teaching of programming in universities worldwide. In [23] from Standford University initially developed the Turtles graphics library in ANSI C and subsequently translated it into Java. He encountered some problems in the implementation of the library because the earlier version of Java introduced in 1995 was relatively unstable.

Author in [22] developed Turtles package for use in teaching CS1 courses using Java at the University of Aarhus, Denmark. He uses the inverted curriculum approach proposed by [24]. With this approach important topics and concepts are introduced first and the details are explained later. The Turtles package has been used in all topics in the course. They reported positive effects, but no analysis was performed to show its effectiveness. Author in [9] also use turtle graphics in programming courses for prospective teachers. They report that this approach increases student motivation.

Another very similar approach is Karel the Robot in which the character is a robot named Karel in a simple world represented as a grid indicating streets and avenues. The Karel the Robot character can move one step forward, turn left, place and collect an object known as a beeper [25]. Author in [25] has used the Karel the Robot in programming courses at the University of Waterloo. Karel the Robot was first developed in 1981 by Richard Pattis to introduce Structured Programming courses using Pascal Language. Another popular variations of Karel are Robots is Karel J Robot [26] which are widely used in teaching basic programming courses using Java.

D. Objects-First Approach

Object-oriented programming is a widely used programming paradigm in both industry and education [27]. Almost all universities have object-oriented programming courses in the curriculum. Object-oriented programming is initially considered as an advanced course and is included in the middle or at the end of the curriculum. This situation has changed and today many universities have introduced object-oriented programming as their first course of programming. Among the object-oriented programming languages are Java, C++, C#, Eiffel, Python and Ruby.

The objects-first approach which was introduced in the ACM Computing Curricula 2001 document emphasized that the principles of object-oriented programming are introduced from the very beginning. The strategy was to begin introducing the concept of objects and classes and then followed by the structure of control, repetition and subsequent topics as in the traditional method.

Many studies highlight that teaching the basic concepts of object-oriented programming is difficult [27]–[30]. This is because many different concepts need to be understood as well as the skills that must be learned before students can write the program. Author in [27] emphasizes that object-oriented teaching is best used as an object-first approach compared to the object-later approach, by starting with a procedural programming approach and then switching to

objects. This is because the transition from procedural to object is more difficult compared to the difficulty of learning from the beginning. However, many textbooks use the object-later approaches.

The traditional approach to programming courses that produce text output is less appealing to today's generation Z students who are more inclined to visual learning styles. However, popular visual programming approaches such as Scratch and Alice are not suitable for higher education particularly for Computer Science students. Therefore, a new approach to the generation Z student learning style is needed. The proposed module named VJava Module uses the graphical and animation library and objects-first approaches. This module allows students to write Java programs textually that uses objects from the developed MJava graphics library to produce visual outputs in the form of graphics and animations. This VJava Module aims to increase students' interest and reduce anxiety at the beginning of the programming course which is perceived as difficult. To develop the VJava Module, two learning theories are applied; the cognitive load theory and constructivism theory. Both theories are discussed in the following sections.

III. Learning Theory

Learning theory describes how knowledge is absorbed, processed, and stored during learning. Learning theories need to be taken into account in designing the module to make them more effective and to achieve the objectives. This section discusses the learning theory applied to develop the VJava Module and the methodology of the research.

A. Cognitive Load Theory

Cognitive load theory was introduced by Sweller [31], [32] in the 1980s as a study of problem solving. It emphasised that all information is processed in a working memory and then stored in long-term memory for later use. Cognitive load theory basic premise is that the capacity of working memory is limited and can only process some information in a short duration of time.

Computer programming is a skill based course that is difficult and challenging which places a heavy cognitive load on the learners. Learning will be restrained by limited information processing capacity. If a learning task or activity requires cognitive capacity beyond its limits, that learning will be hindered [33].

There are three types of cognitive load which are intrinsic, extraneous and germane [32]. Internal cognitive load is related to the complexity of learning material and existing knowledge. Someone expert in programming and have extensive knowledge will learn easily compare to a student who has no direct knowledge of programming. This means that internal cognitive loads cannot be modified through instructional design [33].

Extraneous cognitive load is related to the design of teaching materials which can be modified by organizing the content of the materials. Novice students frequently use the means-ends analysis strategy in solving problems in the problem-solving approach [32]. Using this strategy will result in high usage of working memory resources resulting in a lack of existing cognitive resources. This causes cognitive activity to fail in working memory and thus impedes learning.

Germane cognitive load is an important cognitive load to explore in this study. This load refers to the construction of subsequent schemes as the primary goal of learning [32]. For example, giving students an example to solve a problem will help them understand the important steps in solving the problem and subsequently develop the problem solving scheme. The instructional design should guide the students to develop a scheme to increase the germane cognitive load.

The relationship between intrinsic, extraneous and germane cognitive load can be seen in the following three situations: (1) For situations where intrinsic cognitive load is low (easy learning content), and sufficient memory resources, students will be able to perform the learning process despite the high extraneous cognitive load (poor presentation of teaching material); (2) In situations where high intrinsic cognitive load (difficult teaching content) and high extraneous cognitive load, the cognitive load overcome mental resources and learning processes may fail; (3) Situations in which the external cognitive load in (2) is reduced, and the germane cognitive load is enhanced to facilitate the learning process [34].

Intrinsic cognitive load cannot be changed with the design of teaching materials. To produce meaningful learning, the design of instructional materials should reduce extraneous cognitive load and nurture germane cognitive load. This is because the extraneous cognitive load does not have a positive effect on the learning process, in contrast to the germane cognitive load that can help to improve the learning process.

B. Constructivism Learning Theory

The constructivism learning theory pioneered by Jean Piaget is based on the premise that knowledge is built by a person as a result of his mental activity rather than being conveyed by an instructor. Learning happened by interpreting the meaning of a concept based on existing knowledge and experience [14], [15]. Teachers encourage students to explore how an activity helps them to understand a concept. Constructivist teachers provide learning environments based on problems that need to be solved individually or collaboratively, while students produce their meaningful artifacts [15]. Learning occurs actively in solving problems with teachers acting as facilitators in nurturing meaningful learning.

Constructivism does not deny the role of lecturers or knowledge expert. Constructivism has changed that role, so lecturers helps students to build knowledge instead of just presenting facts. Constructivist lecturers provide tools such as problem solving and inquiry-based learning activities, sharing experiences, discussions, creating concept maps and building a broader picture of concept [35]. Students formulate and test their ideas, draw inferences and conclusions, and integrate their knowledge in a collaborative learning environment. Constructivism transforms students from passive recipients to active participants in the learning process. Guided by lecturers, students actively build their knowledge rather than

receiving knowledge directly from a lecturer or textbook. With a well-designed classroom environment, students will learn how to learn.

One face of Constructivism learning theory is the Constructionism introduced by Seymour Papert [36], [37] that asserted learning occurs "especially felicitously" when learners engaged in constructing artifacts. Papert introduced constructionism in association with LOGO, a programming language designed to enable the study of abstract concepts in mathematics, geometry, physics and others by manipulating computational objects [38]. The most common artifacts in constructionism today are in digital form.

C. Methodology

The VJava Module is developed using ADDIE, the five-phase development methodology [39]. The first phase is the analysis phase which comprises of literature review and preliminary study to determine the problems faced in the learning of programming, determine the appropriate approach and relevant learning theories to apply in the proposed method. The second phase is the design phase, which involves the design of the graphics library and the learning module. This is followed by the third phase, which is the development phase to develop the module. At the end of this phase, the process of verification of the module by the experts is conducted. The next phase is the implementation phase where the updated modules based on expert reviews are tested in the pilot study before being implemented in the actual learning environment. The final phase is the evaluation phase on the students' response after learning using this module. This paper discusses the results of a study that covers phases one to three.

IV. Results and Discussions

This study applies two learning theories, namely cognitive load theory and constructivism in designing and developing VJava Modules for basic programming course. The VJava Module consists of two submodules; the VJava Module I consists of five chapters and the VJava Module II that consists of three chapters. This module uses the MJava graphics library to produce graphics and animated outputs.

A. MJava Library

The MJava library consists of two Java packages, the MTurtle package and the MGraphics package. The MTurtle library implements the turtle graphics concept introduced by Seymour Papert. The output generated by programs using MTurtle library are display in animated form as the turtle moves. The MGraphics library can produce basic graphical shapes output such as line, rectangle, oval, polygon and text. Combination of these basic forms can produce complex drawing.

B. Application of Cognitive Load Theory in VJava Module

The application of cognitive load theory in the VJava Module are (1) graphical and animated output; (2) learn programming by examples; (3) program tracing; (4) step by step guide. These feature are aim to reduce the students' cognitive load in learning programing.

1) Graphical and animated output: Using the MJava library, programs written by students can produce graphical and animated output. Students can better understand the basic concepts of Java such as sequence, repetition and selection structure by associating the graphical and animation output with the written program. The output helps students to understand the program flow especially when they need to correct errors.

Fig. 2 shows an example of program code and the output generated written in the traditional approach as compared to the graphical and animated output approach using MTurtle library. This method will increase students' interest and facilitate the understanding of basic concepts. Hence, this method will reduce students' internal cognitive load in understanding basic programming concepts.

2) Learn programming by example: All concepts in the VJava Module are presented using appropriate examples which include programs to solve specific problems and the output produced. For example, a program to draw a rectangle is shown in Fig. 3 while the output is shown in Fig. 4.

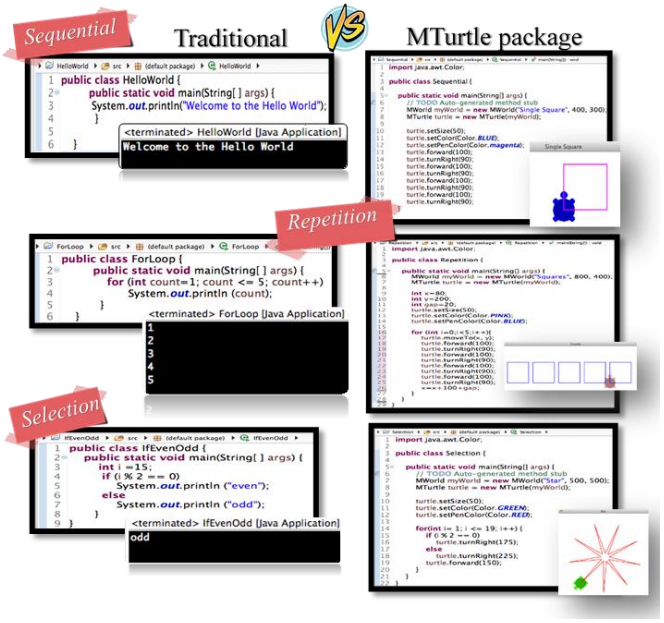


Fig. 2. Traditional Approach vs Graphical Output Approach.

```
public class MyTurtleApp {
    public static void main(String[] args) {
        MWorld myWorld = new MWorld("My First Turtle World");
        MTurtle turtle = new MTurtle(myWorld);          // (a)
        // draw a square
        turtle.forward(100);                            // (b)
        turtle.turnRight(90);                           // (c)
        turtle.forward(100);                            // (d)
        turtle.turnRight(90);                           // (e)
        turtle.forward(100);                            // (g)
        turtle.turnRight(90);                           // (h)
        turtle.forward(100);                            // (i)
    }
```

Fig. 3. Program Example.

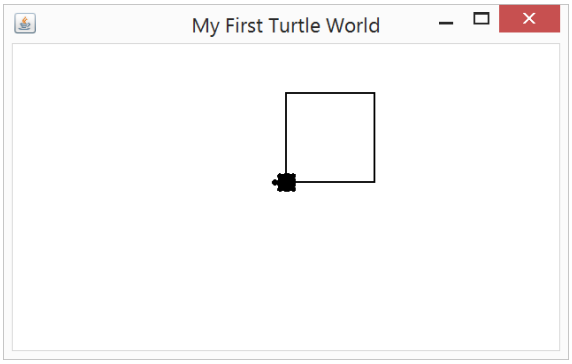


Fig. 4. The Output.

3) Program tracing: Programs in the VJava Module are explained by tracing the program line by line. Program tracing is a method to simulate how the program is executed on paper, by step through it line by line. With this explanations students can understand the program discussed before they can write their code. Fig. 5 shows an example of tracing a program to draw a square.

4) Step by step guide: The step by step guide is to explain important tasks. For example, to describe the process to install JDK and Eclipse on a computer. The step by step guide enables students to follow it to perform the tasks on their own.

C. Application of Constructivism Theory in VJava Module

Features of the theory of Constructivism applied in the module are (1) scaffolding; (2) forming new ideas; (3) exploration of new ideas; (4) construction of new knowledge.

1) Scaffolding: The MJava Module implements scaffolding learning that consisting of 5 stages namely requirements, concepts and syntax, reinforcement, program segment exercise and programming exercise (Fig. 6).

The five stages are explained with an example as follows:

a) Requirements: At this stage the requirements for a topic are described. For example, in Topic 4 Repetition Structure, it is explained how a repetition structure can simplify the writing of the program without duplicating the program statements. This is demonstrated by showing an example of a program to draw two squares side by side by copying the program segment to draw a square and repeating it twice with some changes to set the location and direction of the turtle of the second square. This is discussed in detail with program examples until the purpose is achieved.

The next query put forward is "how to output 5 or maybe 10 squares in a row. Do we need to repeat the program segment 10 times?" (Fig. 7) This situation justifies the need for a repetition structure.

a) Concepts and syntax - In this stage the concepts of a topic are discussed with simple examples and the syntax of the statement is also explained. For example, the program to draw 10 squares in a row as discussed in (a) is shown in Fig. 8. In this example, a repetition statement, that is for loop is introduced and the syntax is explained in Fig. 9.

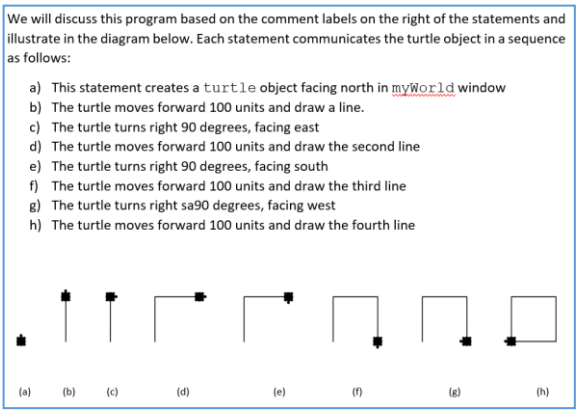


Fig. 5. Tracing a Program that Draws a Square.

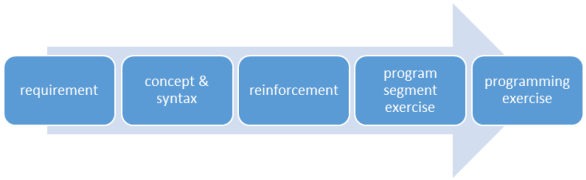


Fig. 6. Scaffolding.

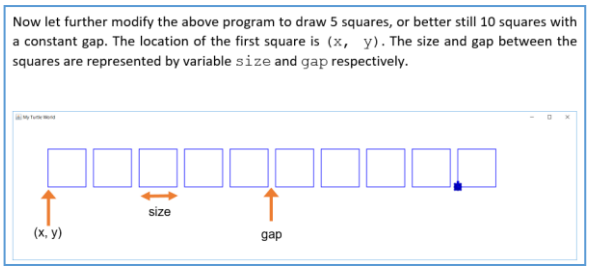


Fig. 7. Question Put Forward to Draw 10 Squares Side by Side.

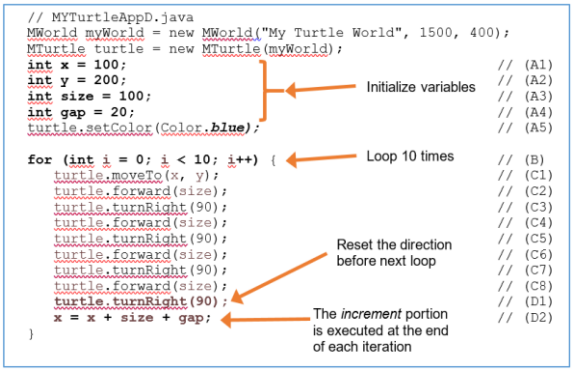


Fig. 8. Program to Draw 10 Squares in a Row using for Statement.

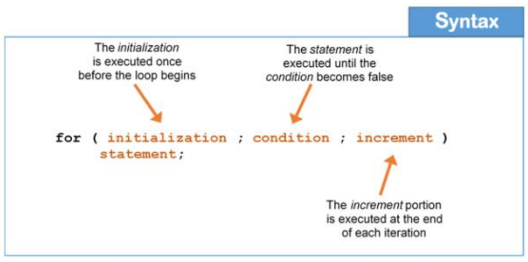


Fig. 9. Syntax for "for Loop".

b) Reinforcement- This stage further explains the program example by tracing the program. Using the same example to draw 10 squares, the program is traced for each loop to show the drawing displayed and the value and calculation of variables representing the coordinates and direction of the turtle.

c) Program segment exercise - Simple questions such as to trace some program segments were given to familiarize students with the concepts being discussed. Further exercises related to the discussed problems are also provided, for example to draw 5 horizontal lines. Using the program to draw 10 squares, students can write a solution to this problem.

d) Programming exercise - The question given in this stage requires students to apply the concepts discussed to solve the problem on their own. An example of the question is to draw 5 cascading squares.

These five scaffolding stages of learning are emphasized in the learning outcomes of the related topics. Examples of learning outcomes for Topic 4 Repetition Structure are shown in Fig. 10. Stages 1 to 5 refer to learning outcomes 1 to 5, respectively.

2) Forming new ideas: Forming new ideas is a key feature of the theory of constructivism. Based on the examples given, students are anticipated to form new ideas in solving problems. For example, based on the given example to draw a square, students should be able to form new ideas to draw a triangle (Fig. 11). The idea needed is to determine the degree of angle for the turtle to turn to draw the lines.

This chapter demonstrates the concept of repetition control structure using MTurtle class library. The learning outcome of this chapter is:

- understand the need of loop control structure
- understand the concept of loop control structure
- understand the syntax of `for` loop using `MTurtle` objects
- understand the flow of control of `for` loop using `MTurtle` objects
- able to apply `for` loop to draw multiple `MTurtle` objects

Fig. 10. Learning Outcome for Topic 4 Repetition Structure.

Task 3: Draw equilateral triangle

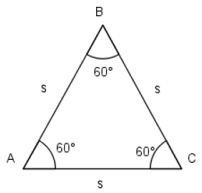
1. Create a new project and name it **AxxxxxxTask3** (Axxxxxx is your student number). In that project, add a new class called **AxxxxxxTask3**.

   Start with a new `main` method. Create a `turtle` object in `MWorld`.

   a. Set pen width to 3.

   b. Draw an equilateral triangle with sides of length = 150.

      An equilateral triangle has three sides of equal length, connected by three angles of equal width.



Question: What is the statement to set pen width to 3?

Fig. 11. Example of Question to form New Ideas.

3) Exploration of new ideas: In VJava Module, students were given examples based on the concepts discussed. In the programming exercises, students are asked to explore the MTurtle and MGraphics software library by referring to the library documentation. The documentation listed all the methods that can be used to create a drawing using the respective library. For example, in an exercise, students are asked to change the color and line thickness of the turtle. Students are guided to explore by referring to the MTurtle documentation.

4) Construction of new knowledge: One of the exercises in Topic 4 Repetition Structure is to trace a program segment that produces a pattern called polyspiral. Students were given a note that describes polyspiral as a hint. This is then followed by a programming exercise where students were asked to write a program that produces the polyspiral pattern. New knowledge constructed in this exercise is that pseudo-code for the polyspiral that can be converted into program code. Students can refer to the program tracing exercise and modify the program code to solve this problem.

Other ensuing exercise questions require students to change the values of some variables to produce interesting polyspiral shapes. A new knowledge constructed that polyspiral patterns can be produced with different variable values that applies the geometric coordinate concept in mathematics.

D. Validation of Module by Experts

The expert validation process is conducted on the developed VJava Module, to determine the validity, and usability of the module. The evaluation was performed by three programming experts and one instruction design expert as shown in Table I. PK1, PK2 and PK3 are the programming experts, while the PRB is the instruction design expert.

The questionnaire for programming experts consists of three sections, Part A for the demographics of respondents, Part B for module content and Part C for module usability. The survey for instruction design experts also comprises three sections, Part A for the demographics of the respondent, Part B for module design and Part C for module usability.

1) Module Content Validation: Content validity is based on the mean score of learning outcomes of the topics in the VJava Module. Findings for expert evaluation of learning outcomes are summarized based on topics as shown in Table II.

The measure used is a 5-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = moderate agree, 4 = agree, 5 = strongly agree). The descriptive range for the mean score is divided into three levels: low (1.00 - 2.33), medium (2.34 - 3.66) and high (3.67 - 5.00). Based on the expert evaluation, all chapters showed high scores, in the range of 3.67 to 5.00. The highest score was for Topic 4 Repetition Structure with a mean score of 4.93 while the topic with the lowest score was Topic 8 Arrays with a score of 3.89. Topic 4 is the topic that introduces the concept of repetition that is discussed in detail. Topic 8 is on arrays which is considered as a difficult topic for novices.

TABLE. I. EXPERT PROFILE

No.	Designation	Expertise	Experience
PK1	Senior lecturer	Programming education, software testing	20 years
PK2	Professor	Artificial intelligence	24 years
PK3	Associate professor	Programming language design, software testing	17 years
PRB	Lecturer	Education technology, computational thinking	11 years

TABLE. II. EXPERT EVALUATION ON LEARNING OUTCOME

Topic	Title	Mean score
1	Introduction to Programming and Java	4.40
2	Objects : Using Turtle Graphics	4.47
3	Input and Output	4.58
4	Repetition Structure	**4.93**
5	Selection Structure	4.67
6	More Repetition Structure	4.87
7	Repetition Structure : nested loop	4.67
8	Arrays	**3.89**

Expert reviews of the overall content of the module are shown in Table III. The reviews are positive with some suggestions for improvement.

2) Modul design validation: The design expert evaluates and validates that module design includes features of cognitive load theory, constructivism theory. The overall review by the design expert are as follows:

"The content structure of this module is clear and provides some examples of incremental program development: 1) examples to illustrate concepts; 2) guided examples for reinforcement (work-example); and 3) activities/exercises that students need to complete without guidance. Need to include more exercises - a broader context that resembles the program used in the industry/field (suggestion)."

3) Usability of the module: All experts have evaluated and validated the usability of the module in terms of learnability, efficiency, memorability and satisfaction. As for the error aspects, experts recommend the need for minor changes in terms of spelling and grammar. An overall review of the usability of the module is shown in Table IV. All of the experts confirm that the VJava module is suitable to be used in basic programming course.

E. Limitation of the Study

This study focuses on the early part of a basic programming course in which most novice learners have difficulties in understanding the basic concept of programming and the syntax of a programming language. Using a simple approach to write programs that produce graphical and animated output will increase students' interest and reduce anxiety, hence will reduce the cognitive load in learning programming. The next part of the programming course which focuses on problem solving and programming skills is part of a different study.

TABLE. III. EXPERT REVIEW ON MODULE CONTENT

Expert	Remarks
PK1	1. This module is very interesting. It's generally very interesting and easy to understand. 2. Only the topic on array seem difficult.
PK2	1. An interesting learning method using visuals and graphics. 2. Simple examples are used to introduce a concept.
PK3	1. This is a great effort to improve the teaching quality of programming. The proposed concept can be further refined to increase students' understanding of Java. 2. I strongly agree with the concept of 'tracing' used and recommend to emphasis on this. 3. Need to include additional references especially from youtube to enable students to learn more about each topic. 4. Overall, I agree with this module and hope to improve it in the future.

TABLE. IV. EXPERT REVIEW ON USABILITY OF THE MODULE

Expert	Remarks
PK1	1. Practical. The use of visuals makes it easier for students to remember/understand what to do (visual/visual algorithms)
PK2	1. Support materials for the laboratory are very suitable 2. It should be used with textbooks and reference books for an in-depth explanation. 3. Overall, it is simple, interesting and appropriate approach for 21st-century learning
PK3	1. This module is simple to understand.
PRB	1. Generally, this module is easy to use. The content is presented in a structured way. 2. The use of icons to mark 'sections' in module content is good.

V. CONCLUSIONS

This article describes the development of the VJava Module that uses visual output approaches for a basic programming course. This module uses the developed MJava library that can be integrated into student programs to produce graphical and animated outputs. Cognitive load theory and constructivism are applied in the design of this module. The VJava module has been validated by programming and instruction design experts in terms of content, design and usability of the module. All experts have responded positively and agreed that this module is suitable for use in the teaching and learning of programming in higher education institutions.

In the next phase of this project, the VJava module will go through a pilot test process before being implemented in a real learning environment and evaluated for its effectiveness. The results of this evaluation will be reported in the subsequent paper.

ACKNOWLEDGMENT

Appreciation to the Ministry of Education of Malaysia for supporting this work through its research grant FRGS/1/2016/ICT01/UKM/02/3.

REFERENCES

[1] M. M. Bashir and A. S. M. L. Hoque, "An effective learning and teaching mod el for programming languages," J. Comput. Educ., vol. 3, no. 4, pp. 413–437, 2016.

[2] A. Luxton-Reilly et al., "Introductory programming: A systematic literature review," in Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE, 2018, pp. 55–106.

[3] M. Noone and A. Mooney, "Visual and Textual Programming Languages: A Systematic Review of the Literature," Oct. 2017.

[4] P. C. Tavares, P. R. Henriques, and E. F. Gomes, "A computer platform to increase motivation in programming students-PEP," in CSEDU 2017 - Proceedings of the 9th International Conference on Computer Supported Education, 2017, vol. 1, pp. 284–291.

[5] R. Latih, M. Abu Bakar, N. Jailani, N. M. Ali, S. M. Salleh, and A. M. Zin, "A Design for Challenge-based Learning of Programming," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 5, pp. 1912–1918, 2018.

[6] R. Smetsers-Weeda and S. Smetsers, "Problem solving and algorithmic development with flowcharts," in ACM International Conference Proceeding Series, 2017.

[7] A. Vihavainen, J. Airaksinen, and C. Watson, "A Systematic Review of Approaches for Teaching Introductory Programming and Their Influence on Success," in Proceedings of the Tenth Annual Conference on International Computing Education Research, 2014, pp. 19–26.

[8] R. M. Kaplan, "Using Problem-Based Learning in a CS1 Course -Tales from the Trenches," in Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS), 2015, pp. 86–90.

[9] R. Horváth and S. Javorský, "New Teaching Model for Java Programming Subjects," Procedia - Soc. Behav. Sci., vol. 116, pp. 5188–5193, 2014.

[10] R. P. Medeiros, G. L. Ramalho, and T. P. Falcao, "A Systematic Literature Review on Teaching and Learning Introductory Programming in Higher Education," IEEE Trans. Educ., 2019.

[11] M. Yousoof and M. Sapiyan, "Optimizing instruction for learning computer programming- A novel approach," Commun. Comput. Inf. Sci., vol. 516, pp. 128–139, 2015.

[12] S. Mohd Salleh, Z. Shukur, and H. Mohamad Judi, "Scaffolding Model for Efficient Programming Learning Based on Cognitive Load Theory," Int. J. Pure Appl. Math., 2018.

[13] X. Li, "Application of Cognitive Load Theory in Programming Teaching," J. High. Educ. Theory Pract., vol. 16, no. 6, pp. 57–65, 2016.

[14] M. Guzdial, "Constructivism vs. Constructivism vs. Constructionism," Computing Education Research Blog, 2018. [Online]. Available: https://computinged.wordpress.com/2018/03/19/constructivism-vs-constructivism-vs-constructionism.

[15] M. Rob and F. Rob, "Dilemma between constructivism and constructionism: Leading to the development of a teaching-learning framework for student engagement and learning," Journal of International Education in Business, vol. 11, no. 2. Emerald Group Publishing Ltd., pp. 273–290, 05-Nov-2018.

[16] A.-M. Gasaymeh, I. A. AlJa'afreh, A. Al-Dmour, and M. A. Alrub, "Higher Education Students' Preferences for Applying the Principles of Constructivism in Learning Programming Languages with the Use of ICTs," J. Stud. Educ., vol. 6, no. 3, pp. 168–187, 2016.

[17] M. A. Bakar et al., "Kerangka Bagi Persekitaran Pembelajaran Berpusatkan Pelajar untuk Latihan Pengaturcaraan Kendiri," ASEAN J. Teach. Learn. High. Educ., vol. 10, no. 1, pp. 24–37, 2018.

[18] J. M. Sáez-López, M. Román-González, and E. Vázquez-Cano, "Visual programming languages integrated across the curriculum in elementary school: A two year case study using 'scratch' in five schools," Comput. Educ., vol. 97, pp. 129–141, Jun. 2016.

[19] H. Husain, N. Kamal, M. F. Ibrahim, A. B. Huddin, and A. A. Alim, "Engendering problem solving skills and mathematical knowledge via programming," J. Eng. Sci. Technol., 2017.

[20] M. Armoni, O. Meerbaum-Salant, and M. Ben-Ari, "From Scratch to 'Real' Programming," ACM Trans. Comput. Educ., vol. 14, no. 4, pp. 1–15, 2015.

[21] S. Cooper, W. Dann, and R. Pausch, "Teaching objects-first in introductory computer science," ACM SIGCSE Bull., vol. 35, no. 1, p. 191, 2003.

[22] M. E. Caspersen and H. B. Christensen, "Here, there and everywhere - on the recurring use of turtle graphics in CS1," in Proceedings of the Australasian conference on Computing education - ACSE '00, 2000, pp. 34–40.

[23] E. Roberts and A. Picard, "Designing a Java graphics library for CS 1," ACM SIGCSE Bull., vol. 30, no. 3, pp. 213–218, 1998.

[24] B. Meyer, "Towards an object-oriented curriculum," J. Object-Oriented Program., vol. 6, no. 2, pp. 76–81, 1993.

[25] B. W. Becker, "Teaching CS1 with Karel the Robot in Java," Proc. thirty-second SIGCSE Tech. Symp. Comput. Sci. Educ. (SIGCSE '01)., pp. 50–54, 2001.

[26] J. Bergin, M. Stehlik, J. Roberts, and R. Pattis, Karel J Robot : a gentle introduction to the art of object-oriented programming in Java. Dream Songs Press, 2013.

[27] M. Kölling, "The problem of teaching object-oriented programming. Part 1," J. Object Oriented Program., vol. 11, no. 8, pp. 8–15, 1999.

[28] N. Thota and R. Whitfield, "Holistic approach to learning and teaching introductory object-oriented programming," Comput. Sci. Educ., vol. 20, no. 2, pp. 103–127, 2010.

[29] M. Berry and M. Kölling, "Novis: A notional machine implementation for teaching introductory programming," in Proceedings - 2016 International Conference on Learning and Teaching in Computing and Engineering, LaTiCE 2016, 2016.

[30] E. J. Johan, S. Idris, M. A. Bakar, and M. Mukhtar, "Persuasive Object Oriented Programming Lab Assignment Framework," Int. J. Technol. Incl. Educ., vol. 4, no. 1, pp. 557–565, 2015.

[31] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," Learn. Instr., vol. 4, no. 4, pp. 295–312, 1994.

[32] J. Sweller, J. J. G. van Merriënboer, and F. Paas, "Cognitive Architecture and Instructional Design: 20 Years Later," Educational Psychology Review. 2019.

[33] T. de Jong and T. Jong, "Cognitive load theory, educational research, and instructional design: some food for thought," Instr. Sci., vol. 38, no. 2, pp. 105–134, 2010.

[34] N. H. Jalani and L. C. Sern, "The Example-Problem-Based Learning Model: Applying Cognitive Load Theory," Procedia - Soc. Behav. Sci., 2015.

[35] T. Pittayapiboolpong and P. Yasri, "Development of an Integrative Learning Unit to Enhance Students' Conceptual Understanding of Dissolution and Their Reasoning Sophistication," J. Res. Sci. Math. Technol. Educ., vol. 1, no. 3, pp. 283–310, 2018.

[36] S. Papert, Mindstorms: Childern computer and powerful ideas. 1980.

[37] G. Walton, M. Childs, and G. Jugo, "The creation of digital artefacts as a mechanism to engage students in studying literature," Br. J. Educ. Technol., vol. 50, no. 3, pp. 1060–1086, May 2019.

[38] M. Lodi, D. Malchiodi, M. Monga, A. Morpurgo, and B. Spieler, "Constructionist Attempts at Supporting the Learning of Computer Programming: A Survey," Olympiads in Informatics, 2019.

[39] R. M. Branch, Instructional design: The ADDIE approach. 2010.

Smart Age Detection for Social Media

Using Deep Learning Techniques via Ear Shape

Manal Alghieth[1], Jawaher Alhuthail[2], Kholod Aldhubiay[3], Rotan Alshowaye[4]
Information Technology, Qassim University, Qassim, Saudi Arabia

***Abstract*—Over the recent years, there has been an immense attraction towards age detection due to its raised implementation in various sectors. Such as government regulations and rules, security control, and human computer interaction. Popular human features such as the face and fingerprints can be modified or changed with time. However, ear has a stable structure that does not change with time and have unique features that satisfies the requirements of a biometric trait. This research presents a detailed analysis extracting the features of the human ear only by applying Deep Learning techniques. In particular, Convolutional Neural Network (CNN) is applied on large datasets which have multiple layers to extract the features and classify them. The proposed methodology increased the number of the dataset by collecting more private children datasets, and consequently achieved high accuracy by 98.75% along with amending the architecture of the selected neural network compared to previous studies. This research can be benefited to control the contents of social media by detecting the age of group whether it is under 18 or above 18.**

***Keywords*—Age detection; deep learning; ear features; CNN; control social media**

I. Introduction

The rapid development of technology and science has benefited humanity in different fields. Simultaneously, there has been increasing interest in age estimation.

Ear is rich in unique features and it is one of the most usable forms of authentication. Among various biometric traits, ear is a stable structure that does not change with age and satisfies the requirements of a biometric trait. This gives it an advantage compared with other popular human features such as the face and fingerprints, which can change with time or can be modified. Ear images can be analyzed using deep learning techniques to detect age of a person.

The main idea of deep learning is that, a different type of data can be implemented into a machine to learn certain features which can lead to an excellent decision making for the classification. Deep Learning is considered as a new field of machine learning, it has achieved great results and high accuracy in various areas compared to human experts such as image classification which takes an input and predict a particular class, Object detection, Image retrieval, Semantic segmentation and Human pose estimation. Deep learning can be divided into four categories as shown in Fig. 1. The most appealing characteristic to use deep learning, the fact that it can be applied on large datasets, the increasing ability of processing, the more layers a network has the higher the level of features it will learn. One of the most powerful techniques in deep learning is neural networks that contains numerous amounts of neurons which offers answers to different kinds of problems. Neural network is a combination of algorithms that is applied in identifying various patterns, like numerical, images, audio, text or time series. The NN types are: Convolutional Neural Networks, Recursive Neural Networks, Unsupervised Pre-trained Networks and Recurrent Neural Networks [1].

The Convolutional Neural Networks (CNN) is an outstanding deep learning algorithm. It consists of three layers: the convolutional layer, the pooling layer and the fully-connected layers. It can be applied to large dataset like images and videos to extract features and classify it. CNN require much less preprocessing compared to other algorithms. Recent articles have proved that CNN is highly effective and already exceeded the accuracy of human experts [2].

Many techniques have been developed to authenticate and verify the identity of humans and age detection.

For instance, Decision tree J48 classifier applied on FG_NET. It consists of three phases image preprocessing, data mining, estimation and evaluation [3].

Deep CNN model called VGG-Face. It contains 11 layers, eight convolutional layers and 3 fully connected layers that were applied on a database for face recognition task. It was modified and fine-tuned to perform age estimation on the Adience database [4].

Convolutional Neural Networks architectures were applied to classify ear appearance-based features. These architectures include AlexNet, VGG, GoogLeNet and SqueezeNet [5].

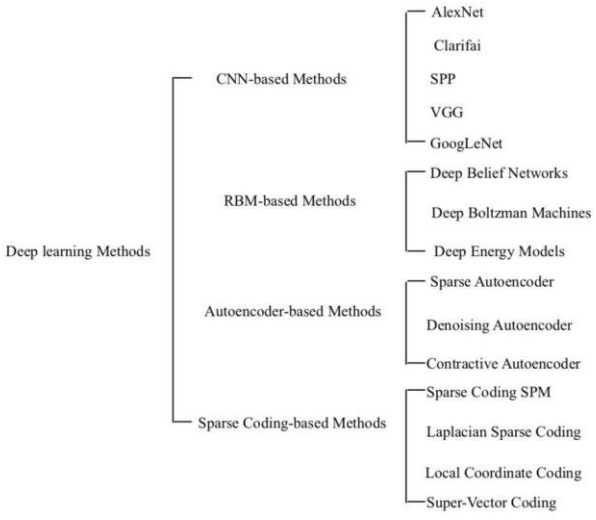


Fig. 1. Deep Learning Categories.

Other techniques were applied on ear geometric features. Such as, Logistic regression provides a linear discrimination between different classes. Its objective is to minimize the logistic function. Random Forests is an ensemble machine learning algorithm and was used for classification as well as feature selection. It works with sub-trees that are learned on a part of the training data that uses different samples from training set instead of using all of them and the average result is calculated from different sub-trees. Support vector machine is a supervised machine learning algorithm which can be used for both classification and regression. It is based on the theory of statistical learning [5].

II. Literature Review

A few attempts have been done on detecting the age via the ear shape. For instance, a detailed analysis on identifying soft biometric traits, age and gender, from ear images was presented. In [5], it uses geometric features and appearance-based features to illustrate the ear morphology. These features have been classified by employing four different methods: logistic regression, random forests, support vector machines, and neural networks. The gender classification resulted in accuracy of 94% and for the age classification achieved a 52% accuracy. Furthermore, the work in [6] focused on using profile face images and ear images as input to build a multimodal deep neural network for age and gender classification in order to enhance the accuracy. It trained their model on the UND-F, UND-J2, and FERET datasets for gender classification and on the FERET dataset for age classification. The experiment indicated that profile face images have numerous amounts of features that can support age and gender classification and merging profile face and ear image can reach great accuracies. Compared to the reported age classification accuracy in the previous work, the study has increased the accuracy by 9% for five class age group classification on FERET dataset and reached 60.97% of accuracy.

To support the proposed methodology, studies about ear features have proved ear is a unique body part that can provide numerous information. M. Taura, L. Adamu, A. Gudaji, and M. Modibbo [7], discussed the possibility to determine age from ear shape by testing the correlation between age and ear length, width and index for both left and right sides. The study applied the regression equation for age detection. It proved the application of external ear morphometry that can serve as additional tool for age estimation of individuals. The human ear is stable and does not change over time and that have been presented. In [8],[9] and [10], it has been proven through studies that the shape of ear of the same person show some similarities but are not strictly symmetric. Ear verification can also be applied to differentiate between identical twins, due to its unique features; there are no two ears that have the same helix.

On the other hand, face images can be used to detect age where in research [4], Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, proposed a model to detect age using deep CNN model called VGG-Face that was applied on a database for face recognition task. It was modified and fine-tuned to perform age estimation on the Audience database. A trained CNN model for face recognition was used in this research for age detection and resulted in great outcomes. This work proved that employing a pre-trained CNN on large data set can overcome over fitting issues. The proposed model achieved a higher accuracy by 9%. The work in [11] applied three methods on facial images to detect age: Feature Extraction , Hierarchical age estimation and Age Estimate Fusion. The research proved that using face image can be employed effectively in various areas to detect age. In [3], it discussed the potential of determining age from facial images using machine learning technique, and applied J48 classifier on FG_NET dataset. The experiment proved that the proposed system can give results with high precision and low time complexity, using decision tree J48 classifier resulted in high precision rate with low time taken on the FG_NET dataset with accuracy of 89.13 % and time taken is 0.023.

Recently, there has been a growing interest in age estimation, due to its raised implementation in various sectors such as government regulations and rules, security control and human computer interaction. In spite of the advances in age estimation, it is still a challenge. There are two main problems in detecting the age by ear shape, lack of ear dataset collection and age classification accuracy. One of the previous studies [5] have reached 52% accuracy and they claim that it is because of the short number of datasets. Their theory states that the more datasets used the more accuracy they will get. Hence, this research will attempt to improve accuracy by using deep learning techniques. It requires a large dataset to give more accuracy and can also extract ear features by itself. There are several ways applied on ear features to detect the age, such as logistic regression, random forest, support vector machines and neural network.

The popular human features such as face and fingerprints can be modified or changed with time. However, ear has unique features and it is a stable structure that does not change with time and satisfies the requirements of a biometric trait. Nowadays, users of social media can give false information without validation, hence users can explore any content available, whether it is age appropriate or not. This research contributes in controlling the content of the watched media depending on the user's age via the ear shape only.

III. Proposed Methodology

This research proposes a system that detect the age by applying ear dataset into deep learning algorithms which will analyze the images and extract the features via multiple hidden layers of deep learning and classify them. The model was trained on AMI, AWE, and a Private Children dataset. The resulted classification will give the system the output needed which is the age group whether above 18 or under 18. The proposed model is shown in Fig. 2.

The implementation begins with collecting ear images from different sources (This model used 3000 ear images which was divided into 80% for training and 20% for testing) and using csv files to handle the data. Importing the necessary python libraries for deep learning, load and process the data, then defining the architecture of CNN model and how to apply it. Fig. 3 shows the algorithm used in this research.

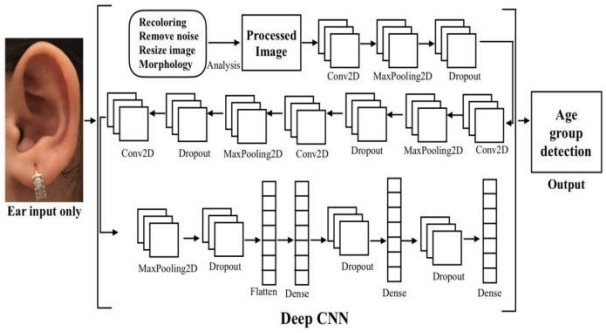


Fig. 2. Proposed Methodology Model.

```
Import libraries
Load training images
Read training csv file
Start
for counter to (images range)
train images + labels preprocess images
Stop
Drop (image name, gener) columns from csv file for validation
Randmly validate 10% of training images
Build CNN architecture
Compile CNN model
Load testing images
Start
for counter to (images range)
test images
preprocess images
Stop
Predict age calss
Output accuracy
```

Fig. 3. Smart Age Detection Algorithm.

The images were different in sizes and shape for that a preprocessing step was applied before the training and testing.

Training the model using the training images set and their corresponding true labels, then it will pass the validation images to assure the performance of the model. Also, in this phase the number of epochs is defined.

The architecture of the applied model includes deciding the number of hidden layers, number of neurons in each layer and what activation function embedded in each layer to convert an input signal to an output signal, to be use the output signal as an input in the next layer. For the 2D convolution layer it uses many filters in a single layer that learns ear features, the extracted features used as an input to the next layer. Following conv2D, a max pooling layer was performed to reduce the amount of pixels computation that needs to be done and sends the most important data to the next layer. The utilized hidden layers are Dropout (drops random unit of CNN model and to prevent overfitting) and Flattened (flattens all the features maps into a single column). Lastly, the fully connected layer (Dense layer) is the output layer that will classify the ear images either above 18 or under 18 years. This research contributes in applying two activation functions in different layers, ReLU (Rectified Linear Unit) and Sigmoid were applied. ReLU activation function was embedded in each layer; except the output layer a Sigmoid activation function was performed to predict the probability as an output, which is excellent for the binary classification task.

For testing, loading 20% of the testing images set with testing csv file without labels. Before making predictions, the model will preprocess these images using the same steps mentioned in the training phase. Then it will predict the age for these ears of human using the trained model. Finally, the model will classify the tested images either more than 18 or less than 18 years old.

IV. Results

The major findings were classifying images to detect the age either above 18 or under 18 for social media and the exceptional accuracy of 98.75%. Table I shows the result related to previous studies and proven the higher accuracy that has been achieved in this research.

TABLE. I. Comparison between the Proposed Method and Previous Studies

Title	***Dataset***	***Overall Accuracy***	***Method***	***Year***	***Age Group***
The Prediction of Old and Young Subjects from Iris Texture[12]	100 Subjects	64%	Rando-mForest Algorit-hm	2013	Young: (22-25), Old: (35+)
Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model[4]	Adience database	59.9%	Deep CNN	2017	0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+
Intelligent Age Estimation From Facial Images Using Machine Learning Techniques[3]	FG_NET dataset	89.13%	Decisio-n tree J48 classifie-r Algorit-hm	2018	Class1: (3-7,26-30), Class2: (8-25), Class3: (31-50)
Age and Gender Classification from Ear Images[5]	UND J ear Database	52%	Deep CNN	2018	18-28, 29-38, 39-48, 49-58, 59-68+
Multimodal Age and Gender Classification Using Ear and Profile Face Images[6]	FERET	60.97%	Deep CNN	2019	18-28, 29-38, 39-48, 49-58, 59-68+
Smart Age Detection For Social Media Using Deep Learning Techniques via Ear Shape (The proposed method)	AMI, AWE & Private Children datasets	98.75%	Deep CNN	2019	Above 18 & Under 18

V. Discussion

CNN is one of deep learning architectures and it was applied in this research to classify the ear images. The lack of classified dataset limited the research from identifying age in different kinds of groups such as child, teenagers, adult or elderly. However, above 18 or under 18 age group can be extremely useful in applying this to the social media and the most important thing is to identify adults or young generation, that can allow controlling the appropriate watched content. In the future utilizing the medical aspect can help in improving the extraction of more features and producing better accuracy result for different group of ages. This research used the shape of ear only that contains unique features and the chosen architecture succeeded and resulted in a great accuracy. A collaboration with biologist could be made to perform a better accuracy based on the features of the ear itself.

VI. Conclusion

This research proposed a model to identify human age either above 18 or under 18 based on ear shape only by using deep CNN model. It applied most appropriate activation function to make the model more accurate. This study contributed in applying two activation functions, ReLU which was embedded in each layer except the output layer and Sigmoid activation function was performed to predict the probability as an output.

The proposed model reached an exceptional accuracy of 98.75%. Due to the high accuracy of detecting the age groups, this research can be applied in social media platforms to authenticate user's information.

References

[1] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, vol. 187, pp. 27–48, 2016.

[2] J. Gu et al., "Recent advances in convolutional neural networks," Pattern Recognit., vol. 77, pp. 354–377, 2018.

[3] A. R. Abbas, "Intelligent Age Estimation From Facial Images Using Machine Learning Techniques," Iraqi J. Sci., vol. 59, no. 2A, 2018.

[4] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model," no. 1, 2017.

[5] D. Yaman, F. I. Eyiokur, N. Sezgin, and H. K. Ekenel, "Age and gender classification from ear images," IWBF 2018 - Proc. 2018 6th Int. Work. Biometrics Forensics, pp. 1–7, 2018.

[6] D. Yaman and F. I. Eyiokur, "Multimodal Age and Gender Classification Using Ear and Profile Face Images."

[7] M. Taura, L. Adamu, A. Gudaji, and M. Modibbo, "Application of external ear morphometry in age prediction: a pilot study," Int. J. Res. Med. Sci., vol. 3, no. 7, pp. 1775–1779, 2015.

[8] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," IET Biometrics, 2018.

[9] A. Sabet, K. Kamal, A. Ghany, and H. Elmahdy, "Human Ear Recognition Using Geometrical Features Extraction," Procedia - Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 529–537, 2015.

[10] I. Omara, F. Li, H. Zhang, and W. Zuo, "A novel geometric feature extraction method for ear recognition," Expert Syst. Appl., vol. 65, pp. 127–135, 2016.

[11] S. E. Padme and P. S. Desai, "Estimation of Age from Face Images," Int. J. Sci. Res., vol. 4, no. 12, pp. 1927–1931, 2016.

[12] A. Sgroi, K. W. Bowyer, and P. J. Flynn, "The prediction of old and young subjects from iris texture," Proc. - 2013 Int. Conf. Biometrics, ICB 2013, pp. 1–5, 2013.

Automated Methodology for Optimizing Menus in Personalized Nutrition

Valery I. Karpov[1], Nikolay M. Portnov[2]
Department "Information Systems and Technologies"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Igor A. Nikitin[3]
Department "Technology of Grain Processing, Bakery Pasta and confectionery industries"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Yury I. Sidorenko[4], Igor V. Zavalishin[9]
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Sergey M. Petrov[5]
Department "Automated Control Systems"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Nadezhda M. Podgornova[6]
Department "Technologies of Production and Organization of Catering and Merchandising"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Mikhail Yu. Sidorenko[7], Sergey V. Shterman[8]
GEON Limited Liability Company
Moscow, Russian Federation

***Abstract*—In the personalized nutrition rationalized management system the central practical task is to compile an optimized menu that provides the best value for a multi-criteria set of assessments. These are nutrient composition, cost (economic acceptability), energy value, food intolerances, individual preferences, etc. To solve the problem, a combined optimization method is used. It includes preliminary ordering of options and controlled enumeration. The result of solving the problem of developing a personalized nutrition menu is a diet that meets the needs of a particular diet, taking into account its nutritional status, individual preferences and intolerances, medical appointments. In connection with discrete values of the outputs of dishes and recipes, the task of optimizing the human diet is in practice a combinatorially integer, and for its solution the method of computer modeling and controlled enumeration of options was used. Evaluation of the effectiveness of optimization is carried out by external experts. The developed menu design system makes it possible to repeatedly solve the problem of optimizing a personalized menu when changing incoming data for reasons of changing dietary tasks, introducing new products, changing food preferences, etc. With this approach, the system is a personalized food model that is regularly used for rational planning and allows to achieve a reduction in labor costs (compared to the "manual" compilation of a menu in a computer system) by 2-3 times. An additional way to use such a model is the targeted design of functional product formulations. Moreover, the properties of the product are not evaluated in isolation, but as part of a specific diet.**

***Keywords*—Personalized nutrition; menu optimization; nutrition management; practical nutrition**

I. Introduction

Managing the nutrition of an individual or an entire team is one of the complex tasks of an interdisciplinary nature, which use knowledge of medical dietetics, technologies of culinary and food production, organic chemistry, psychology and other sciences [1,2,3,4].

In the complex of tasks to be solved in the system of rational nutrition management, the key one is the development of a menu corresponding to medical goals or the goals of meeting the human body's need for energy, nutrients, micronutrients. The implementation of individual characteristics of food perception is also very important. In personalized nutrition [5], such a key task is to compile a set of daily menus for a given period using specific recipes and food outlets for a given composition of meals, menu sections and the total mass of food.

At the moment, optimization calculations are carried out using spreadsheets (which is a form of manual calculation) [3]. In some cases specialized computer systems are used, including means of nutrient calculations [6, 7]. Here the menu compilation procedure is implemented in computer forms, calculation of nutrient composition is carried out automatically, the decision on compliance of the developed diet with the initial requirements is made by the technologist who develops the menu, and if necessary, introduces corrections into the next cycle of calculation and evaluation [8, 9].

There are also approaches based on the automatic generation of a diet as a mixture of food products or dishes, which do not provide an appropriate level of consumer quality and, as a result, also require the manual work of a specialist.

Common disadvantages of these solutions are: significant labor costs, duration of work, lack of an objective assessment of options being developed. These drawbacks not allow to realize the possibility of a full-fledged launch of these methods on a massive scale, required for personalized nutrition.

A numerical representation of parameters for evaluating nutrition and its components provides an objective ordering of options, and makes it possible to calculate the measure of proximity of estimates, including the main assessment of the compliance of the actual indicator with the norm. As the main meters in assessing the quality of nutrition, in the sense of its usefulness for the body, the energy value and content of the main (proteins, fats, carbohydrates) and minor (vitamins and minerals) nutrients are used. Many of these indicators will be called nutrients. To assess consumer preferences (for example, taste) and expert ratings, point scales are used. For other indicators, conventional numerical meters are used such as price, weight.

The nutrient composition values of the formulations are determined by the composition of the ingredients [10, 11]. Calculation of the nutrient composition of the formulation is carried out according to the formula (1):

$$H(n) = \sum_{i=1}^{K} Np(i,n) * p(i) * (1 - Pp(n)/100) \quad (1)$$

Where:

H(n) - is the mass of the *n*-th nutrient present in the dish;

K - is the number of ingredients;

i - is the ingredient number;

Pp(i, n) - *n* nutrient content in 100 grams of the *i* ingredient;

p(i) - is the mass of the *i*-th ingredient in grams;

Pp(n) - is the percentage of losses during the processing of the nutrient *n*.

The menu optimization criteria include: compliance of the nutrient composition with an individual consumption standard, standards for natural indicators, cost (economic acceptability), caloric structure by meals, food intolerances, individual preferences, variety, and food compatibility.

Additional estimates (indicators): balance of proteins/fats/carbohydrates by weight and share in calories; protein quality (amino acid rate [12] as well as utilization coefficient, balance coefficient, index of essential amino acids [3, 13, 14, 15, 16]); quality of the fatty acid composition (ratio of fatty acids [17, 18]); achievement of target designations for minor nutrients [1]; variability of values by days (permissible unevenness). In some cases, individual indicators can become part of optimization criteria, for example, for a ketogenic diet [19], an increased proportion of fat and a decrease in the proportion of carbohydrates are required.

Accepted axioms: food, diet - a set of complexes of dishes (not a set of products); specificity of norms (indication of average target values); determination of menu parameters by a set of days (not per day or a separate meal); the concept of "consistent with the norm" (a measure of proximity to the norm, an optional match); limiting the nutritional model just by consumption (excluding the digestibility by the body, the actual properties of raw materials and cooking); ignoring the combined effects of food components; restriction of the list of methods for evaluating nutrition only having a scientific justification.

The aim of the study was to develop a computer system prototype that simulates the development of a personalized menu, including its optimization according to a multi-criteria set of indicators, to test the methods of the most automated development of personalized diets. The computer system includes real data sets on food recipes, the nutritional composition of products and recipes, practical nutritional calculation algorithms, diet quality assessments according to certain criteria (protein biological value, fat quality, etc.) for the most adequate reproduction of the traditional menu development process. Modeling the development process in various conditions (taking into account the individual characteristics of eating, differences in standards for categories of eating, differences in assessment methods) is necessary to develop a methodology for developing diets, to achieve the maximum degree of automation, to check the effectiveness of optimization tools.

II. Statement and Analysis of the Problem of Menu Variants Design with given Characteristics

The result of solving the problem of developing a personalized nutrition menu is a diet that meets the needs of a particular diet, taking into account its nutritional status, individual preferences and intolerances, medical appointments. The degree of closeness of indicators of the actual diet to a predetermined need is estimated in numerical terms (objective function), taking into account a number of criteria (particular assessment parameters): nutrient standards, balance of basic nutrients, cost, consumer assessment, protein quality, quality of fats, uniformity of indicators for days and meals; the mutual relation of particular criteria is determined by weight coefficients for the components of the objective function, which can vary in accordance with a specific situation.

The developed diet is described by a set of technological documentation, including a menu for a period (week or more) with an indication of specific recipes for meals combined for day meals including dish outcome. The recipes used in the menu are documented by routings. The documentation set also includes the used product range, information on the nutritional composition of foods and dishes, nutritional calculations, reflecting the nutritional structure by days and meals, and evaluating the variability of numerical indicators, calculations of compliance of the diet with consumption standards.

The initial data is: 1) objective data on the person - anthropometric measurements, medical history and medical appointments (in terms of nutrition), the study of actual nutrition, body composition data; 2) information on individual preferences and intolerances of products and dishes; 3) official

consumption standards, scientifically based methods for assessing food quality (amino acid rate, fatty acid composition, balance of proteins/ fats/ carbohydrates, etc.); 4) information from reference scientific literature on formulations, nutrient composition.

Directly for the optimization problem, the source data is a set of values of individual nutritional appointments (a vector of values for a set of nutrients) that reflects characteristics of personalized nutrition (in contrast to the classical task of developing collective/average nutrition).

Due to the discrete values of the outputs of dishes and recipes, the nature of human nutrition in the form of complexes of dishes, taking into account their mutual influence, the task of optimizing the human diet in practice is a combinatorially integer. To solve it controlled enumeration of options and computer simulation methods are applied rather than analytical mathematical methods.

III. Comprehensive Optimization Algorithm for the Development of the Menu with the Given Characteristics of Nutrition Value

To determine the optimal value on the set of menu options, an integral estimate (convolution) - Q is used, a linear function with weight coefficients that reflect the significance of each individual component of the estimate (2):

$$Q = \sum_{i=1}^{N} Bkr(i) * Z(i) \quad (2)$$

Where:

N - is the number of criteria;

Bkr(i) - the value of the assessment in points according to the *i*-th criterion normalized on a common scale (0-10);

Z(i) is the weight of the *i*-th criterion that determines its importance, 0 <= Z (i) <= 1.

The use of a simplified linear objective function is explained by the insufficient amount of scientific knowledge to build more flexible models and is a primary approximation. To justify the weight coefficients of the integral function, an expert approach is used; a more reasonable estimate can be obtained by the method of hierarchy analysis [20].

Because of a large set of heterogeneous initial requirements for the menu optimization algorithm, the requirement of model stability (algorithm convergence) is presented to obtain the optimal solution even with a possible initial formal incompatibility of conditions when the found solution provides the closest approximation to a given set of conditions.

To ensure stability, a combined optimization method is used. Preliminary optimization includes: clustering the initial set of dishes to reduce the dimension of the task; selection of candidate dishes; ordering lists of dishes in the direction of local optimization by individual nutrients that are main components of optimization. Final optimization is performed by combinatorial sorting according to pre-ordered lists of dishes, in combination with local optimization for maximum compliance with the norms (partial change of the bookmark norms within the permissible limits, outputs across the cluster of permissible values). This method, in addition to stability, also provides practical results with limited time. With an increase in the search time, in the general case, the value of the integral estimation function is improved (practical convergence).

Prior to optimization, a check is carried out for completeness of the initial data and compatibility of conditions with the possibility of reformulating mutually exclusive requirements. An example of such “optimization of the initial data” is the rejection of the use of natural norms that are incompatible with the norms for nutrients and their transfer from the category of optimization criteria to indicators. The reasons for such inconsistencies may be inconsistency in official standards, errors in determining individual consumption standards.

For practical implementation in the form of a computer program, the rapid development tool “Platform 1C: Enterprise” (version 8.3) is used. For editing of recipes the “Technological chart” module is used; for editing the menu and related calculations the “Development of ration” module from the program “1C: Medicine. Dietary nutrition” [7] is used. As a supplier of raw data for dishes - electronic recipe collections in the format of “Recipe-3” is used. The module "Development of the diet" is also used to describe the source data for subsequent optimization: the menu template (meals and sections), the composition of the food components, the distribution of calories in meals. Nutrient consumption standards are filled in the directory “Categories of persons”, the substantiation for individual consumption standards is carried out in the computer system “Monitoring of physical development and nutritional status” [21].

To solve optimization problems, a standard 1C system (supplied with source codes) was amended:

- Features of the diet for personalized nutrition are implemented in the form of directories of personal preferences and intolerances.
- For reference sets of individual recipes the "Dishes" reference book was modified.
- To visualize a set of estimates of a particular menu option, the report “Assessment of a menu option” is used, which also contains means for decrypting (explicating) the calculated values.
- Optimization of the menu (“processing” in terms of the 1C platform) is carried out by the “Menu Optimization” subprogram, which includes the stages of preliminary optimization (ordering of the original lists of dishes), enumeration of menu variants, calculation of private estimates and the value of the integral function (final optimization). To control the convergence and the possibility of the resumption of enumeration, a convoluted identification of options and a variant assessment log are used.

- To visualize the progress of evaluating a number of options, the “Report on menu options” report has been developed.
- To control the composition of optimization parameters and composition of indicators, the parameter settings form is used, in which weight coefficients of the integral objective optimization function are also introduced (see Fig. 1 and 2).

To assess the effectiveness of the optimization machine, an external expert assessment of the quality of complexes of dishes in one meal, the compatibility and variety, the correspondence of the actual menu indicators and norms (proximity measures) is used. In addition to the composition of the complexes of dishes, i.e. the menu itself experts also use calculated indicators that are not involved in optimization such as uniformity of values by day, balance. For convenient and informative perception of such information, in addition to numerical arrays, means of infographics and decoding of the final values are used. The accumulated expert assessments are used to train optimization tools (they are used at the preliminary optimization stage) in subsequent work sessions. Experts in the fields are involved in evaluating effectiveness: medicine (assessment of the impact on health and treatment), food preparation technology (practical feasibility, risks), as well as the consumer himself.

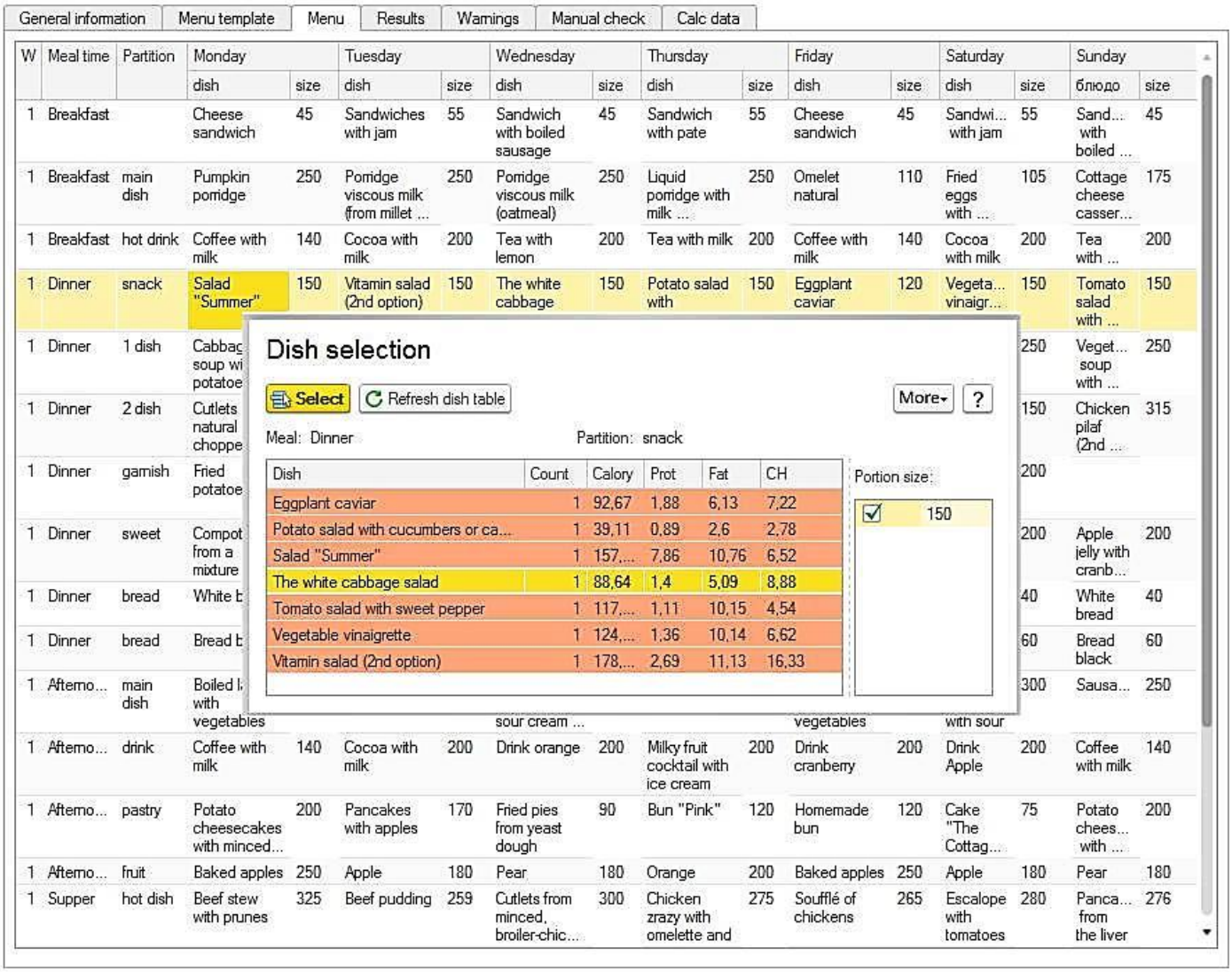


Fig. 1. Interface of the Menu Editor in the Module "Development of the Ration".

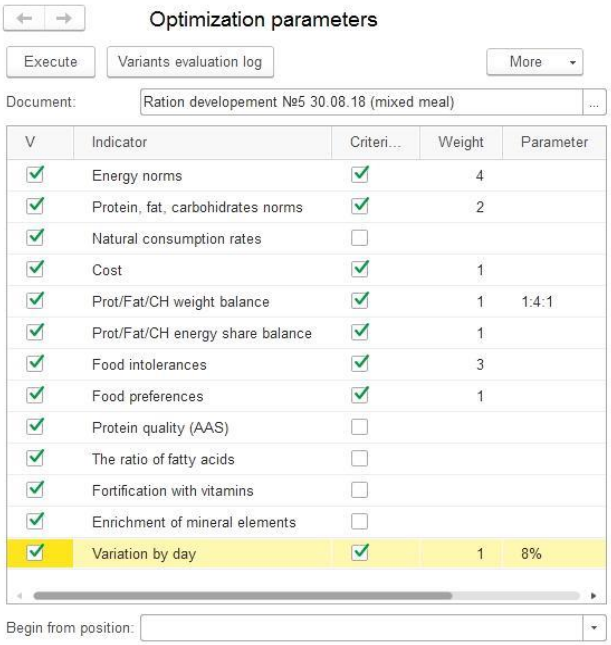


Fig. 2. Parameters of the Optimization Module.

IV. Implementation of the Program Discussion of Results

The developed system has been tested for cases of compiling a personalized menu (individual and micro-teams) for a healthy person and patients with alimentary-dependent diseases (with the indicated diets, medical prescriptions), taking into account individual preferences in dishes, for individually calculated nutritional needs (ICS - individual consumption standards were determined using the NutriMon computer system [21]) with various correlations of importance of particular criteria of the integral objective function. Based on the results of testing, changes were made to the system. The new version takes into account individual characteristics of recipes and cooking methods, typical servings, sets of products with individual binding. By means of automatic optimization in comparison with the "manual" menu making in a computer system a reduction in labor costs was achieved by 2,3 times.

The system provides the ability to repeatedly solve the problem of optimizing a personalized menu when changing incoming data for reasons of changing dietary tasks, introducing new products, changing food preferences, etc. With this approach, the menu design system is a personalized food model that is regularly used for rational planning.

An additional way to use such a model is the targeted design of functional product formulations, including the case of targeted development for personalized nutrition. Moreover, the properties of the product are not evaluated in isolation, but as part of a specific diet.

Directions for the development of the model: improvement of primary information about dishes and recipes and their official publication; taking into account the variability (statistical nature) of the evaluation indicators; comparison of actual consumption with a planned menu as part of an overall assessment of the effectiveness of diet therapy; use of nutribiomics and nutrigenomics; psychology of food consumption; family model of nutrition (optimization on a microgroup, taking into account individual characteristics); analysis of the structure of a multi-criteria set of assessment indicators to identify the main components; porting to a mobile platform and to a cloud service.

V. Conclusion

Software has been developed for the automated development of personalized diets with the possibility of automatic optimization to achieve maximum proximity of the menu indicators to the desired nutritional settings. Evaluation of the quality of the developed diet is carried out according to a set of criteria using a generalizing (integral) function, the behavior of which is adjusted by a set of component importance coefficients.

The result was achieved using: ready-made computer programs "1C: Planned Nutrition" and "NutriMon" (monitoring of nutritional status and physical development); electronic collections of recipes in the "RecipeBook" format; development of additional optimization modules and database modification by means of the 1C: Enterprise platform.

The materials for state registration of the developed computer program are prepared.

Further directions for the development of the model are: improving the primary information about dishes and recipes and their official publication; taking into account the variability (statistical nature) of the evaluation indicators; comparison of actual consumption with a planned menu (as part of an overall assessment of the effectiveness of diet therapy); use of nutribiomics and nutrigenomics; psychology of food consumption; family model of nutrition (optimization on a microgroup, taking into account individual characteristics); analysis of the structure of a multi-criteria set of assessment indicators to identify the main components; porting to a mobile platform and to a cloud service.

Acknowledgment

This article was written with support from the Government of the Russian Federation (Resolution of 16.03.2013), agreement and subsidies for the fulfillment of a fundamental part of a state order under Project 15.9195.2017/5.1.

References

[1] Methodological recommendations 2.3.1.2432-08 Norms of physiological requirements for energy and nutrients for various population groups of the Russian Federation.

[2] Baranovsky A.Yu., Dietology, St. Petersburg: Peter, 2017.

[3] Lisin P.A., Computer modeling of production processes in the food industry, St. Petersburg: Lan, 2016.

[4] Nikitina M.A., Chernukha I.M., "Multicriteria optimization of the formulation of the product," Theory and practice of meat processing; 3 (3): 89–98.DOI 10.21323 / 2414–438X 2018–3–3–89–98, 2018.

[5] Sidorenko M.Yu., Personified Nutrition, M .: DeLi Plus, 2016.

[6] Krutko V.N. et al., "Evaluation and optimization of nutrition using the computer system "Nutrition for Health and Longevity"/ "Computer Science of Health and Longevity". Proceedings of the ISA RAS. Vol. 13. Moscow.: KomKniga., 2005.

[7] Computer program "1C: Medicine. Dietetic nutrition", On the Internet: http: // http://diet.1cp.ru/ Date of access 04/15/2019.

[8] Mosov A.V., Portnov N.M. The methodology for the development of collective diets, Moscow: Eidos, 2015.

[9] Barry C et al, Optimization of the Order Menu in the Electronic Health Record Facilitates Test Patterns Consistent With Recommendations in the Choosing Wisely Initiative / Am J Clin Pathol. 2019 Aug 1. doi: 10.1093/ajcp/aqz134.

[10] Reinivuo H., Laitinen K., Proposal for the harmonization of recipe calculation procedures, EoroFIR, 2007.

[11] Vasquez-Caicedo et al., Report on collection of rules on use of recipe calculation procedures including the use of yield and retention factors for imputing nutrient values for composite foods, EuroFIR, 2008.

[12] FAO food and nutrition paper 92. Dietary protein quality evaluation in human nutrition/ Report of an FAO Expert Consultation, Geneva: FAO, 2013.

[13] Lisitsyn A.B., Nikitina M.A., Sus E.B., "Evaluation of the quality of protein using computer technology" Food industry. - 2016. - No. 1. - P.26-29.

[14] Muratova E.I. et al., Automated design of complex multi-component food products, Tambov: FSBEI HPE "TSTU", 2011.

[15] Musina O.N., Computer design of recipes, M.: Direct Media, 2015.

[16] Nadtochy L.A., Orlova O.Yu., Innovations in biotechnology. Part 2. Food Combinatorics: Educational Method. manual, St. Petersburg: ITMO University, 2015.

[17] FAO Food and nutrition paper 91, Fats and fatty acids in human nutrition. Report of an expert consultation, Geneva: FAO, 2011.

[18] Simonopolus A.P., Visioli F., More on Mediterranean Diets, Basel: KARGER, 2007.

[19] Kossoff E.H. et al., «Optimal clinical management of children receiving dietary therapies fir epilepsy: Updated recommendations of the International Ketogenic Diet Study Group» Epilepsia Open, 3(2): 175-192, 2018.

[20] Nikitina M. et al., «Application of the Hierarchy Analisys Method at the Foodstuff Quality Evaluation» International Journal of Advanced Computer Science and Applications, т. 9, № 5, pp. 51-59, 2018.

[21] Computer program "Monitoring of physical development and nutritional status", on the Internet: http://www.1cp.ru/solutions/NutriMon/ Date of access April 15, 2019.

Node Relocation Techniques for Wireless Sensor Networks: A Short Survey

Mahmood ul Hassan[1], Amin Al Awady [4]
Department of Computer skills, Deanship of preparatory year, Najran University
Najran, Kingdom of Saudi Arabia

Shahzad Ali[3]
Department of Computer Science
Jouf University
Tabarjal, Kingdom of Saudi Arabia

Khalid Mahmood[2]
College of Science and Arts, Department of Information Systems, King Khalid University
Tehama Branch, Kingdom of Saudi Arabia

Muhammad Kashif Saeed[5]
Community College, Department of Information Systems
King Khalid University
Tehama Branch, Kingdom of Saudi Arabia

***Abstract*—Sensor nodes in a sensor network often operate in harsh and challenging environments and this leads to frequent failure of sensor nodes. Failure of sensor nodes leads to partitioning in the network connectivity. For significant effectiveness of applications of sensor networks, the inter-sensor connectivity among sensors is vital. Some sensors are also involved in sustaining the flow of information from the sensor to unapproachable end users. The network can be split up into multiple incoherent blocks and cease working due to physical damage or onboard energy depletion. To deal with such scenarios, a plethora of node repositioning techniques are proposed in the literature. In this article, the recent and up to date mode of research on dynamic sensor repositioning in WSN is discussed. This article classifies sensor repositioning methods into on-demand and post-deployment repositioning based on whether the optimization is accomplished at deployment time or while the network is functioning.**

***Keywords*—Node failure; sensor node; post-deployment; on-demand relocation; internode connectivity**

I. Introduction

Wireless Sensor Network (WSN) is marked as an influential technology and considered as a well-established research area for its wide range of applications. WSN while having certain lower energy and reasonable densely inactive compact mobile nodes is typically organized well in the practices like environment reconnaissance, to assign data and manage communications amongst one another via radio-based technology [3]. WSN is known to be very useful in the environments where human intervention is mostly impractical like wild fire detection, battlefield surveillance, deep ocean exploration, etc. [11]. Mainly, these sensor nodes applications are associated with armed forces, households, or medical care. Moreover, in the armed services, for instance, through its timely practice, defective endurance quality, and independent operation, WSN has revealed its extraordinary applicability relating the systems of armed forces involving inspection, scheming, communication, targeting, and commanding. In health applications, these nodes help monitor the patients and support the patients with special needs (disabled). Besides, these sensor nodes are influential in the areas of commerce and finance like in supply chain management, etc. Moreover, they can also be utilized for monitoring the areas affected by disaster [10].

Fig. 1 shows the sensors random deployment in a region. All of these nodes are responsible for performing the given task independently in a distributed manner and communicate with each other using multi-hop communication. Sensor nodes send back the sensed data to the sink node in a multi-hop fashion as shown in Fig. 1. Eventually, this data is sent to the user by means of internet afterwards [3].

The dynamic relocation of sensors in the operation of the network is very significant to enhance its performance [25]. During the operation of WSN, some of the redundant sensors located at different locations will likely be repositioned in order to replace the sensors with depleted batteries for rectification of network stability [26].

Besides, this type of dynamic repositioning might be appropriate in a mechanism that supports the target movement; also the moveable ones [1]. To further clarify, it approaches forth, eventually gets feasible to reposition particular sensor nodes close-by to the target to attain the improvement of node's data accuracy. In relevant applications [4] and in order to maintain the availability, installing base-station into secure position from ambiguous targets like a tank of an enemy, at the present, could be a better accomplishment [13]. The resolution of the disputation to reposition dynamic sensor node has been analyzed in [22, 23].

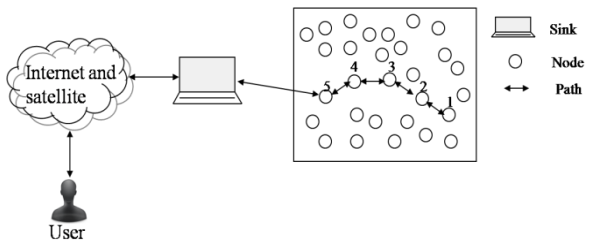


Fig. 1. Wireless Sensor Network (WSN).

This work is sponsored by Jouf University Saudi Arabia, under the research project number 40/117.

A variety of node positioning/relocation techniques are present in the literature. In order to facilitate the research community, it is important to provide them with a comprehensive state-of-the-art knowledge about the recent proposed mechanisms for node relocation. The main contributions of our article can be summarized as follows:

- Varieties of techniques in the literature on dynamic node positioning are characterized by explaining the consequences of dynamic node relocation in wireless sensor network field. Consequently, during the usual network operation or after deployment, this may turn into a practical standard to achieve better improvement in the performance.
- In the sensor node repositioning, it is conceivable to handle the dynamic variation in the network resource surrounding environment as it is dissimilar from the initial deployment. In this article many technical issues are highlighted for sensor node relocation. These technical issues define where to move the sensor node and how to accomplish the network when the sensor node is mobile?
- Several dynamic node positioning techniques are compared and reviewed based on their application, objective, and methodology. As a result, it is supposed that the sensor nodes which are movable can probably be dynamically relocated in order to fill the connectivity and coverage by managing the variations in the user importance.
- The problem of managing multi-node reposition is acknowledged as an open research area and an important attention is paid to this interesting and challenging area.

The rest of the paper is organized as follows. Nodes repositioning issues are summarized in section II. Dynamic node repositioning schemes are discussed in section III and subsequently, section IV concludes this paper.

II. Node Repositioning Issues

In order to improve the network performance, it is necessary to dynamically relocate the sensor nodes while the network is operational. For instance, because of consumption of batteries many sensor nodes stop functioning in the neighborhood of base-station, consequently to improve the lifetime of network replacing the dead sensor nodes by some redundant sensor nodes from other monitored region can be one solution. The dynamic relocation is much valuable in those applications where the target is movable. For example, to increase the sensor’s node fidelity, few of the sensor nodes might be relocated in the sensing region, e.g. the relocating of the enemy tank to the safer region by confirming its availability. Therefore it is very challenging to relocate the sensor nodes during the regular network operation. In response to environment-based or network incentives, such type of relocation is followed. So the continuous monitoring of the nodes in the sensing region is necessary.

III. Dynamic Node Relocation Schemes

Purposefully, for dynamic sensor node's relocation, various techniques have been introduced in the research. These techniques can be categorized into two different types as shown in Fig. 2. The post-deployment relocation (established at the completion of repositioning) undergoes relocation only through the procedure that is implemented at the initiation while the on-demand relocation is the movement carried out by the nodes at any moment of the repositioning of the sensor nodes.

A. Post Deployment Repositioning

Generally, the deployment of WSN in a conventional manner involves locating the sensor nodes either randomly or by the ad-hoc procedure for the appropriate and convenient outdoor application [12]. However, a sensor node or piece of equipment in an industrial setup is designed to be positioned at the same point for several years. Therefore, it is not possible to contemplate the luxury of relocating nodes in a random way. Additionally, connectivity and coverage are essential for sending information to the required destination effectively. To overcome such difficulties, systematic and well-planned design and arrangement is required. This arrangement and planning helps in the reduction of cost and is ultimately time-saving for designing a WSN. Hence, it is necessary to organize and simulate the WSN before deploying it.

VECtor-centered technique (VEC), Minimax technique, and VORonoi-centered technique (VOR) [24] involves the conceptualization of [6, 8] associated with this type. There is a firm bond of all these three techniques to sensor node's Voronoi polygon or position in which the sensor node or the point that is in proximity rather than to the sensory border. In this scenario, the VECtor-based technique can be handled by applying Coulomb's law as the corresponding procedure so as to demonstrate deterrence amongst electrostatic components. The sensor node that is meant to dispossess a constituent of Voronoi polygon of the node subsequently is aimed to get eliminated from other sensors close by via vigour analogous to interlude that it acquires probably from sensors or polygon's angular points.

Fig. 3 shows how the Voronoi technique detect the coverage hole and how to point out the hole size.

Knowingly, for reducing the covered area for the route, [8] has suggested a technique depending on the proxy that lacks the physical movement of the nodes, apart from the computed destination. At this juncture, it is found that the author emphasis to a great extent on the methodology of authentication of movable or stationary sensors. In this regard, the movable nodes are responsible for occupying the level of that point where sensors are missing in distributed pattern evaluated by stationary sensors, indicating that movable sensors are in the consistent mobility and locate stationary sensors in the form of an agent. Following this frame of reference, there will be a major depletion in distance whether from the totality or the calculated set of numbers in which the mobile sensors [10] direct alongside in retaining the coverage at equal rate [7]. By implementing it using a methodology

approved from a widely accepted design for the load balance amongst the sensors in conformations number on parallel computing through the process is separated into multiple parts, concurrently carried out various processors. This strategy, successively, is used in multi-cluster WSN in which the square cell that forms 2-D mesh will specify each single cluster and several nodes added on the single cell shall categorize cluster's load. The position in an inter-connection either in vertical or horizontal indices and the sensor node quantity in its cluster signify that the only communication the cluster-head could have it is by using its same location in other nearby cells. Consequently, the corresponding attained coverage is relevant to the problem of balancing the level of energy to stabilize the distribution of nodes among clusters. Fig. 4 shows the post deployment repositioning techniques.

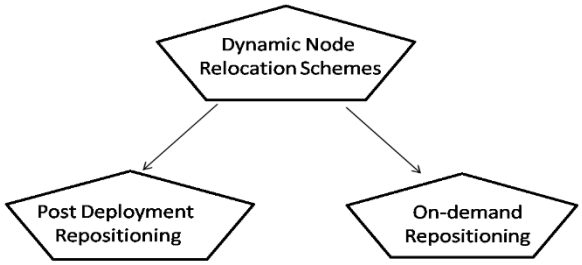


Fig. 2. Categories and Evolution of Dynamic Sensor Node Relocation.

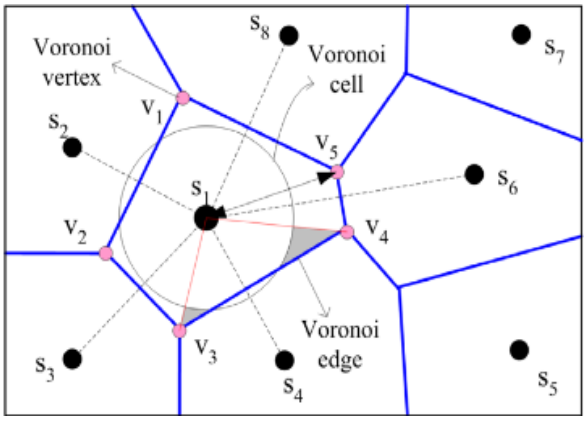


Fig. 3. Voronoi Diagram for the Detection of the Coverage Hole and the Hole Size [6-8].

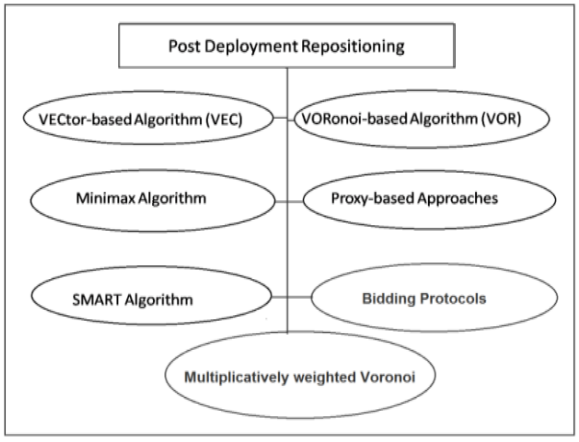


Fig. 4. Post Deployment Relocation Schemes in WSN.

B. On-Demand Relocation

The effectual network for relocation/repositioning emerges as the objective of this category. The effectiveness at the outset of network application lessens essentially for the faulty sensor node. Besides, the essence of the output can also be affected by several fluctuations in the requirements of the application. Moreover, in other cases, it is assumed that sensor nodes should move methodically for keeping the network arrangement efficient. During sensor relocation, rather than repositioning the sensors at the deployment stage, a definite number of sensors are alternatively moved by on-demand relocation to reinforce the specific performance metrics [21, 30].

The existing connectivity restoration techniques [14, 28] mainly consist of two steps i.e. failure detection and recovery. Failure detection is used to discover a node or a component failure. It can be classified into single and collaborative diagnosis depending on the number of nodes involved. Single diagnosis can further be categorized in to self and non-self-diagnosis. In self-diagnosis, a node can only determine some of its component failures. For instance, some of the faults can be determined by the actor node [5] itself such as loss of actor action ability, remaining battery power, communication faults, etc. Non-self-diagnosis involves one of the neighbor actor nodes detecting the failed actor based on the heartbeat messages that nodes exchange with each other as part of the network operation. For example, approaches like PADRA [14, 21] involve one of the neighbors detecting the failure. On the other hand, collaborative diagnosis employs more than one node to determine the failure of a particular actor. It can further be categorized based on the number of nodes involved. Whereas, the MPADRA employs two actor nodes and RIM [22] involves all the neighbors in order to detect the failure.

In [21] authors have addressed the substitution of failed node by putting forward a mechanism called Cascaded Movement, in which, nodes use a recurrent substitution towards sensor close to another redundant one.

Fig. 5 shows the Cascaded Movement where the sensor s1 characterizes redundant sensor. Moreover, sensors s4 and s1 are faulty and redundant sensors. The sensor s3 after finding s4 failure will be searching for the redundant sensor. Using the sensor s2, as per presumption, the sensor s3 discovers s1 that is the redundant node, and the failed node will be superseded by the cascaded movement of sensors s3, s1, and s2, latterly (movements of s1-s2, s2-s3, and s3-s4 occur instantaneously).

Correspondingly, connectivity is taken into account by the further mechanism whereas demonstrated in [15, 23], one technique prefers on keeping up the 2-degree connectivity even under node or link failure depending on the mobility of a subset of the sensor node. While the relocation time can be reduced by using cascaded movement; due to disarrangement of the redundant sensor node, the reviewers are incapable of reducing the time of finding out the redundant sensor node.

In the meantime, post-deployment coverage and connectivity were studied by [20] and [21]. With Connected Coverage and Latency-aware Actor placement (COCOLA), architecture of the network is designed in such a way that is

ranked orderly. Following the aforementioned methodology, high-tiered sensors are relocated, progressively, in order to maximize the coverage but without extension to the data route of a sensor on 1st tier. Purposively, this is for maintaining the desirable bound on latency of data [9]. In spite of that, the influence exerted by faulty sensor note remains unhandled by both COCOLA [21] and C^{2}AP [2].

In [22] authors have proposed RIM (Recovery through Inward Motion) and NN (Nearest Neighbor) techniques; both of which are conformed to the avid heuristics. As a result of failure of node in NN, it will proceed to progress towards Failure Neighbor Node (FNN) being one in close proximity to F position is for restoring the connectivity failure immediately. As a result, a movement of the highly adjacent node among the FNN neighbors takes place, which, subsequently, will position as a replacement of FNN and ultimately the neighboring FNN responds to its elimination.

Fig. 6 provides a diagrammatic representation of working of NN technique in which there exists a topology hole, once node F fails. This is portrayed in the Fig. 6(a). A notification message is sent from node B to D and C previous to it makes a move and H, G and A once it arrives at another location. Moreover, as shown in the Fig. 6, the parts (b) & (c) refer to the interconnectivity of network subsequent to the replacement of node F by B and also B substituted by D. As per illustrated in (part d) it is just node E that gets affected and takes over the initial position of D.

For the purpose of establishing network connectivity, RIM as illustrated in Fig. 7, a procedure is established via inner movement that acquires a prominent method of displaying the inward direction of neighbors of F node to the particular point after the failure occurs in order to set up an interconnection amongst the nodes. This phenomenon is determined by neighbor nodes instantaneously influenced by the failure of F node, and subsequently, once approaching one another, inter-node connectivity will be restored aimed to its former place before the breakdown.

The repositioning is done by recursion to any sensor node that is not relevant to the neighbor node's movement, supposedly, nodes that have moved out towards the damaged one. NN consequently ceases its operation either due to the absence of the neighbor node for a dead sensor (approaching the network borderline), or else, after every node's movement in the network. Provided that RIM enforces the list of one-hop neighbor, the procedure in NN technique, sequentially, postulates that each node should be aware for its two-hop neighbor, subsequently, it can empower to discover the neighbor in the close proximity earlier than the breakdown of F sensor node. In this scenario, there is no relevance of NN and RIM with the aftermath that probably appears in restoring the connectivity of the enclosed network.

Fig. 5. Sensor Nodes Cascaded Movement after Failure.

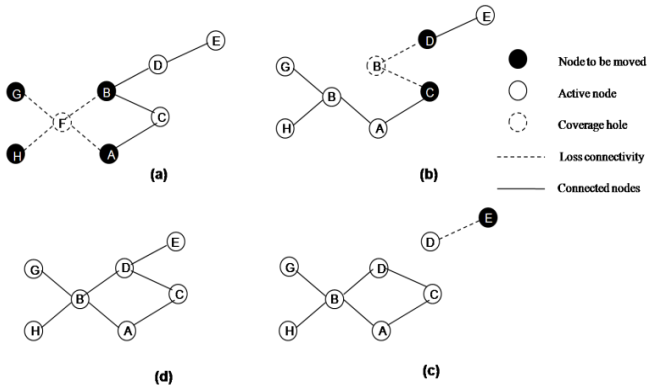


Fig. 6. The Example Illustrates How NN Procedure Reestablishes Connectivity by Substituting Failed Sensor with its Adjacent Neighbor and Iterating this Procedure for Every Sensor that Plants its Location.

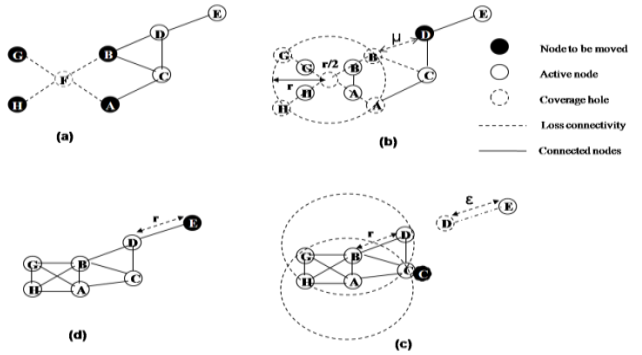


Fig. 7. (a)-(d) is an Example of the Restoration Process of RIM; Every Shaded Sensor Node Moves Depending on its Neighboring Position, Represented in the Circles of Dotted Line.

This cascaded repositioning could be prevented by C^{3}R procedure, also known as Coverage-Conscious Connectivity Restoration [16]. Despite the fact that if replacing the neighbor node by other is restore the connectivity among nodes, it clearly modifies the coverage hole in the sensor's area partially in point of fact, either in the network's interior part or mobile sensor relocation along with border to tackle efficiently with connectivity and coverage from the sensor node's breakdown. The sensors, instead of network topology reconfiguration, work back and forth to substitute failed sensor so as to provide sporadic recovery rather than a permanent one.

Fig. 8 illustrates how C^{3}R operates based on the patterned network. Moreover, it can be seen that the capacity of neighbor node's connectivity with n9 sensor node will be diminished and rapidly goes through the internal movement after detecting the defect of n9. This is carried out to design a pattern that overcomes the breakdown of the node. In addition, as a matter of fact, that n11 sensor node is lying in close range to n9, it swifts speedily than other neighboring nodes towards n9. Besides, n11 serves as a coordinating node as represented in Fig. 8(b). In this scenario, several nodes will assemble the scheduled data from n11 and go back to the respective places excluding n2 node because it serves as an initiating node organized for restoration and acquire a strong overlapping coverage as indicated in Fig. 8(c). Fig. 8(d to g), furthermore, depict to and fro motion of various neighbors of n9 sensor node successively for the recovery of the network from failure of

node n9. This approach noticeably unfolds numerous periodic changes in the inter-connection and possibly can inflict a considerable amount of overhead. It is therefore unambiguous about this mechanism being transient in terms of anticipating the efficiency of the spare nodes. C^{3}R, furthermore, handles only a single failure individually and doesn't cope with the simultaneous and multiple node failures.

The DCRS (Distributed Connectivity Restoration Strategy) [20] has been put forward to deal with the faulty critical node. The mechanism is illustrated in Fig. 9. Due to energy limitation of the sensor nodes, the energy overhead of the repairing process should be reduced to prolong the network's lifespan [27]. For implementation, authors initially designed an algorithm for the identification of two critical nodes that merely rely on the factual data of position of the one-hop and a few two-hop neighbors, and a methodology of choosing the pertinent backup has been introduced for every critical node. A distributed Resource Constrained Recovery (RCR) technique illustrates the procedure of restoring the network connectivity by the strategic relocation of nodes [17].

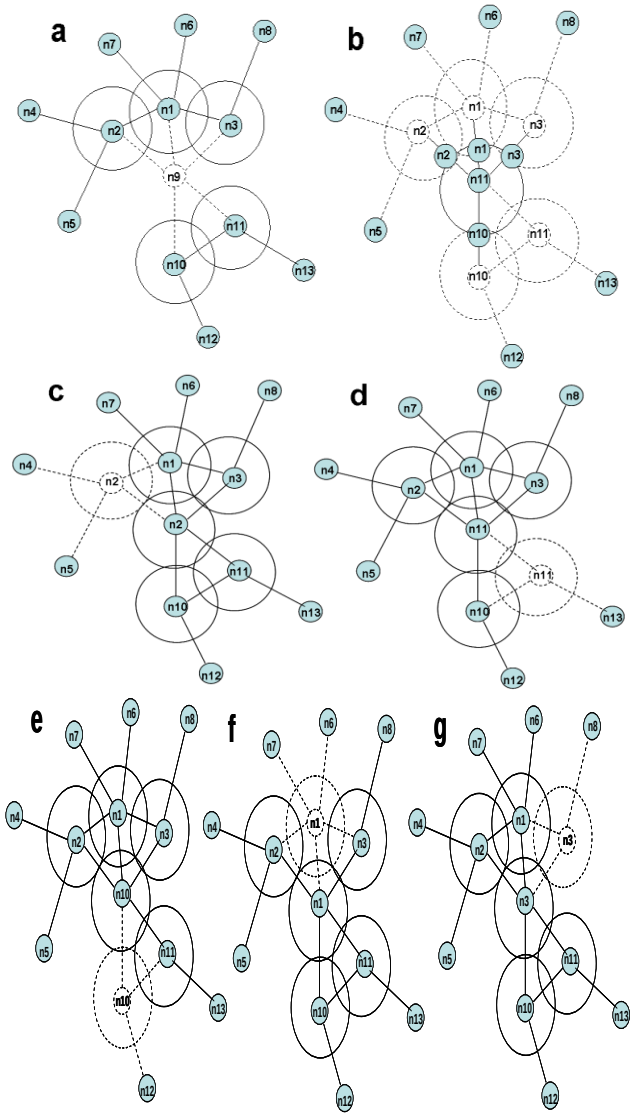


Fig. 8. Working Flow of C^{3}R Technique.

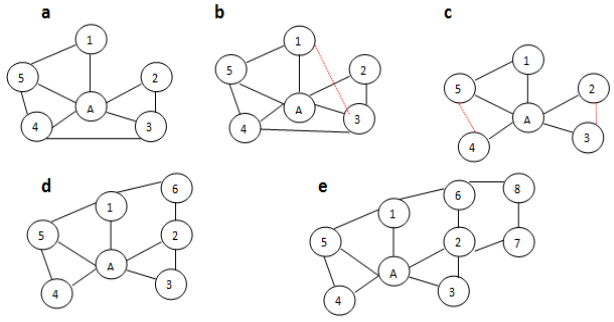


Fig. 9. The DCRS Technique Illustration.

The GSR is proposed that exploits form of deployment region to facilitate the restoration of the connectivity in the partitioned WSN in distributed pattern. This approach disintegrates the area of deployment into the equivalent 2D skeleton outline, with a mobile transmission that is occupied by the survival of the disjoint parts to recover the connectivity [19].

The RRLC-CFP (Restore Relay Lost Connectivity using Concentric Fermat Points) algorithm is proposed for the network restoration which is based on the transmitted node's placement. In this technique network partitioned due to multi-node failure, the problem is considered and solved by a novel approach named Concentric Fermat Points. This approach is used in data transmission for the reduction of the distance of data transmission among the nodes in order to boost the life time of the network [18].

So in this section, different connectivity restoration techniques were presented based on on-demand relocation. In the next section, we present a comparison of all the existing techniques presented in this paper and summarize all the information in tabular form.

Fig. 10 shows the list of on-demand repositioning techniques discussed in this paper.

C. Design Goals for Optimization of a Node Relocation Algorithm

While designing a new algorithm for node relocation, following points must be addressed. The descriptions of these points are illustrated in Table I.

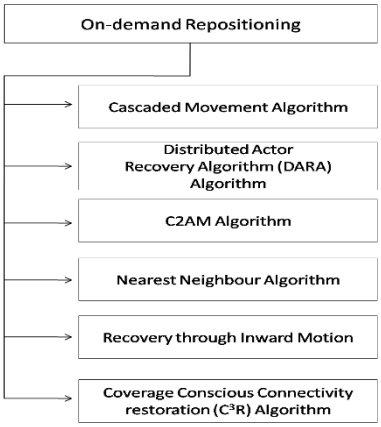


Fig. 10. On-Demand Relocation/Repositioning Schemes.

TABLE. I. CORE OPTIMIZATION POINTS FOR NODE RELOCATION

Optimization Point	Description
Fast hole healing	If any of the sensor node fails, or there is a network hole, then neighbour of the failed node immediately recover this network hole. The technique for identifying network hole must be fast to sustain the network coverage and connectivity for the recovery process.
Minimum message overhead	In order to restore the network connectivity, this method shows the total number of exchange messages by all the sensor nodes. The lower the number of messages the lower the overhead.
Energy efficiency	Energy efficiency is the most important factor of wireless sensor network [29]. If the sensor nodes are the mobile nodes then the primary design objective is to enhance the energy level during the movement and transmission of sensor nodes in sensing region. Different energy efficient algorithms or techniques are used for computing optimum ranges of transmission in order to make the network energy efficient

TABLE. II. DYNAMIC RELOCATION SCHEMES COMPARISON

Reference	Optimization objective	Relocation Technique	Types of nodes	Mobility of Sensors	Number of hops	Limitations/Constraint
8, 24	Restoration of Connectivity	Direct technique	Sensor nodes	Robots	2	Time of convergence, no deliberate coverage
14, 15	Restoration of Connectivity	Shifted technique	Data collector	Robots	2	Time of convergence, doesn't reflect coverage
6	Restoration of Connectivity	Shifted technique	Sensor nodes	Mobile sensors	2	Connectivity not included
25	Coverage	Direct technique	Sensor nodes	Mobile sensors	1	Connectivity not considered
20	Restoration of Connectivity	Shifted technique	Sensors	Sensors	2	Time of convergence, coverage not included
22,28	Restoration of Connectivity	Direct technique	Sensors	Moveable sensors	2	Connectivity not considered
16	Connectivity and Coverage	Direct technique	Sensors	Moveable sensor	1	Multi-node and simultaneous node failures not considered
17	Connectivity	Direct technique	Sensors	Mobile sensors	1&2	Connectivity not considered
18	Connectivity And Network Lifetime	Shifted technique	Sensor	Mobile sensors	1	Multi-node and simultaneous node failures not considered

D. Comparison of Node Repositioning Techniques

Dynamic node repositioning patterns comparison is shown in Table II. All the considered protocols are compared on the basis of various characteristics including optimization goals, migration techniques utilized during the operation, types of sensor nodes considered, the type of mobility of sensor nodes, the number of hops that are considered, and any limitations/constraints associated with the specific works. All the considered works are also provided with references so that if a reader is interested to know more details about a specific work, he/she can refer to the paper and read the full details.

IV. CONCLUSION

A lot of research has been put into the Wireless Sensor Networks (WSNs) due to its immense potential for a variety of applications. However, the resource-limited sensor nodes, ad-hoc network formation, commonly coupled with the unattended procedure, leads to a number of challenges. One of such challenges is restoration of connectivity in case of failure of sensor nodes. This paper elucidates the influence of vibrant node repositioning in the sensor area by incorporating a series of methods and strategies from various researches done previously, that either subsequent to the process of deployment or through conventional operation of the network. Additionally, this research also signifies the number of technical problems encountered during node relocation, indicated as a resolution of the target of repositioning and evaluating the method of network management in relocating the nodes. Successively, numerous publications relating the techniques for dynamic node relocation have been interpreted together with the analysis followed by contemplation of the objective, mechanism, and implementation. A presumption developed here is that the mobile nodes can relocate dynamically in order for filling in the gap in interconnection and coverage and ultimately to cope with the changes. To sum up, the integrated multimode relocation issue has been described and this intriguing and challenging aspect has been elaborated.

ACKNOWLEDGMENT

We would like to express our gratitude to Najran University, Jouf University and King Khalid University for providing us administrative support and healthy research environment.

REFERENCES

[1] Y. Zhang, J. Wang, G. Hao, “An Autonomous Connectivity Restoration Algorithm Based on Finite State Machine for Wireless Sensor-Actor Networks”, Sensors (Basel), 2018 doi: 10.3390/s18010153.

[2] Abbasi, A., Baroudi, U., Younis, M., Akkaya, K., “C2AM: An algorithm for Application-Aware Movement-Assisted Recovery in Wireless Sensor and Actor Networks”. In Proceeding of the International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly, ACM, Leipzig, 2009.

[3] Akkaya K, Senel F, Thimmapuram A, and Uludag S. 2010. Distributed Recovery from Network Partitioning in Movable Sensor/Actor Networks via Controlled Mobility. IEEE Transactions on Computers 59(2):258-271.

[4] Chen, B., Jamieson, K., Balakrishnan, H., & Morris, R., 2002. Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. Wireless Networks, 8(5), 481–494.

[5] Chen, J., Diaz, M., Llops, L., Rubio, B., & Troya, J. M., 2011. A survey on quality of service support in wireless sensor and actor networks: Requirements and challenges in the context of critical infrastructure protection. Journal of Network and Computer Applications (Elsevier), 34, 1225–1239.

[6] Wang, G., G. Cao, T. La Porta, and W. Zhang, Sensor Relocation in Mobile Sensor Networks, in the Proceedings of the 24th Annual IEEE Conf. on Computer Communications (INFOCOM'05), Miami, FL, 2008

[7] Wang G, Cao G, Berman P, La Porta TF. Bidding protocols for deploying mobile sensors. IEEE Transactions on Mobile Computing. 2007 May;6(5):563-76.

[8] Wang, G., G. Cao, T. La Porta., 2004. Proxy-based sensor deployment for mobile sensor networks, in: Proceedings of the 1st IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS'04), Fort Lauderdale, Florida.

[9] Hyewon, J., Wenrui Z, Mostafe H., Ellen W. Zeguar, Chungki L., 2007. Trading latency for energy in densely deployed wireless ad hoc networks using message ferrying, Journal of Ad Hoc Networks 2007; 5(4): pp. 444–61.

[10] Rizzo, A. G., Mancuso V., Ali S., Marsan, A M., Stop and forward: Opportunistic local information sharing under walking mobility, Ad Hoc Networks, Volume 78, 2018, Pages 54-72.

[11] Akyildiz, IF., W. Su, Y. Sankarasubramaniam, E. Cayirci., 2002.Wireless sensor networks: a survey," Computer Networks 38 (4), pp. 393–422.

[12] Chu, M., H. Haussecker, F. Zhao., 2010. Scalable information driven sensor querying and routing for ad hoc heterogeneous sensor networks, The International Journal of High Performance Computing Applications 16 (3), pp. 293–313.

[13] Younis, M., P. Munshi, E. Al-Shaer., 2003. Architecture for efficient monitoring and management of sensor networks, in: Proceedings of the IFIP/IEEE Workshop on End-to- End Monitoring Techniques and Services (E2EMON_03), Belfast, Northern Ireland, pp. 140-146.

[14] Younis, M., S. Lee, A. Abbasi., 2010. A Localized Algorithm for Restoring Internode Connectivity in Networks of Moveable Sensors, IEEE Transaction on Computers, Vol. 59, no. 12, pp 1669-1682.

[15] Heo. N., and P. K. Varshney., 2005. Energy-Efficient Deployment of Intelligent Mobile Sensor Networks. IEEE Trans. On Systems, Man, Cybernetics, Part A, Vol. 35, No. 1, pp. 117-127.

[16] Neelofer, T., Mohamed Y., 2010. Coverage-aware connectivity restoration in mobile sensor networks, Journal of Network and Computer Applications 33, pp. 363–374.

[17] Ma, Guizhen, et al. "Distributed connectivity restoration strategy for movable sensor networks." China Communications 11.13 (2014): 156-163.

[18] Virender, R., Mayank, D., & Kumar, V. A. "Lost Connectivity Restoration in Partitioned Wireless Sensor Networks." Proceedings of International Conference on ICT for Sustainable Development. Springer Singapore, 2017.

[19] Yatish K. Joshi, Mohamed Younis, Feb, 2016 " Exploiting skeletonization to restore connectivity in a wireless sensor network." Computer communications vol.75, Issue C, pp. 97-107, Elsevier Science Publishers B. V. Amsterdam, The Netherlands.

[20] Akkaya, K. and Younis, M., 2007, April. C2AP: Coverage-aware and connectivity-constrained actor positioning in wireless sensor and actor networks. In 2007 IEEE International Performance, Computing, and Communications Conference (pp. 281-288). IEEE.

[21] Akkaya, Kemal, and Mohamed Younis. "Coverage and latency aware actor placement mechanisms in wireless sensor and actor networks." International Journal of Sensor Networks 3, no. 3 (2008): 152-164.

[22] M. Younis, S. Lee, S. Gupta and K. Fisher, "A Localized Self-Healing Algorithm for Networks of Moveable Sensor Nodes," IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference, New Orleans, LO, 2008, pp. 1-5.

[23] M. Younis and K. Akkaya, "Strategies and Techniques for Node Placement in Wireless Sensor Networks: A Survey ," The Journal of Ad-Hoc Network, 6(4): 621-655, 2008.

[24] G. Wang, G. Cao, T. La Porta, and W. Zhang, "Sensor Relocation in Mobile Sensor Networks," in the Proceedings of the 24th Annual IEEE Conf. on Computer Communications (INFOCOM'05), Miami, FL, Mar. 2005.

[25] Moh'd Alia, Osama. "Dynamic relocation of mobile base station in wireless sensor networks using a cluster-based harmony search algorithm." Information Sciences 385 (2017): 76-95.

[26] Ong, Keat Ghee, and Craig A. Grimes. "Network of sensor nodes assemblies and method of remote sensing within liquid environments." U.S. Patent No. 7,319,411. 15 Jan. 2008.

[27] A. Abro et al., "Minimizing Energy Expenditures using Genetic Algorithm for Scalability and Longlivety of Multi hop Sensor Networks," 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2019, pp. 183-187.

[28] Arain. Q , et al., 2018. "Location monitoring approach: multiple mix-zones with location privacy protection based on traffic flow over road networks". Multimedia Tools Appl. Vol 77, no 5.

[29] ul Hassan, Mahmood, Khan, M.A., Ali, S., Mahmood, K. and Shah, A.M., 2018. Distributed Energy Efficient Node Relocation Algorithm (DEENR). International Journal OF Advance Computer Science and Applications, 9(3), pp.95-100.

[30] Mahmood, K., Khan, M.A., Shah, A.M., Ali, S. and Saeed, M.K., 2018. Intelligent On-Demand Connectivity Restoration for Wireless Sensor Networks. Wireless Communications and Mobile Computing, 2018.

Comprehensive e-Learning System with Simulation Capabilities for Understanding of Complex Equations

Kohei Arai[1]
Graduate School of Science and Engineering
Saga University, Saga City, Japan

***Abstracts*—A comprehensive e-learning system with simulation capabilities for understanding of complex equations is proposed. Through experiment with the proposed e-learning system, it is found that the proposed system is much effective and comprehensive than the conventional e-learning system without any mathematical expressions of processing and simulation capabilities by 14.5 %, the time required for learning increased by 18.2% though. In addition, the proposed system not only helps students to understand subjects and complex equations but also can be effective to increase students' motivation to learn.**

***Keywords*—SCORM; e-learning; simulation; Physics subject; action script**

I. Introduction

The use of e-learning is increasing every year. Convenient tools have been provided. Standardization is being promoted by various organizations. SCORM (Sharable Content Object Reference Model) is one of those which is advocated by ADL (Advanced Distributed Learning). There are some standardized e-learning systems such as AICC (Aviation Industry Computer-Based Training Committee), SCORM, and xAPI/Tin Can API (the successor to SCORM) SCORM.

SCORM[1] is a standard adopted worldwide. Its aims are reusability, accessibility, interoperability, and tolerance. Meanwhile, The AICC[2] is dissolved, but having an LMS that supports AICC is still crucial. These universal specifications set forth by the AICC soon reached far beyond the aviation community and into the corporate training world, making AICC the first e-learning standard in the world.

The SCORM standard is all about tracking the status of big and chunky e-learning modules, with the e-learning module and the learner record usually residing in a single LMS or Learning Management System. Tin Can API[3] however, rightly recognizes that most learning happens away from the LMS. So the focus has moved away from e-learning modules towards learning activities, be these offline or online, tutor-led or collaborative, real-world or virtual. It doesn't matter where the activity takes place; what matters is that some remote system with knowledge of that activity can send a simple statement to a central learner record store (LRS) containing some very basic details of what the learner did.

In these standardized e-learning system, comprehensive e-learning contents and procedure as well as processes have to be provided for improving learning performance. Most of e-learning contents for scientific subjects such as mathematics, physics, chemistry, etc. provide formula directly to learners even if the learners are not good at mathematics. It is not comprehensive at all. Therefore, comprehensive e-learning contents and procedure as well as processes in particular for Newtonian equation subject as just an example in this paper.

In order to create a comprehensive e-learning, the proposed e-learning system provides simulation capabilities. There are other ways for improving comprehensiveness of e-learning contents for the lessons of the Newtonian equation. Virtual Reality: VR, Augmented Reality: AR based approaches are typical ways. Motion based Human-Computer interaction based approaches are another ways. The proposed e-learning system can be used together with these typical ways of learning processes, procedure and contents. Namely, the proposed e-learning system provides simulation capabilities followed by AR, VR as well as motion based HCI utilizing system.

The purpose of this research is to develop a comprehensive e-learning system that enables mathematical expression processing and simulation, and to confirm the effect of the developed system. Students can understand complex equations by using simulation capabilities which are included in the proposed e-learning system. It is possible that the coefficients in the complex equations can be changed and also look at the simulation results easily. Thus the students can understand the meaning of the complex equations. The proposed function added to the proposed e-learning system is based on SCORM standard obviously.

The following section describes research background and related research works as well as the proposed method and system followed by the method of experiments. Then the experimental results are described together with some remarks. Finally, conclusion is described together with some discussions and future research works.

II. Related Studies and Research Background

The eLearning Designer's Handbook: A Practical Guide to the eLearning Development Process for New eLearning Designers 1st Edition was published for fundamentals of e-learning system design and e-learning content creations [1]. Adaptation of composite e-learning contents for reusable in smartphone based on learning system is proposed for comprehensive and convenient e-learning system for mobile

[1] https://elearningindustry.com/primer-youll-need-elearning-standards-scorm-xapi-aicc
[2] https://www.elearninglearning.com/aicc/
[3] http://www.open-thoughts.com/2012/10/understanding-tin-can-api/

learners [2]. Practical report on hybrid learning with e-learning and simulation is reported for confirmation of effectiveness of simulations in e-learning [3]. Instructional design for e-learning practice is reported for enhancing an effectiveness of simulations in e-learning [4].

III. Proposed Method

A. Basic Design Concept of the Proposed e-learning System

The basic design concept and the feature of the proposed e-learning system is as follows:

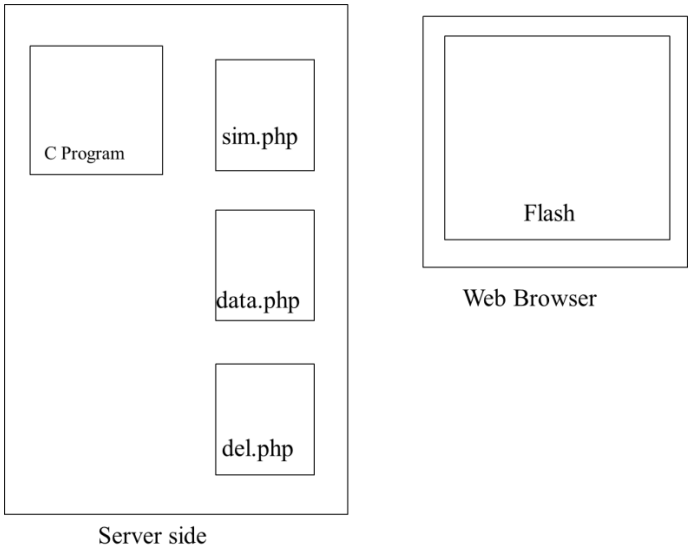
1) Client-Server Model
2) PHP based programing for user interface
3) C programing language is available for users
4) There are three major capabilities, simulation functions (sim.php), data handling (data.php), and data deletion (del.php)
5) Web based browser with Flash

Students can use the proposed e-learning system through Web browser in comprehensive manner. There are many simulation software tools under the sim.php. Therefore, students may use one of the simulation software on their own. If students would like to create their own simulation software tools, they can use the C programing language and register their simulation software tools in the sim.php. Meanwhile, they can use the data under the data.php for simulations. Also, they can delete the data through the del.php.

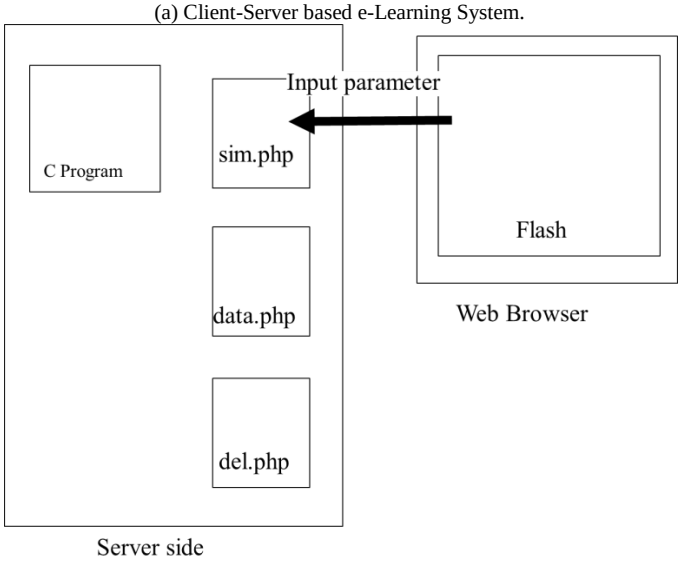
B. Proposed e-Learning System and Procedure

Fig. 1 shows the proposed e-learning system configuration and procedure. It is essentially client-server model. In the server side, sin.php, data.php, del.php are stored together with c programs. On the other hand, Flash based display functionality with web browser is functioned in the client side as shown in Fig. 1(a). Input parameters are input from the client side as shown in Fig. 1(b). Then parameters are transferred to c programs as files as shown in Fig. 1(c). In the same time, coordinate data is created in the server as shown in Fig. 1(d). After that, the coordinate data is transferred as the files as shown in Fig. 1(e). Then, the data file name is transferred to the client side as shown in Fig. 1(f). The data is, then, transferred to server side of data.php as shown in Fig. 1(g). At the same time, coordinate data is transferred to data.php from the file as shown in Fig. 1(h).

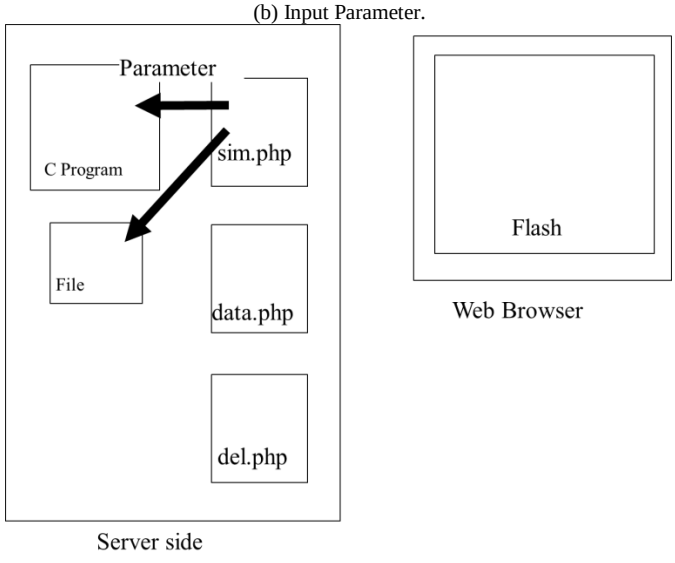
Fig. 1(i) shows coordinate data transfer to client followed by execution of the simulator as shown in Fig. 1(j). Then, data file name is transferred to del.php as shown in Fig. 1(k) followed by data storing through del,php in the files as shown in Fig. 1(l). This is the proposed e-learning system with simulation capability and its procedure.



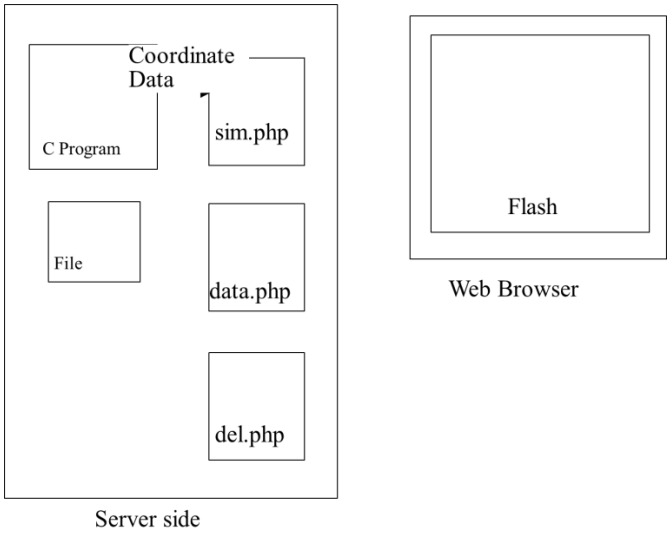
(a) Client-Server based e-Learning System.



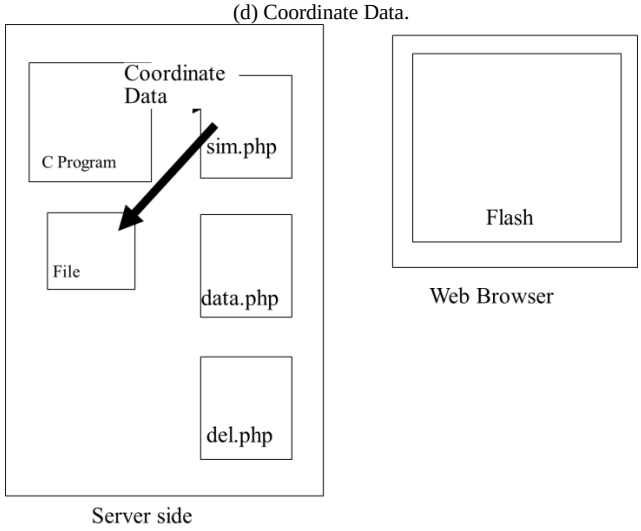
(b) Input Parameter.



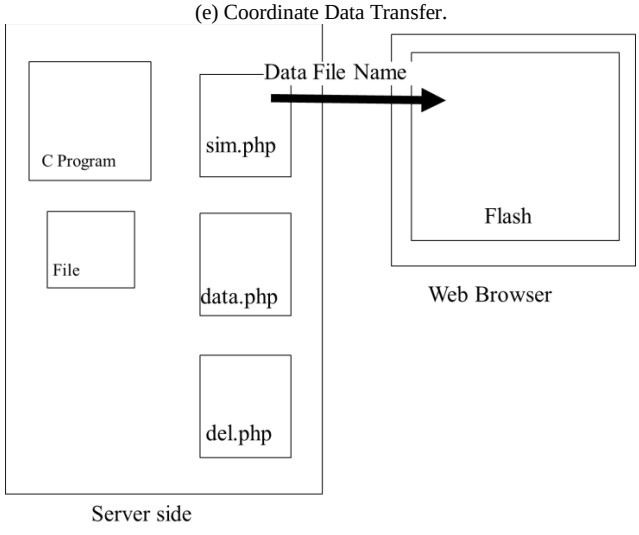
(c) Parameter Files.



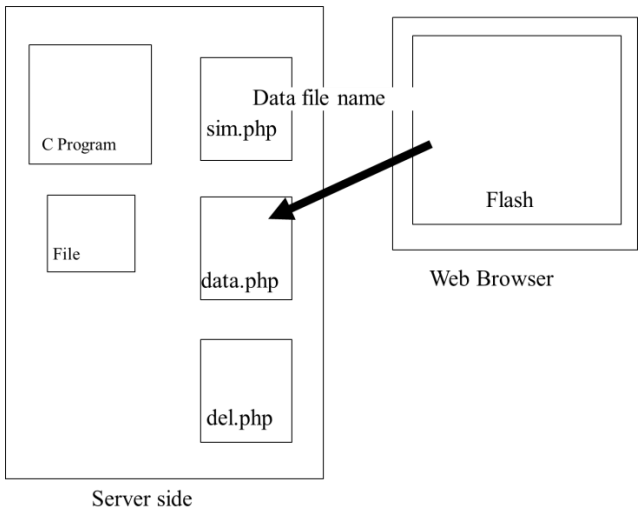
(d) Coordinate Data.



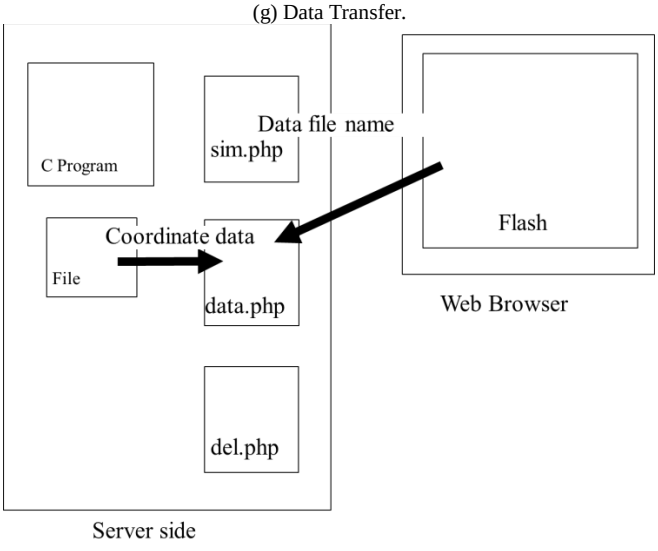
(e) Coordinate Data Transfer.



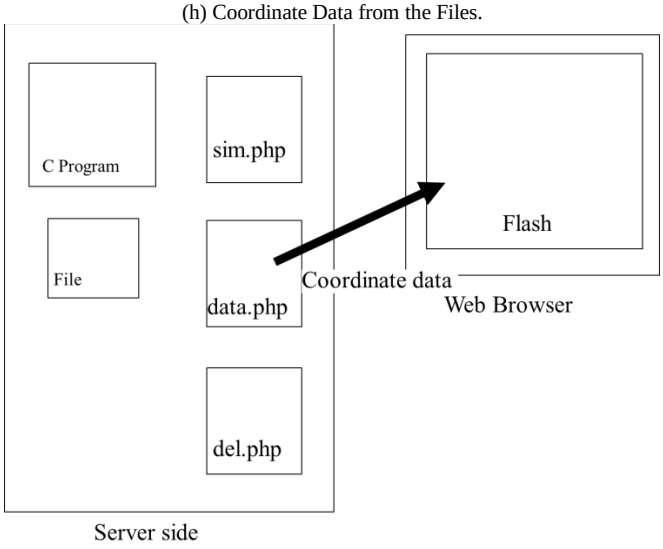
(f) Data File Name.



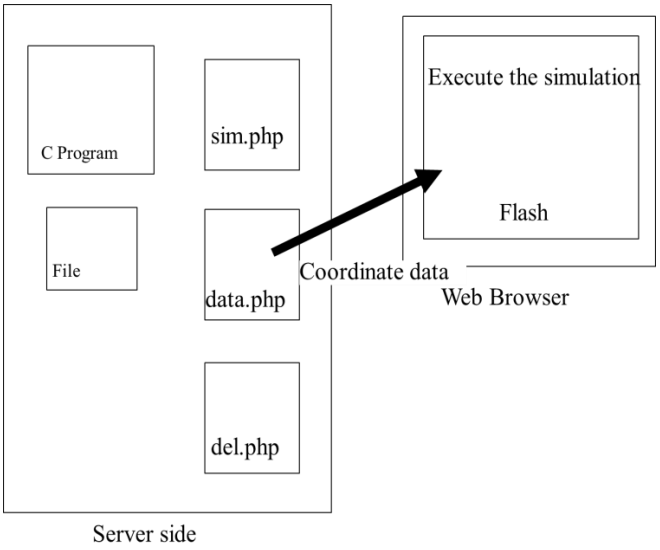
(g) Data Transfer.



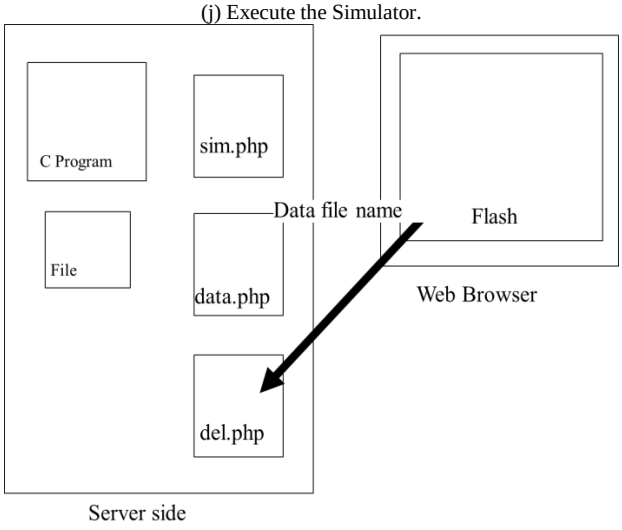
(h) Coordinate Data from the Files.



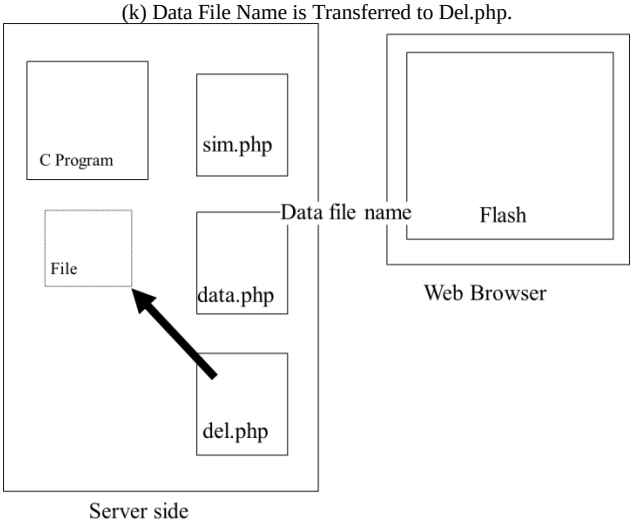
(i) Coordinate Data Transfer to Client.



(j) Execute the Simulator.



(k) Data File Name is Transferred to Del.php.



(l) Data is Stored Through Del,php in the Files.

Fig. 1. Proposed e-Learning System with Simulation Capability and its Procedure.

C. Software Tools and Language used

The tools and language used are as follows.

FlashMX[4] (ActionScript)
User interface
PHP[5]
File operation
Interface between Flash and C program
C, C ++

FlashMX (ActionScript) can create simple animations to complex web applications. The SWF: Small Web Format / Shockwave Flash[6] file made with FLASH-MX can be seen if Flash player (free distribution) is installed. The penetration rate is high (98%). It can be displayed in the same manner on various platforms and browsers, and it is difficult to be influenced by the user's environment.

PHP is free software, there is no need to compile. Also, grammar is similar to C language, easy to understand. In addition, it is a server side script, not dependent on the user's environment.

In this way, since data exchange and file operation between programs are performed by PHP, the system can be reused relatively easily. Fig. 2 shows an example of the snap shot image of the client display.

D. Learning Procedure

Firstly, students choose the subject to learn. The descriptions, the content, and the video content are displayed onto e-learning system display. The description usually contains complex equations. It is not always that the equations are comprehensive. Therefore, students may try to use simulation software tool under the sim.php. Once students choose one of the simulation software tools in the simulation software tool database, the coefficients in the equations can be changed through dialog box in the Web browser. Then, simulation results can be displayed after the simulations. Therefore, students can understand the meaning of the equations in a comprehensive manner.

E. Example of Simulation

One of the examples is introduced here. That is simulating the movement of a satellite around the earth. Students can set the distance from the center of the earth and initial velocity when the artificial satellite is laid horizontally on the ground.

The movement of the artificial satellite around the earth is a two-body problem. The equation of the universal gravitational force acting on the satellite of the mass m is expressed as the motion relating to the relative coordinate vector r of the artificial satellite as follows.

$$\mathrm{m}\frac{d^2t}{dt^2} = -\frac{GMm}{r^2}e_r \quad (1)$$

[4] https://www.updatestar.com/ja/topic/flash%20mx%20 無料
[5] http://php.net/manual/ja/intro-whatis.php
[6] https://dic.nicovideo.jp/a/swf

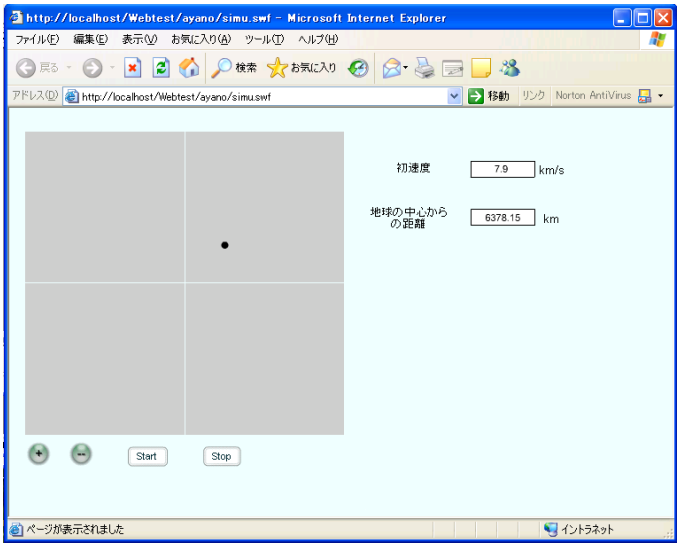


Fig. 2. Example of the Snap Shot Image of the Client Display.

where, $e_r = \boldsymbol{r} / r$ is the unit vector in the direction r, $G = 6.67259\ e^{-11}$ Nm ^{2} / kg ^{2} is the gravitational attraction constant, m is the mass of the satellite, $M = 5.9742\ e^{+24}$ kg is the mass of the earth, r is the equatorial diameter. It is the distance between the center of the earth of 12,756.3 km and the artificial satellite. The satellite size shall be negligibly small. When the above equation is expressed by (x, y) coordinates,

$$F_x = -GMm\frac{x}{r^3} \quad (2)$$

$$F_y = -GMm\frac{y}{r^3}$$

However, the radius of the earth is $r = \sqrt{x^2 + y^2}$. From this equation of motion,

$$\frac{dx}{dt} = v_x m\frac{dv_x}{dt} = -\frac{GMmx}{(x^2+y^2)^{3/2}} \quad (3)$$

$$\frac{dy}{dt} = v_y m\frac{dv_y}{dt} = -\frac{GMmy}{(x^2+y^2)^{3/2}}$$

It can be represented by simultaneous differential equation. If students change the initial speed from this formula using the Euler method [7], Runge-Kutta method [8], students can draw various satellite orbits. In order to fly satellites orbiting the earth on the earth's surface, a speed of about 7.9 km / s is required, which is called the primary space velocity. Although these equations are not comprehensive, variables in the equations can be changed and simulation results can be made available to display results in these equations are going to be understandable comprehensively.

On the other hand, the following questions are effective for deep understanding of the Newtonian equations.

F. Examples of the Questions in a Lesson

Examples of the questions in a lesson for Newtonian equation are as follows:

- Question 1

In the above-mentioned differential equation, the earth escape speed is about 7.9 km / s, and check the satellite orbit by the program to solve this by the Euler method and the Runge-Kutta method. However, there is no need to consider air resistance.

- Question 2

Consider the result of the trajectory obtained at various initial velocities.

- Question 3

Change the step size of the Euler method and the Runge-Kutta method, and investigate the change of the trajectory. Consider the difference in the results of the trajectory calculation.

One of the answers of the satellite orbit trajectory is as follows. Fig. 3 shows the locus of the artificial satellite turning around the earth solved by Runge-Kutta method.

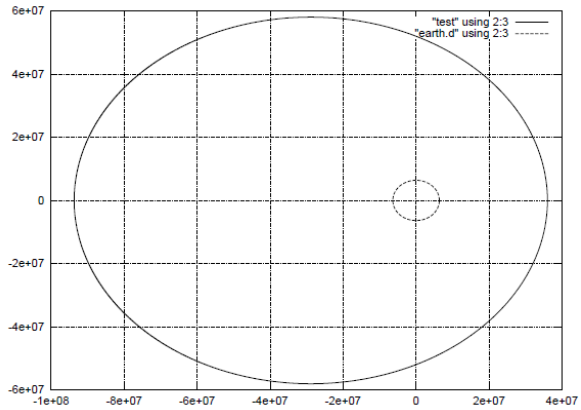


Fig. 3. The Locus of the Artificial Satellite Turning Around the Earth Solved by Runge-Kutta Method.

Thus, the students may understand the meaning of the equations comprehensively and deeply by changing the variables in the equations.

IV. Experiment

Achievement tests were conducted with and without the system created. The 10 questions are there in the achievement test. I also conducted 10 subjects and compared the results. Table I and Fig. 4 shows the achievement test result.

Most of the students gained their score about 9 point with the standard deviation of 14.491 when they use the simulation function. The time required for lesson, however, is increased about 34.3 in average with the standard deviation of 133.911 when they use simulation function.

When using simulation function Compared to when not in use, Eight out of ten people increased (maintain) the score. In addition, learning time has increased more than half. For those who reduced the score when using the simulation function, both the average score and average learning time increased. The reason why the average point rose was thought to be that it was easier to intuitively understand the movement of the artificial satellite by using the simulation function.

[7] http://lpsa.swarthmore.edu/NumInt/NumIntFirst.html
[8] https://en.wikipedia.org/wiki/Runge–Kutta_methods

Regarding the increase in learning time, it is conceivable that simulation must be executed in addition to learning by text, and that the motivation for students' learning has improved due to the addition of a simulation function. Moreover, the tendency that the original score is slightly higher is seen for those who have decreased learning time. Furthermore, I think that the second motivation to learn has been deprived because I took high score earlier.

From the table, it is found that the probability of obtaining a high score is higher as the percentage of learning time with simulation is larger. Moreover, the relation between the time required for lesson and their score without simulation functionalities shows less dependency on the students' potential capabilities (R square value is less than 0.3) while that with simulation functionalities shows much large dependency (R square value is greater than 0.5). This implies that lessons with simulation capabilities give the students very well understandings comprehensively about the Newtonian equations through attractive learning processes.

TABLE. I. ACHIEVEMENT TEST RESULT

	Without Sim.		With Sim.		Difference BTW Both	
Student No.	Required time	Score	Required time	Score	Required time	Score
1	214	70	175	60	-39	-10
2	204	70	387	90	183	20
3	208	60	246	70	38	10
4	307	70	117	60	-190	-10
5	197	70	372	70	175	0
6	209	60	246	70	37	10
7	220	50	108	60	-112	10
8	56	60	260	70	204	10
9	105	40	221	80	116	40
10	275	70	226	80	-49	10
Average	199.5	62	235.8	71	36.3	9
St.Dev.	72.724	10.328	92.294	9.944	133.911	14.49

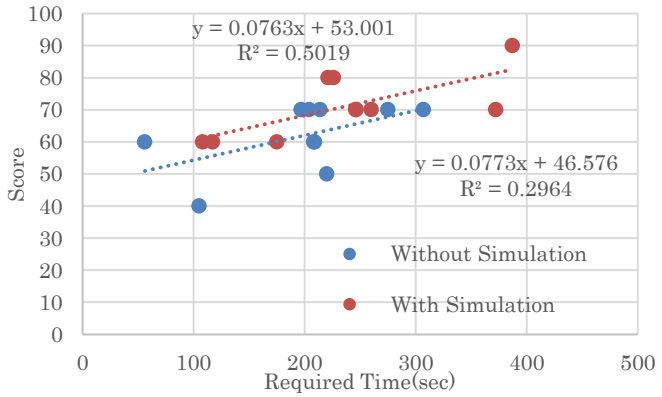


Fig. 4. Achievement Test Result.

V. CONCLUSION

A comprehensive e-learning system with simulation capabilities for understanding of complex equations is proposed. Through experiment with the proposed e-learning system, it is found that the proposed system is much effective and comprehensive than the conventional e-learning system without any mathematical expressions of processing and simulation capabilities by 14.5%, the time required for learning increased by 18.2% though. In addition, the proposed system not only helps students to understand subjects and complex equations but also can be effective to increase students' motivation to learn.

Experimental results show that the e-learning system using the simulation function is effective for the learning of the mathematical system. In addition, this system not only helps students to understand but also can be said to be effective in terms of "motivation to learn".

Further investigations are required for another simulation capability with another subject.

ACKNOWLEDGMENT

The author would like to thank Keiko Ayano for her effort to conduct the experiments.

REFERENCES

[1] Tim Slade, The eLearning Designer's Handbook: A Practical Guide to the eLearning Development Process for New eLearning Designers 1st Edition, ISBN-10: 1983447439, 1983.

[2] Herman Tolle Kohei Arai, Aryo Pinandto, Adaptation of Composite E-Learning Contents for Reusable in Smartphone Based on Learning System, Proceedings of the International Conference on Advanced Computer Science and Information Systems 2014, 502-507, DOI10-1109/ICANSIS 2014, 7065895, IEEE, 2014.

[3] Yoshikazu Hamada, Yoshihiko Suzuki, Go Hasegawa: "Practical Report on Hybrid Learning with e-Learning and Simulation", Research Report of Educational Systems Information Society, Vol.27, No.1, pp.77-80, 2012.

[4] Katsuaki Suzuki, "Instructional Design for e-Learning Practice", Journal of Japan Society for Educational Technology, Vol.7, No.3, pp.197-205 2005.

AUTHOR'S 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 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 37 books and published 570 journal papers. He received 30 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.

Challenges in Wireless Body Area Network

Muhammad Asam[1], Tauseef Jamal[2], Aleena Ajaz[6]
PIEAS University, Islamabad, Pakistan

Shariq Aziz Butt[5]
The University of Lahore, Pakistan

Muhammad Adeel[3], Areeb Hassan[4]
Superior University
Lahore, Pakistan

Maryam Gulzar[7]
The University of Lahore,
Software Engineering Dept., Pakistan

***Abstract*—Wireless Body Area Network (WBAN) refers to a short-range, wireless communications in the vicinity of, or inside a human body. WBAN is emerging solution to cater the needs of local and remote health care related facility. Medical and non-medical applications have been revolutionarily under consideration for providing a healthy and gratify service to humanity. Being very critical in communication of the data from body, it faces many challenges, which are to be tackled for the safety of life and benefit of the user. There is variety of challenges faced by WBAN. WBAN is favorite playground for attackers due to its usability in various applications. This article provides systematic overview of challenges in WBAN in communication and security perspectives.**

Keywords*—WBAN (Wireless Body Area Network); denial of service attacks; resource management; cooperation; security

I. Introduction

Wireless communication brought numerous benefits to our society. Technology up gradation has made this communication possible by the help of 4G, LTE-A, 5G and so on. Recently, Machine to Machine (M2M) communication has been a favorite area of research in past few decades. Communication between machines and the human was next destination. T.G. Zimmerman proposed Personal Area Network (PAN) [1]. Low power, lightweight and miniature physiological sensors has made it possible to connect them to form a Body Area Network (BAN). This connection is supplemented by the wireless technology and the WBAN is formed. This network represent the natural union between connectivity and miniaturization [2].

WBAN comprises multiple sensors. These sensors sample, process and communicate vital sign like heart beat rate, vascular blood pressure and or blood oxygen saturation. Same can be done by the sensors for environmental parameters like location, temperature, humidity and light. These sensors as an attachment with the body and sometime within the clothes. Implants inside the body are getting more attention [3].

Communication network in WBAN can be divided into two major parts or tiers, one is the communication between the sensors and the second is the distribution network as shown in the Fig. 1 [4]. Three-tier architecture is mostly agreed upon by inserting another layer of communication between WBAN coordinator and WBAN gateway or sink node.

In the remaining section describe the architecture of WBAN, while Section 2 details its applications. In Section 3, explain major challenges faced by WBAN. Section 4 discusses the open research issues and our findings in this area.

WBAN communication commonly comprises of three tiers communications as shown in Fig. 2 [5].

- First tier of WBAN architecture is realized by body sensor units which are placed outside or inside of human body. These sensors are responsible for detecting the physiological data signals, converting the signals to digital form and then transmitting through wireless media it get from human body. Then sends it wirelessly to the next tier. This communication is also referred as intra-BAN communication.
- Second tier is comprised of personal server units. These units get data from sensors and process it. This tier formats the processed results to convey to the upper, third tier if necessary. Communication with both the first and third tier is done wirelessly. This communication is also known as inter-BAN communication.
- Third tier comprises of user machines, where end users are data experts who can take some decision, or can conclude some results from this data. This inference may be about someone's health in hospital or at home. It may be sending some caretaker or ambulance to the patient. It may be about taking some specific diet for sportsman. It may be about some artillery movement command from the army head quarter.

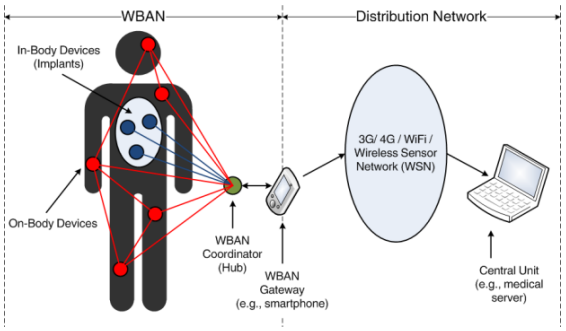


Fig. 1. Two Tier Architecture for WBAN [4].

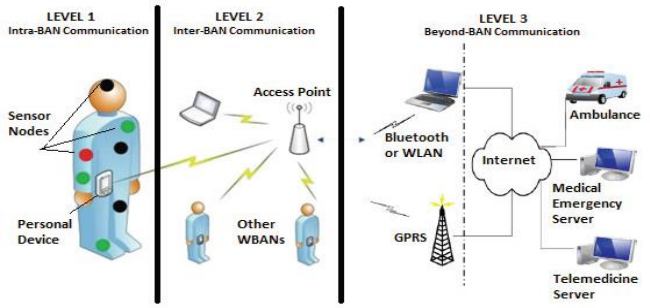


Fig. 2. Three Tier Architecture for WBAN [5].

Despite of the two-tier and three-tier architecture, can be distinguish WBAN entities into two major categories, sensor node and sink or gateway node. Former entity is responsible for data collection from the human body through sensors while the later entity sends it to other servers and communication networks. These communication networks can be mobile network, WLAN, hospital, military's base station or sports training center, etc.

IEEE 802.15.6 on the other hand is an international standard for WBAN, specifies the wireless communications near body or inside the body. It provides low power, short range, and extremely reliable wireless communication within the surrounding area of the body. Different applications can enjoy a vast range of data rates. This standard specifies the wireless communications in the vicinity of body or inside the body. This standard is not limited to humans only but could support any living or non-living thing. It defines the Physical (PHY) and Medium Access Control (MAC) using the frequency bands which are approved by regulatory authorities. This standard considers effects on portable antennas due to the presence of a person (varying with male, female, skinny, heavy, etc.), radiation pattern shaping to minimize the Specific Absorption Rate (SAR) into the body, and changes in characteristics as a result of the user motions. In the next section explained detailed some of its applications.

II. Applications

There are many useful and innovative applications of WBAN. As the WBAN is closely attached to acquire the body parameters so its most favorite applications are in medical field. We classify these applications into two broad categories, i.e. Medical and Non-Medical applications (c.f. Fig. 3).

A. Medical Applications

- A number of sensors are attached to the body like ECG, pulse oxi meter and heart beat sensor on the patient's body. These sensors used to immediately inform the corresponding medical staff about the irregularities and heart rate in advance.
- Cancer can be detected by the help of nitric oxide. Sensor is attached to the affected area, which has the ability to detect nitric oxide emitted from cancer cells.
- WBAN helps to monitor and track the patient's movement which is necessary in home based rehabilitation scheme.
- Allergic sensors used to automatically detect the allergic agents in the air and will immediately report it to the patient or his physician.
- Implanting a bio-sensor in the patient's body to monitor the glucose level and inject insulin automatically when the glucose level is at a certain threshold.
- The solution to all these problems is placing the ambient sensors at home to measure the physiological data of the patient. This data is stored or transmitted to a control unit/healthcare center in regular intervals. This helps the patients to stay at home and get continuous healthcare support without visiting the hospital. Moreover, these sensors, placed on the patient's body, will raise an alarm or urgent notification to the nearby healthcare center in case of any emergency.
- Telemedicine helps remote diagnosis and treatment of patients using telecommunication technologies. WBAN technology can used in telemedicine sector by online consultation of patients with their doctors, transmission of the patient's medical reports and remote medical diagnosis (c.f. Fig. 4).

B. Non-Medical Applications

- Heart rate sensors along with some additional sensors can be used to provide information like speed, body temperature, heart rate, oxygen level, timer and location.
- WBAN can be used to safeguarding the personnel like soldiers, policemen and firefighters. Sensors can be placed on their uniforms in order for them to attain facilities. The WBAN sensors can monitor the level of toxics in the air and warn the firefighters or soldiers if a life-threatening level is detected. WBAN sensors can also monitor health of the uniformed personnel especially soldiers who need medical assistance during war.

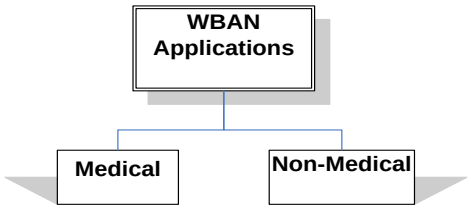


Fig. 3. WBAN Applications.

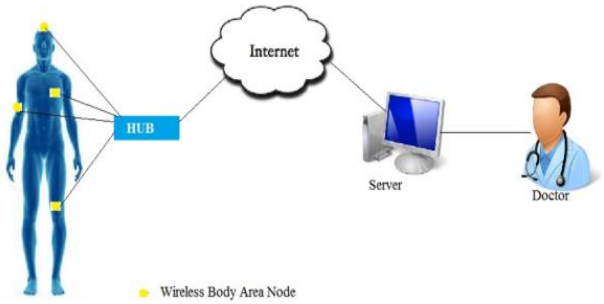


Fig. 4. Application of WBAN in Telemedicine.

III. Challenges in WBAN

Effectiveness of the WBAN is important from both patients and healthcare perspective. As the time passes, challenges to the emerging technologies increases along with the advancements. There is variety of challenges faced by WBAN as explain below. These challenges are classified in six major classes such as energy, mobility, security and communications (i.e., networking, QoS and cooperation), as shown in taxonomy provided in Fig 5. Security is the major issue need to be tackle in parallel with any of other issues.

A. Energy Requirements

Since, most of the devices in WBAN are using the wireless medium, therefore they are portable. Such devices are small in size and carry power source too. Hence, the power is always limited. Wireless natures made them roam free, meaning the devices are free to move. So the power to the device of the network is provided with the help of batteries. Things are not simplified by allowing the power from battery but is encompasses some more challenges of power management of the battery supplies especially in case of implants. Since the sensors that are implanted in the body are so small that the battery cannot sustain more than a month [6]. Removing the implants and re-installation require even more management of the complications generated. Different parameters that alter the power consumption include communication bandwidth and processing power. So there is need to have better scheduling algorithm along with better power management schemes.

Different equipments and sensors for monitoring the body parameters are called body nodes. Each body node has different power consumption profile. To entertain all the body nodes, a reasonable power source is required to work effectively. As a rough estimate, weight of the battery is directly proportional to the power of it. So we may not increase the weight to increase the power as it is to be carried out by the human body and the case is more severe if it is to be implanted inside.

Energy harvesting technique is one solution to the power issue [7]. Energy present in the vicinity of the nodes is converted into electrical energy by the help of specific devices or techniques. The energy harvesting can eliminate the batteries charging either full or partial, based on technique. Such solutions are more clean and green. Vibration, electrostatic, electromagnetic, solar, thermoelectric, pyro and kinetic energy are candidates for harvesting.

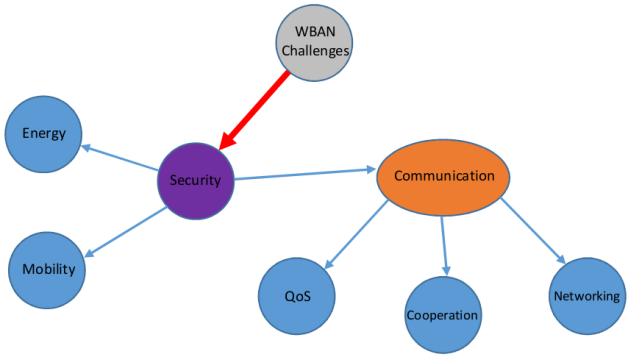


Fig. 5. WBAN main Challenges.

B. WBAN Security

In any network, communication data is of worth importance. In case of WBAN, it becomes more critical as it has been connected to the Physical system. These communication channels are very much visible to the attacker and if not securely implemented it could any of the attack including eavesdropping on traffic between the nodes, message injection, message replay, spoofing and off course compromise the integrity of physical devices. Upon successful attack, such actions not only invade privacy but may lead to catastrophic situation [8, 11]. As reported in Healthcare IT news in February, 2014, hackers accessed a server from a Texas healthcare system, compromising the protected health information of some 405,000 individuals, which was one of the biggest HIPAA security breaches. Even worse, it was demonstrated that implantable cardiac devices can be wirelessly compromised [5]. Security measures are necessary to protect the users from potential risks. Security architecture for WBAN is more challenging than other networks. Efficiency, scalability and usability are performance requirements for the security architecture for WBAN. Regardless of the architecture of the WBAN can coarsely divide the communication into two parts, internal communication between WBAN and external communication between WBAN and external users.

1) Security requirement in internal communication

It includes the following:

- Data authenticity means that data is coming from the claimed source. An attacker may inject bogus data into the WBAN. Public key cryptography schemes are used for data authenticity
- Data confidentiality leads to information disclosure to unauthorized entities. Encryption is also done to achieve this.
- Data integrity is achieved through Message Authentication Code (MAC) or by the help of hashed MAC. Integrity is made sure by doing the reverse process of generating the authentication codes.
- Data availability is the most pervasive security requirement for WBAN. Due to its criticality of the physical system in WBAN, availability of data is made sure. Denial of Service (DoS) attack is the favorite place for the attackers over here.

Along with completing the security requirement, WBAN protocols must be efficient enough to fulfill its desired mandate. In [9, 10], the authors suggested secure and reliable routing framework for WBAN. They demonstrated that it can significantly counter the data injection attacks.

2) Security requirements in external communication: Utility of WBAN in healthcare system may include the self-monitoring patients, network service provider for data transmission, application support and local/remote personnel who offer medical services. Considering the privacy and significance of patient-related data and medical messages, WBAN may suffer threats such as message modification and

unauthorized access. It is desirable that proper security mechanism should be considered for securing the communication between WBAN and external users, where each user must prove their authenticity and then access the data according to their privileges.

WBAN is suitable and useful for different applications and solution. So it is found favorite playground for attackers. One of the classifications of attack is four part communication implementation stack namely PHY layer, MAC layer, network layer and transport layer attacks.

- Being the radio frequency based, PHY is more prone to attacks like tempering and jamming. In jamming attack, attacker transmits radio signal of random frequency. This signal interferes with the other sensor signals. Eventually, node in the range of the attacker cannot communicate message and become isolated. In tampering attack, the cryptographic keys and even program code can be tampered.
- MAC is dealing with the frame detection, multiplexing and channel accesses. Collision attack at this layer may cause in exponential rise in back-off packet in certain protocols. MAC schemes can be interrupted at this layer to cause unfairness attack. Continuous transmission of corrupted packets may result in DoS.
- In WBAN, routing is carried through the coordination of nodes. A compromised node in a network can spoof, alter or replay the routing facts for the network. Sometime the attacker node may selectively route the packets in the network causing selective forwarding attack. A malicious node may attract all the traffic in the network to itself by claiming it to be the best coordinator in the network. It can do alteration with the data received once it is recognized as best data exchange. A single node may pose to have multiple network identities. This results in Sybil attack. An attacker may send a hello message powerful enough to be selected by the nodes to route their messages. This arise the hello flood attack as shown in Fig. 6.
- End to end connection between the nodes are managed at transport layer. An attacker may send a lot of requests to establish the connection to use its all resources. This results in restricting making legitimate connections of the nodes. This is known as flooding attack. An attacker in de-synchronization attack sends fake control flags or sequence number to both nodes in an active connection.

C. Mobility Support

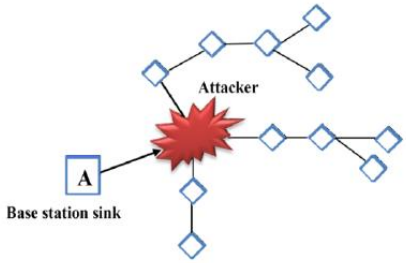
WBAN provides two major advantages, i.e., portable monitoring and location independence. Regardless of the application, these are the key factors due to which WBAN is potential candidate in many venues. But these two advantages put some special limitations i.e., mobility. Mobility can pose serious problem in some application like E-Health care even posture do effect the communication [12].

The mobility is defined between the user and the WBAN as a seamless link. One of the major issues is to reach to sink, which may be single or multi hop. Collier et al. in [13] show that may not stick to single strategy and can find which one is better for the particular case. The same may be applied to optimize the hop count. M. Shanmukhi et al. in [14] proposed a TDMA technique for MAC protocol. Message is flooded to all nodes to reach sink node and the path with minimum delay is selected. Reliable multipath routing is another solution proposed by Birgani et al. in [15]. A path list is maintained depending upon different factors of the routing and the link is established accordingly. Braem et al. in [16] proposed Loose association Implicit reservation Protocol for Mobile WBAN. It works on one hop communication model and has less delay.

D. Quality of Service

Quality of Service (QoS) is the requirements fulfilled by system as requested by the users. For more life critical system, timeliness may be the parameter for the quality. System, that cannot fulfill the said requirement, falls short of providing the QoS. Same is true for other factors like bandwidth, latency, jitter, robustness, trustworthiness, adaptability [17]. Similarly, seamless roaming and end to end wireless connection between the body nodes and the sink nodes is another QoS factor [18].

It is of worth mentioning that system may not be able to fully provide the requested services but the goal of the quality of service may be categorized to Soft QoS, Hard QoS and even no QoS [19]. Challenges to QoS centric WBAN system may be categorized as shown in the



Hello Flood, A Network Layer Attack.

[20, 21].

E. Cooperation between Nodes

When the intermediate nodes help source destination pair in communication, the cooperation occurs. The intermediate nodes may refer as helper or relay as shown in Fig. 7 [22]. Cooperation offers a good solution for many of the limitations in WBAN such as distance, mobility, coverage and channel impairments.

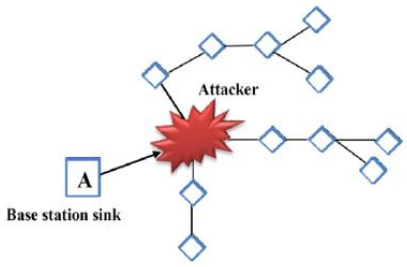


Fig. 6. Hello Flood, A Network Layer Attack.

TABLE. I. QOS PARAMETERS

Parameters for QoS
Limited resources and Capabilities
Scalability
Multi-source multi-sink systems
Node deployment
Dynamic network topology
Various types of applications
Various traffic types
Wireless link unreliability
Real-time system
Data redundancy

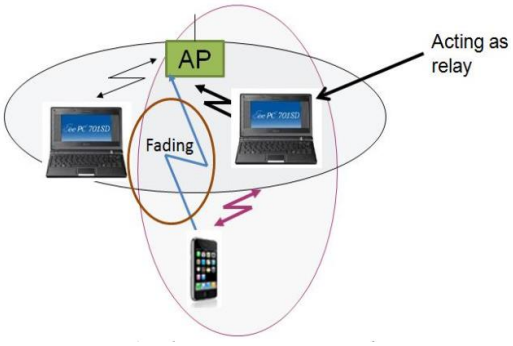


Fig. 7. Cooperative Networking System.

However, introduction to cooperation may leads to additional blockage and overhead. Which limit its benefits and affects the resource management [23].

Communication via relay can add additional interference as well. Therefore, cooperation communication protocols must be devised in way to incur less overhead and interference.

Another issue with cooperation is relay selection. Normally relays are pre-selected, which is based on historic information [24]. Therefore, such relays are not suitable for mobile scenarios. Also, selecting malicious node as relay can make it is easy to launch any attack.

F. Networking Issues

As the size of network grows, it mainly affects the routing protocols performance and throughput of the network. Bandwidth utilization also suffers from links sharing in each connected node; this is one of major cause of slow routing in homogenous channelization.

- Self organization of Mobile nodes in Ad-hoc networks is one of challenging research problem in the context of efficient routing protocols. Multi hop routing is also a promising solution when source and destination are not directly connected to each other. The challenges due to routing protocols are dynamic topology, re-configuration and management, monitoring free, no centralized control, and scalability.
- Multicast routing is a promising solution in ad hoc networks due to frequent attachment and detachment of mobile nodes. Multicast routing is getting special attention in ad hoc networks due to its suitability for link efficiency and central broadcast. Challenges faced by multicasting are lack of QoS, low scalability, frequent updates, delay tolerant multicasting etc.
- The major limitations in WBAN come from limited resource devices and shared medium [25].

MAC layer is very important since it communicate with next hop and access the medium. Therefore, collisions, contention and resource blockage could be handled in case of efficient MAC layer schemes. Hence, backoff algorithms, Carrier Sence Multiple Access (CSMA), contention windows and handshaking need attention while devising MAC protocol for WBAN specially when there is ad-hoc connectivity.

IV. Discussion

This article discussed how different factors affect the performance of the WBAN, its related work and its challenges.

Since health is top priority for all of us. Making the health systems efficient and affective can benefit the human society. Compromising a node could even result in loss of lives (humans or animals). Therefore, security is the utmost challenge for such systems. As a summary, security is needed to be addressed at every level.

The security mechanism must be light weight, since have limited resources. There are various kinds of attacks that can be launched in e-health systems. This could be passive i.e. regarding patients' data confidentiality etc., or active i.e., DoS etc. Issues related to DoS attack is explained below:

A. Denial of Services

Whenever the requested resource is not granted within due time, it can say DoS attack is occurred. DoS attack can active or passive. In former case the attacker launches the attack, while in later case there is depletion of resources. Therefore, DoS attack is very critical to WBAN systems. It can avoid active attacks but better intelligence is required for mitigating the passive attacks such as lazy node behavior etc. There is need to improve the performance of overall network in order to avoid such passive DoS attacks. That's why it can be very easy for adversary to launch DoS attack, as well as it can be launch by itself from inside the network. Therefore, it is very important to design better solutions for handling DoS. This is an open research area within e-health systems.

Most of the attacks we discussed in this paper are kind of active DoS attacks, as shown in Fig. 8.

As discussed earlier, MAC layer is important in term of resource managements. Therefore, depletion of resources leads to DoS attack. One of the solutions in this regard is cooperation, as explained below.

B. Cooperation

As explained earlier, cooperation offers good solution for many problems rises in wireless networks. One of the main advantages is always-connected situation. With the cooperation it can be ensure that patient is always connected even in case of disaster or loss of infrastructure.

Multiple relays assisting source-destination pair could allow higher diversity gain, which could increase the reliability. This way it can achieve performance gain equal to Multiple Input and Multiple Output (MIMO) using single antenna devices without any additional hardware. Cooperation

can also mitigate effects of mobility such that in case of handover or ping pong movement. However, sending data frames via intermediate nodes are not always secure. The relay node may be malicious or its unavailability may lead to DoS itself. Hence, cooperation might itself become security risk.

In our opinion, there is need of good cooperative networking protocol at MAC layer to enable cooperation in case of link failure. Such protocol must be light weight and based on current channel information. Such cooperation can be used instead of layer 3 path re-computation (c.f. Fig 9).

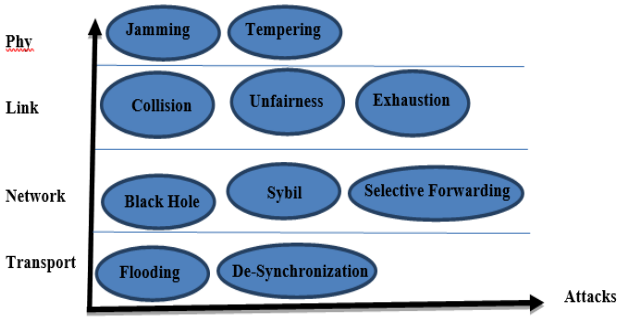


Fig. 8. DoS Attacks on WBAN.

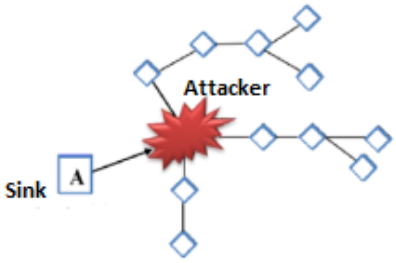


Fig. 9. Attack Mitigation via Cooperation.

V. Conclusions and Future Work

WBAN provides monitoring of human health systems weather inside or outside. However, this is new technology and research is needed to address the conflicting challenges. In this paper we discussed various issues and proposed some taxonomies. However, these issues are categorized into two types: (i) Cooperation and (ii) Security issues.

Security issues need to be tackle at every layer and scenarios of eHealth systems. Similarly, all the networks related issues such as mobility, QoS, energy, routing, and distance could be resolved via cooperation. Cooperation on the other hand adds additional interference, blockage and overhead.

Therefore, in future work we will propose cooperative relaying solution for WBAN to address most of the inherit issues [26]; similarly, as a part of effective security mechanism on various kind of DoS attack's detection and mitigations.

References

[1] T. G. Zimmerman, "Personal AreaNetworks: Near-fieldintrabodycommunication," IBM SYSTEMS JOURNAL, vol. 35, 1996.

[2] M. A Kavitha and S. A Sendhilnathan, "Body area network with mobile anchor based localization," Cluster Computing, pp. 1-10, 2017.

[3] SA Butt, and T. Jamal, "Study of Black Hole Attack in AODV", in Proc. of International Journal of Future Generation Communication and Networking, Vol. 10, No.9, pp. 37-48, 2017.

[4] M. A.-u. Deena M. Barakah, "A Survey of Challenges and Applications of Wireless Body Area Network (WBAN) and Role of A Virtual Doctor Server in Existing Architecture," 2012.

[5] H. S. Sangha and H. Sohal, "Power Challenges in Wireless Body Area Network for Mobile Health Powered by Human Energy Harvesting," vol. 9, December 2016.

[6] A. D and K. K. Venkatasubramanian, "Biomedical devices and systems security,," USA, September 2011.

[7] X. L. X. Liang, Q. Shen, "Exploiting prediction to enable secure and reliable routing in wireless body area networks," 2012.

[8] X. L. M. Barua, R. Lu et al., "Peace: an efficient and secure patient-centric access control scheme for ehealth care system," in Proceedings of the Computer Communications Workshops (INFOCOM WKSHPS), IEEE Conference, 2011, pp. 970-975.

[9] N. Z. C. Hu, H. Li, X. Cheng, and X. Liao, "Body area network security: a fuzzy attribute-based signcryption scheme," IEEE Journal on Selected Areas in Communications, vol. 31, pp. 37-46, 2013.

[10] M. G. Nabi, MCW. Basten, AA., "MoBAN: A Configurable Mobility Model for Wireless Body Area Networks," March 2011.

[11] T. Jamal and Z. Haider, "Denial of Service Attack in Cooperative Networks", in Proc. of ArXiv, arXiv: CoRR Vol. arXiv:1810.11070 [cs.NI], Oct. 2018.

[12] C. J. M. Shanmukhi, "A Review on Mobility Feature in Wireless Body Area Networks," vol. 6, June 2017.

[13] Y. Birgani, N. Javan, and M. Tourani, "Mobility enhancement of patients body monitoring based on," Indonesia, May 2014.

[14] B. Braem and C. Blondia, "Supporting mobility in wireless body area networks: An analysis," Ghent, Belgium, November 2011.

[15] Butt, S. A., Diaz-Martinez, J. L., Jamal, T., Ali, A., De-La-Hoz-Franco, E., & Shoaib, M. (2019, July). IoT Smart Health Security Threats. In 2019 19th International Conference on Computational Science and Its Applications (ICCSA) (pp. 26-31). IEEE.

[16] A. C. W. W. Okundu Omeni, Alison J. Burdett, and Christofer Toumazou, "Energy Efficient Medium Access Protocol for Wireless Medical Body Area Sensor Networks," IEEE Transactions on Biomedical Circuits and Systems, vol. 2, 2008.

[17] M. Asam and Z. Haider, "Novel Relay Selection Protocol for Cooperative Networks", in proc. of ArXiv, arXiv:1911.07764 [cs.NI], November 2019.

[18] N. U. Shah Murtaza Rashid Al Masud, P.O. Box 1988 and S. A. Najran, "QoS Taxonomy towards Wireless Body Area Network Solutions," International Journal of Application and Innovation in Engineering Management vol. 2, April 2013.

[19] T. Jamal, and SA Butt, "Malicious Node Analysis in MANETS", in Proc. of International Journal of Information Technology, PP. 1-9, Springer Publisher, Apr. 2018.

[20] T. Jamal, P. Mendes, and A. Zúquete, "Relay-Based Cooperative MAC Protocol," in MAP TELE Porto, Portugal, June, 2013.

[21] T. Jamal, P. Mendes, and A. Zúquete, "Analysis of Hybrid Relaying in Cooperative WLAN," in IEEE/IFIP WirelessDays, Nov. 2013.

[22] T. Jamal, P. Mendes, and A. Zúquete, "Relayspot: A Framework for Opportunistic Cooperative Relaying," in IARIA ACCESS, Luxembourg, June, 2011.

[23] T. Jamal, P. Amaral, and A. Khan, "Denial of Service Attack in Wireless LAN," in ICDS, Rome, Italy, 2018.

[24] T. Jamal and M. Alam, "Detection and Prevention Against RTS Attacks in Wireless LAN," 2018.

[25] Z. Haider and K. Ullah, "DoS Attacks at Cooperative MAC", in Proc. of ArXiv, arXiv:1812.04935 [cs.NI], Dec. 2018.

[26] Butt, S. A., Jamal, T., Azad, M. A., Ali, A., & Safa, N. S. (2019). A multivariant secure framework for smart mobile health application. Transactions on Emerging Telecommunications Technologies, e3684.

e-Parking: Multi-agent Smart Parking Platform for Dynamic Pricing and Reservation Sharing Service

Bassma Jioudi[1], Aroua Amari[2], Fouad Moutaouakkil[3], Hicham Medromi[4]
Research Foundation for Development and Innovation in Science and Engineering, Engineering research laboratory (LRI)
System Architecture Team (EAS) National and High School of Electricity and Mechanic (ENSEM)
Hassan II University, Casablanca, Morocco

***Abstract*—Parking is a key element of a sustainable urban mobility policy. It plays a fundamental role in travel planning and transport management, as the foremost vector of modal choice, but also as a potential means of freeing up public spaces. In this article we define the smart parking concept, as an application of smart mobility, present a historical analysis of the evolution of smart parking framework and show a statistical analysis of the published patent applications in this field around the world using the ORBIT database. Then, we propose a new smart parking architecture based on multi-agent features. Finally, we introduce the e-Parking system, platform to improve the driver experience of crowded cities. It provides real-time parking prices and offers a reservation and guidance services. In addition, the system assigns an optimal parking for a driver based on the user's requirements that combine proximity to destination, parking cost and dwell time, while ensuring a fair sharing of public space among users and improves traffic conditions. Our approach is based on dynamic pricing policy. Our scheme is suitable for mixed-usage areas, as it considers the presence of reserved and not reserved driver in the same parking area.**

Keywords*—Smart cities; smart mobility; smart parking; dynamic pricing policy; cruising traffic

I. INTRODUCTION

Throughout the 20th century, the private car is gradually becoming the most favorable transport mode. Mobility experts predict that the number of motor vehicles in densely populated areas will increase significantly over the last 30 years which will greatly change the way the territory is managed. The mostly influenced element by this change is parking.

Indeed, urban areas suffer from increasing traffic congestion, followed by economic, social and environmental damage. In addition, in areas where parking capacity is limited and parking information is not available, the situation becomes more challenged.

Moreover, the lack of parking spaces combined with inefficient parking pricing policies increase the difficulty of finding available parking spaces, induce additional traffic congestion and increase the environmental impact of car journeys [1]. Surprisingly, Shoup [2] found that about 8 to 74% of traffic in Central Business Districts (CBD) is due to vehicles looking for parking spaces and the average time to find an available space is about 3.5 to 14 min.

Known as one of the leading causes of congestion in big cities, looking for available parking spaces process has attracted many authors [1]–[10]. The common objective of these works is to reduce as much as possible the time lost in searching for an available place and eliminate the negative consequences on the urban environment and the quality of life.

However, the large number of drivers looking for parking spaces is mainly due to:

- Lack of parking information, so that drivers conduct blind research to find an available parking lot;
- Driver preferences for free or cheap parking lots. So that they move away from the off-street parking to go to the on-street parking lots, which causes many problems of urban traffic;
- Drivers who occupy attractive parking lots for a long time (commuters) and prevent the rotation of parking spaces. Indeed, commuters who arrive early, take the attractive parking places and drivers arriving during the day do not find places available near to their destinations.

To address these problems, we may introduce smart parking solutions, through guidance and booking services. However, this type of solution has some limitations:

- Driver should be informed that the target area only receives customers who have booked in advance, in order to avoid additional flows (drivers without reservations);
- Driver should control his parking dwell time to avoid overcharging;
- Another problem with reservation systems is the phenomenon of multiple-car-chasing-single space, so that the customer who arrives and finds his place occupied by the last driver who has parked previously or by another who has not reserved; and no place is available in the parking, will be asked to leave without parking, which would cause frustration among drivers;
- In order to guide a driver to a vacant space, an availability parking detection system is required. However, the cost of implementing such a heavy solution is important, especially on the streets, given the number of equipment to be installed per lot (sensors, cameras, retractable bollards, etc.).

Based on the above findings, pricing policies are obviously an effective control tool to reduce the search for parking spaces and control parking occupancy and dwell time. In this article, we focus on the smart parking concept as an application of urban mobility that aims to resolve traffic problems. We present a historical analysis of the evolution of the concept and statistical analysis of the published patent applications for smart parking in the ORBIT database around the world. Then, we discuss our proposed smart parking architecture based on multi-agent features. Finally, we highlight the features and services of a Smart parking Web application developed in the Foundation of Research, Development and Innovation in Sciences and Engineering of Casablanca, Morocco (FRDISI). The proposed concept aims to improve the driver experience by combing real time pricing and reservation sharing service.

This article is organized as following: In Section II, analysis of smart parking concept and paper problematic are presented. In Section III, we are interested to the proposed solution architecture and the role of each agent in the system. Then, in Section IV, we present the e-Parking smart parking platform and the different steps that user can be experienced. Lastly, we conclude our work.

II. State of Art

The degree of urbanization in the world is continuously increasing. According to the United Nations forecasts, increasing of urbanization combined with demographic growth will contribute to an increase in the population of cities by 2.5 billion people. In 2013, 50% of the world's population lives in cities, and this proportion is expected to reach 70% in 2050 [11]. In the report on demography in Morocco, elaborated by the Center for Demographic Studies and Research (CERED) [12], it is estimated that Morocco will reach its stationary population (45 million inhabitants) by the year 2050. Overall, during the century from 1960 to 2060, the urban population of Morocco would be multiplied by 10. The rate of urbanization which was 29.2% in 1960 would then be 71.6% according to these projections [13]. This is mainly due to the natural and migratory movements, the urban population increased rate, which creates unprecedented opportunities, but also new difficulties in urban management, that can lead to a strengthening of environmental degradation, poverty and exclusion [4].

In view of these findings, it is imperative for the authorities to develop new approaches and take the necessary measures to implement a territorial development policy to find new reception possibilities, ensure balanced spatial sharing of the population and related economic activities, reduce the environmental impact of cities, which make cities more intelligent and sustainable. This has given rise to the emergence of the smart city concept.

Smart cities constitute a priority concern for large cities leaders, both for the economic development of territories and for the well-being and living environment of citizens. A smart city stands out as a city that takes full advantage of digital transformation as well as Information and Communication Technologies (ICT) to improve public services and make them more personalized and efficient. As a concept, Smart City aims to provide a comprehensive state-of-the-art service platform that aims to improve the quality of life of citizens and the services offered by cities through the adequacy of existing infrastructure, citizen participation and, ultimately, to enable sustainable development of the urban environment in the future [14]–[16]. Article [17] presents the different definitions proposed and states that the concept of a smart city is no longer limited to the diffusion of ICT but consider the needs of citizens and communities.

Given the variety of definitions of the smart city, the report [14] summarizes the different domains and sub-domains identified in the literature according to six dimensions. In line with our research focus, smart mobility is based on the integration of ICT infrastructure, sustainable transportation and logistics systems to support improved urban mobility and prosperity of the urban environment. It allows to: notify drivers in advance of road dangers before they become visible; improve driver safety, for example by keeping vehicles at a safe distance from each other by suggesting optimal speed based on various parameters related to traffic conditions; use communications, including vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) as special cases; inform drivers of the local speed limit; share information specific to an area (parking lots, petrol stations, etc.); provide users with real-time information on public transport services and real-time travel and traffic information, as well as intelligent and transparent ticketing solutions; enable users to plan their trip in a multimodal way; meet the needs of freight operators and customs authorities, this is accompanied by tracking the position and condition of shipments, as well as the orientation of their vehicles along the most efficient, economical and safe route; maximize vehicle efficiency and reduce CO2 emissions by choosing the optimal route and speed to the destination.

Global trends in urbanization and motorization and the influence on climate change make urban mobility sector a challenging task for traffic planners. The major element that the automobile has changed in the use of the ground is parking.

Parking constitutes one of the sources of environmental nuisance and degradation of living environment. This is particularly the case in dense urban areas where the lack of free places significantly affects the fluidity and traffic conditions due to drivers looking for parking space. Beyond these externalities, parking poses significant issues related to land use, occupying public space, pedestrian safety and social equity and it strongly determines the accessibility to different urban functions.

Companies and investors consider parking as an economic necessity and an important opportunity to attract customers to their business. A convenient parking offer is also necessary to retain staff and employees. In the sense that drivers always want to park as close as possible to their destinations, while ensuring safety at the best possible cost. City residences consider parking with a certain degree of emotion. They consider the space near to their residence as private property. From a pedestrian perspective, parking may be undesirable if it blocks walking trails or increases walking distances. Pedestrian safety is also challenged by unplanned or poorly designed parking lots.

Today it seems admitted by authorities and transport planners that parking issues are much more complex and go beyond the simple lack of parking supply. However, due to the increasing demand for parking spaces in cities, finding a parking space is becoming a common challenge faced by millions of drivers every day. Although vehicles are stationary for 95% of the time, and therefore park for most of their life [18], the severe shortage of parking places has created a challenge and problem in managing these areas.

In article [19], four methods of searching parking spaces are cited: blind search (the most common method), the parking information sharing, buffer parking information sharing [20] and reservation-based smart parking systems [21]–[23]. Each of these methods has advantages and conveniences. The last three methods rely on the integration of information technology into the parking process via smart parking platforms.

Smart parking, application of smart mobility in smart cities, represents a potential solution to solve traffic and parking problem. Smart parking roles and key features have been widely discussed in the literature (see, for example, [3], [7], [24]–[33]) in recent years.

A. *Smart Parking Systems*

Smart parking system is becoming an alternative solution to solve parking problems. Fig. 1 illustrates the main components of a smart parking system:

Vehicle entry/ exit detection system: Identification of vehicles in checkpoints is the act by which a vehicle is authorized to enter/exit a restricted area. It is based on the detection of an identifier (i.e. the license plate of the vehicle) and compare it with a pre-established database. Access to the parking is generally managed by issuing a ticket with a digital code or bar code that has a configurable exit time or by the automatic identification of the driver through a token or a specific card. It facilitates several operations such as access control by validating, for example, access only to vehicles already reserved or payment of parking fees by recording the access and exit time of vehicles. Several technologies may be used to identify vehicles. We distinguish.

- Methods that require vehicles to be equipped with embedded devices (RFID, Bluetooth, NFC, etc.);
- Methods that potentially identify all vehicles without installing equipment in vehicles i.e. ALPR technology as an effective solution for identification since each vehicle has unique information.

Vacant parking lot detection system: The monitoring of parking spaces can be used to notify the driver of the number of places available in a parking, to guide drivers to the vacant places and to offer the service of reservation. Techniques can be divided into two types: the first suggests estimating the overall status of the parking, for example by counting incoming vehicles; the second supports the verification of the presence of a vehicle in each parking space to guide drivers to vacant places.

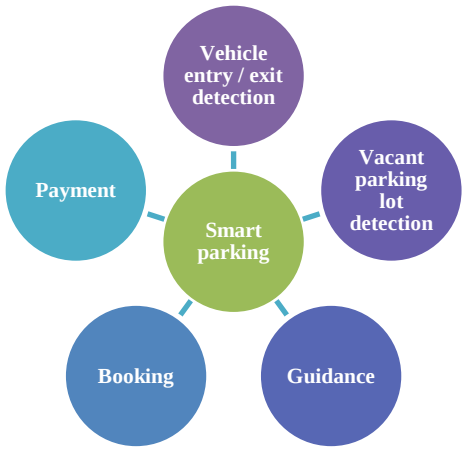


Fig. 1. Smart Parking Components.

Table I illustrates some of vacant parking detection systems and their implementation exigences.

Guidance system: Informs drivers about the availability of parking spaces in real time, via two tools: a mobile application and/or dynamic information panels. It informs drivers about parking capacity, price, parking levels, availability of spaces, directions to vacant spaces. The status of the parking spaces, identified via a detection element, is transmitted in real time to the central data processing and management system, which updates the variable message signs using digital network communications.

In order to ensure clear visibility, the choice of the placement of information signs panels in a parking is essential. Table II shows the types of signs that can be used in parking.

TABLE. I. VACANT PARKING DETECTION RESOURCES

Resources	**Implementation environment**	**Installation**	**Detection type**
Radar sensors	On-street lots, Open outdoor parking lot, Covered parking with high ceiling	Urban lighting columns, facade	RADAR, no image acquisition.
Indoor sensors	For all underground parking lots	Parking Ceiling	Ultrasonic
Outdoor sensors	Ground	on the ground either Semi-recessed, recessed or surface mounted	Infrared, magnetic
Camera	large-scale parking	on building or high mast	Image acquisition and processing
Minute stop station	On-street lots, aeroports, private areas	On street ground	Magnetic
Combined technologies	Combination of more than one detection resource, e. g. video camera and indoor sensors.		

TABLE. II. THE TYPES OF SIGNS IN PARKING.

Types	Description
Plinth external signs	Double or single-sided signage mounted on the ground or on a wall and generally used at the entrances of multi-level parking, or at the end of the aisles of outdoor parking
Wall signs	Provide directions and/or place numbers
LED light panels	Designed for internal marking and mounted on the ceiling
Application mobile	Installed on a mobile device or on-board computer

Booking system: Booking or Reservation is the action of retaining an empty place in a parking for a fixed period of time in order to occupy it. The reservation depends on the architecture of the parking management system, which can be either centralized or distributed. In the first case, the reservation is often made through an application installed on an intelligent mobile device (smartphone, tablet, iPad, etc.). The user sends a request for an available place to a parking management and control system. The management system processes the request and sends a response to the user based on the data and information available on the parking status. In the second case, the use of vehicle-vehicle communication using Ad-Hoc vehicle networks (VANET) is used. There are different reservation systems for intelligent parking, we have been able to classify them into two types: booking systems using the agent approach and non-agent booking systems.

Payment system: Is the purchase transaction that effectively brings a monetary contribution for the parking service. Several payment methods are available and are divided into three classes:

- The user moves: Parking meter;
- The user does not move but performs the remote payment: payment by SMS, on a smartphone application, or in advance (reservation). The user communicates by SMS the license plate number and the code of the parking area concerned;

Payment does not require any user action or movement: payment is made by automatic recognition, via a license plate detector or a Dedicated Short-Range Communications (DSRC) device.

B. Technology Trends for Smart Parking Systems Worldwide

The statistical analysis of the published patent applications for smart parking in the ORBIT database generated several results in the order of 28,474 patented inventions.

The distribution of patent applications in the field of smart parking by country of publication (Fig. 2) reflects the potential markets for technology. Indeed, the fact that an applicant protects its patent application in a country other than the country of filing shows interest and potential profitability of that patent application in that territory. China, Japan and the United States of America remain the most attractive countries in terms of extending the protection of patent applications with 8143 families, 4913 families and 3454 families, respectively. The graph above shows that the African continent is still widely perceived as an inhospitable land for smart parking technologies. The only attractive market for this technology is South Africa by 106 families.

Fig. 3 presents the top depositors in volume on the analyzed theme. This type of representation shows the players who own the biggest portfolios in the field of smart parking. As can be seen, the leading companies in this field occupy the first places in the list of TOP patent holders (Amano, IPS, Panasonic, Dayang Parking, Robert Bosch, Hangzhou Xizi Iuk Paring System). This is confirmed by the smart parking market study published on September 19, 2018 by Market Research Future.

In most modern cities, drivers are provided with guidance systems (variable message signs, parking status displays, etc.). However, despite this, the traffic system and drivers face extreme difficulties, especially during peak hours, or at special events such as festivals. Ultimately, in many urban areas, it is tedious to find an available parking space and, when this is done, it is difficult to know if it meets the expectations of drivers and parking managers.

Thanks to the development of smart reservation-based parking systems, drivers can check the availability of parking in real time, book a place remotely and are no longer required to be physically present to perform these operations.

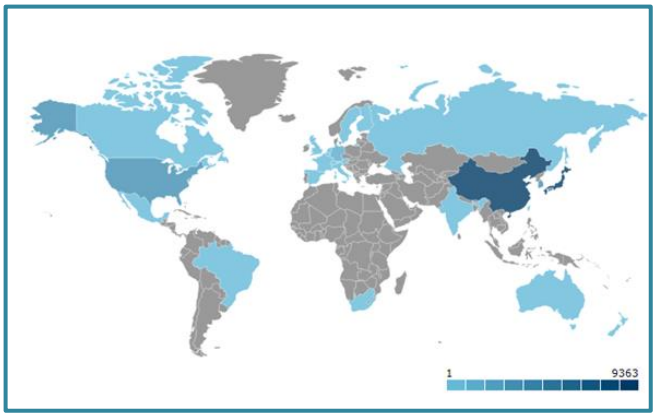


Fig. 2. Countries Producing Smart Parking Technologies.

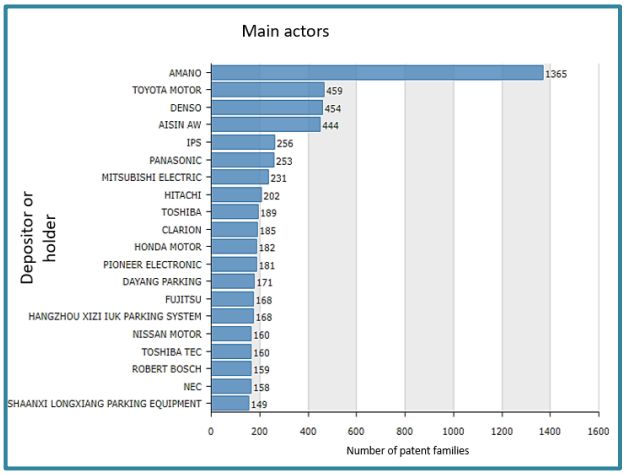


Fig. 3. Top 20 Active and Innovative Structures in the Field of Smart Parking.

For efficient use of such a system, the driver should be informed that the target area only receives customers who have booked in advance, in order to avoid additional flows (drivers without reservations). In addition, the driver should control his parking dwell time to avoid overcharging. Another problem with reservation systems is the phenomenon of multiple-car-chasing-single space, so that the customer who arrives and finds his place occupied by the last driver who has parked previously or by another who has not reserved; and no place is available in the parking, will be asked to leave without parking, which would cause frustration among drivers. Finally, the cost of implementing such a heavy solution is important, especially on the streets, given the number of equipment to be installed per lot (sensors, cameras, retractable bollards, etc.).

Based on the above findings, pricing policies are obviously an effective control tool to reduce the search for parking spaces and control parking occupancy and dwell time. Indeed, by achieving optimal occupancy of the parking (around 70% and 80% occupancy), the problems associated with reservation can be overcome. In addition, the difficulty of obtaining real-time parking information on public roads, unlike off-road parking with detection systems (sensors, cameras, etc.), makes pricing policies a powerful tool for managing supply and demand for parking in dense/very dense urban areas.

C. Parking Pricing Policies

Parking pricing policies constitute a key factor in influencing driver decisions and solving traffic problems without a dedicated infrastructure. In other words, effective pricing policy ensures, through financial constraint, an attractive incentive mechanism in view of the parking supply and maximizes the advantage for the parking operator. However, the adoption of a pricing policy is also a sensitive measure, as it can reduce and/or induce travel demand [34]. It has been proven that the availability of parking places at the workplace and the absence or low cost of parking greatly promote the use of personal cars [35].

Parking pricing has attracted many researchers as an important component of parking policy [2], [36]–[39]. It guarantees, by a financial constraint for the driver, the respect of the parking dwell time and is also a source of profit for the owner of the parking lot or the authorities. In this way, a consistent pricing policy should be well defined to control travel planning and absorb parking demand.

Dynamic pricing has recently entered the language of parking as a component of an effective congestion mitigation strategy. Most Parking meter offer a "static" linear hourly rate, regardless of traffic conditions and parking usage. Static rates generally favor the search for "free" or "cheap" parking spaces, which increases congestion and travel times. Essentially, dynamic pricing is about setting the correct hourly rates that will determine how long a car is parked - which will increase the turnover and availability of parking.

III. Proposed Solution

The proposed solution aims to overcome reservation and guidance systems limitations. The concept combines real time pricing and reservation sharing service; thus, a driver can consult, online and in real time, parking prices within a radius of their destination, and reserve a spot in the selected parking if the services is available, otherwise, he receives the itinerary to the selected parking lot. In addition, a dynamic pricing control unit which dynamically updates the parking prices, based on real time parking occupancy and reservation requests, is integrated. In addition, the system gives drivers the opportunity of choosing an optimal parking lot based on their destination locations and other parameters (parking availability probability, walking distance, cruising time, … etc.) subject of our previous paper.

A. System Framework

The platform connects all the parking lots in a city forming a network of parking lots as shown in Fig. 4. There are one central price control unit, one central booking center, one parking manager (authority) and multiple parking facilities throughout the city.

Driver wishing to travel to a destination, can use the platform to obtain the price of the available parking lot in optimal way. The driver's request process is as following: parker chooses his destination and set the maximum walking distance he can tolerate, expressed as a search radius value. The system responds back with price offer list of available parking lots within the fixed search radius.

To improve driver experience, the system can calculate the best alternative (optimal parking lot) based on a cost assigned to each parking lot alternative.

Driver chooses the desired parking. He can book a parking place in the select parking, if the reservation service is provided. If not, the parking itinerary is indicated.

Parking authority can adjust the relevant parking price by fixing, increasing or decreasing rates; or updates data center with reservation events. Below we describe the main components of the system:

Price control unit: is an application that run a pricing program. Its function is to control parking resource exploitation by updating prices every predefined time interval. The new prices are sent to user through the web application and virtual message sign (VMS) in each parking facilities.

Booking center: manages the parker's booking requests. In order to validate a booking request, the system check the occupancy of the selected parking. If available, the system assigns a parking place to parker after completing the payment transaction. The parker identifications (i.e., driver and car license IDs) are required to finalize the booking request. Finally, the system notifies the driver of information related to parking spot reserved and itinerary.

Data Center: stores all the data (Sensors, RFID, LPR) from all parkings in the city. It contains a pricing table with the up to date information on pricing per parking lot, occupancy table which holds the utilization data, parking location data, parking lots capacities, … etc. It is also responsible for updating virtual message signs and public devices of pricing information and parking availability.

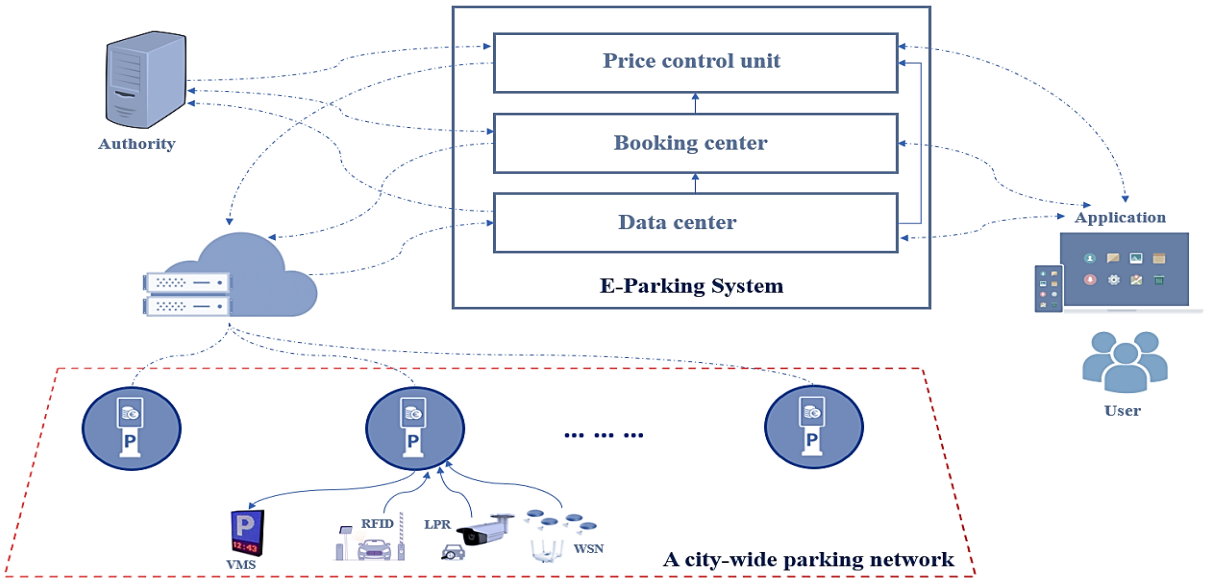


Fig. 4. e-Parking Framework.

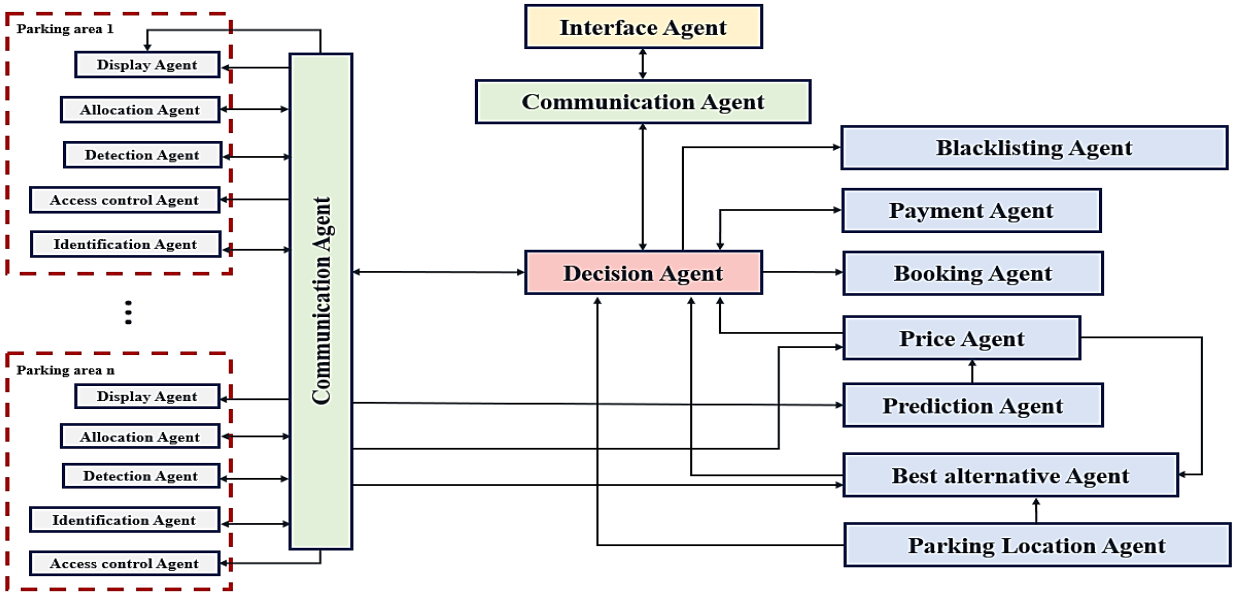


Fig. 5. Multi-Agents Architecture for Managing the Parking Network.

B. Smart Parking Architecture

As part of our work, we were interested in modelling our solution as a distributed multi-agent system. For more details about multi-agent systems, see [40], [41]. For this purpose, a first step, consists of modelling the interactions in a multi-agent system, is illustrated in Fig. 5. Each agent has a mission to perform in this interaction:

- **Display agent**: its role is to display the parking information messages in real time i.e. parking status, prices, etc. The display can be either locally at the parking level or on a city scale via variable message signs and web application.
- **Allocation agent**: As we work in a mixed environment that accepts drivers who have made or not a reservation of parking space, we will need the agent allocation to manage the parking arrival flows. Two missions can be performed by this agent, the first is to assign the available parking places to the reservation requests received; the second is to allocate parking places to drivers who come to the parking without having made a reservation. For the second case, the assignment is based on available places (detection agent) and reservation requests already validated (booking agent).
- **Detection agent**: Its role is to detect the parking availability by checking the status of sensors on each place and the presence of vehicles in the parking entrance and exit.

- **Identification agent**: Its function is to identify the vehicle identity in parking check-point (i.e. parking entrance) via identification technologies i.e. RFID, ALPR or a combination of the two technologies.
- **Access control agent**: It control access/exit to/from the parking. To perform this function, this agent relay on information from detection, identification, booking and allocation agents.
- **Communication agent**: It ensures communication between agents located in different environments.
- **Decision agent**: It collects messages from all agents and makes the final actions and decisions of the system. It represents the authority or parking network manager.
- **Price agent**: Price agent mission is to control parking occupancy and update parking prices every predefined time interval. The new prices are calculated based on the parking availability, prediction information and reservation requests. The role of this agent is to overcome the phenomenon of multiple-car-chasing-single space, by maintaining an optimal level of occupancy via a dynamic pricing policy. When the target occupancy level is achieved, the price agent adjusts (increase or decrease) the prices dynamically. In this way, the parking can always maintain an optimal level of availability able to satisfy its customers.
- **Prediction agent**: It predicts the occupancy level based on real time (detection agent) and historical occupancy (booking agent) data. Its role is to provide price agent by data necessary for price adjustment and maintain an optimal occupancy parking rate.
- **Best alternative agent**: It calculates the best alternative (available parking) based on a cost assigned to each parking alternative. The cost is calculated based on several parameters i.e. walking distance, parking price, parking occupancy, etc.
- **Booking agent**: It manages and stores the validated reservation requests and parking lots information.
- **Payment agent**: It manages the payment transactions for confirmed reservation requests.
- **Parking location agent**: Contains information about parking locations. It displays the available parking lots within the search radius, locates the user and sends the itinerary to the desired parking.
- **Blacklisting agent**: it stores in a black list the customers judged by the decision agent as customers prohibited from accessing the parking platform.
- **Interface agent**: Acts as an intermediary between car parks, users and the system, receives customer requests and displays available processing results.

This architecture guarantees the quality of service by ensuring optimal occupancy of the parking lots.

IV. Implementation

In this section we discuss the implementation of the system in a real-world scenario.

The project consists in developing a web application, shown in Fig. 6, that allows users to consult, online and in real time, parking prices within a radius of their destination.

In order to evaluate the performance of our system under realistic conditions, we have chosen the metropolis of Casablanca, see Fig. 7, as test environment. Casablanca is the largest and most populated city in Morocco. It is also the largest and most important (both economic and demographic) cities in Africa. It is therefore an excellent choice and a perfect place to lead innovative pilot projects and living laboratories. In addition, it shares a context similar to that of many cities in Africa and Europe, which results in effective consequences.

The flow chart in Fig. 8 illustrates and explains the different steps that can be experienced by users:

The application user can consult the different prices of the parking lots located within the radius already defined (Fig. 9 and 10).

Fig. 6. e-Parking Home Page.

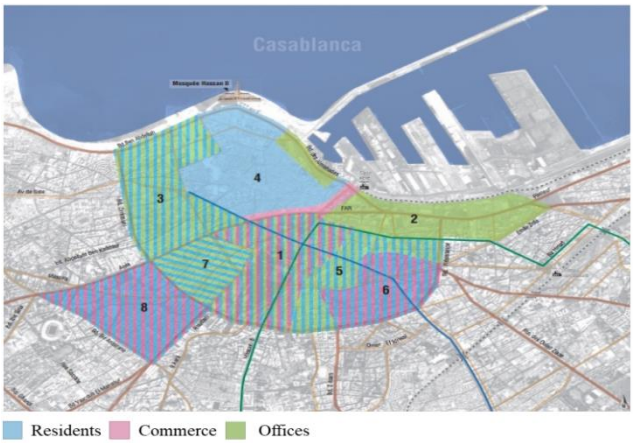


Fig. 7. Anatomy of the Urban Fabric and Activities in Casablanca.

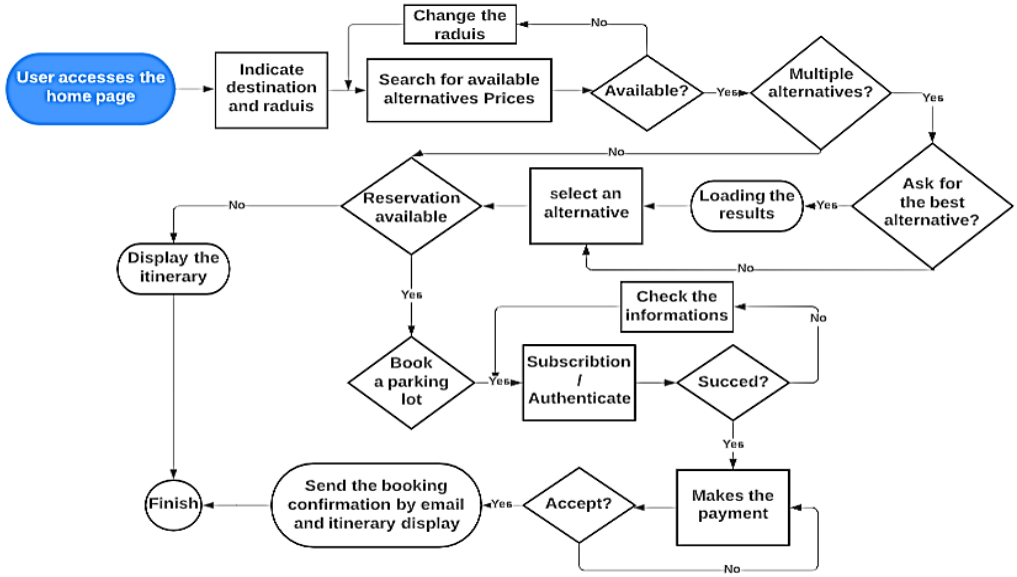


Fig. 8. e-Parking Flow Chart.

Fig. 9. Destination and Search Radius Setting.

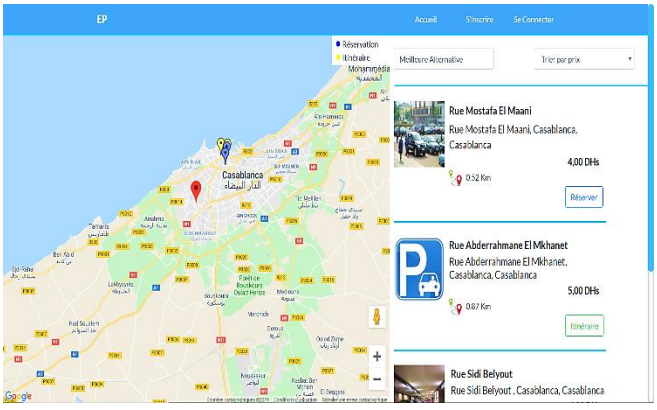


Fig. 10. Search Results.

If several alternatives (parking lots) are available, the driver can request the most optimal alternative in terms of walking distance (calculated from the previously set radius), parking fees (according to its parking time and the price on the database) and other parameters of the available parking (arrival rate, departure rate, real time occupancy rate and parking capacities).

After the most suitable alternative has been calculated, a list is generated that ranks the alternatives according to their total parking costs as shown in Fig. 11. If, among these choices, a parking lot offers a reservation service, an icon is displayed next to it to orient the driver towards the reservation service.

Otherwise the available parking is shown alone with the itinerary icon.

To make a reservation, the user should log in or create an account if he does not been registered. Registration is made by providing the user's information such as first and last name, personal details (address, telephone and email).

An email is sent to our customers who have just registered on our web application, so that they can activate their accounts (Fig. 12).

Interface, in Fig. 13, allows users to book a parking lot by indicating the arrival and departure date and time.

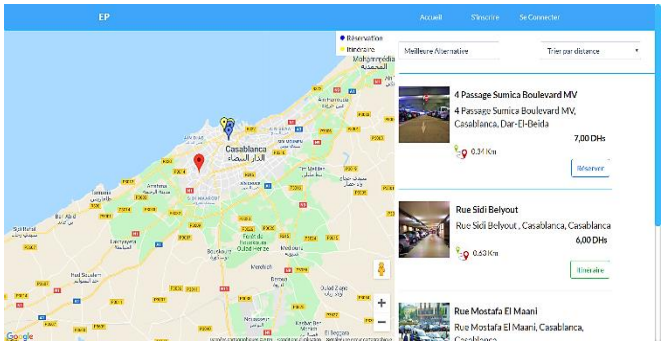


Fig. 11. Best Alternative Results.

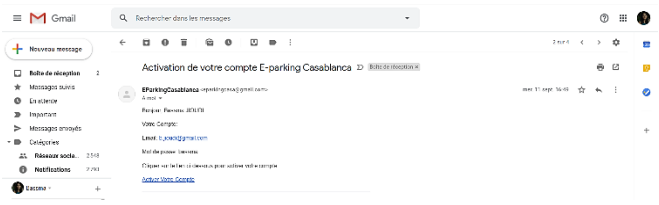


Fig. 12. Activation Email.

Fig. 13. Reservation Interface.

In Fig. 14, the customer has the possibility to pay parking fees online using all possible cards (visa, MasterCard, etc.).

When the payment transaction is completed, the driver receives the itinerary to the selected parking lot and the relative bill (Fig. 15).

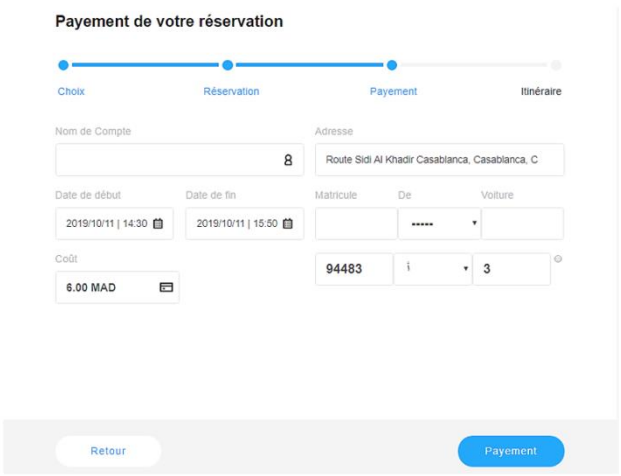


Fig. 14. Payment Interface.

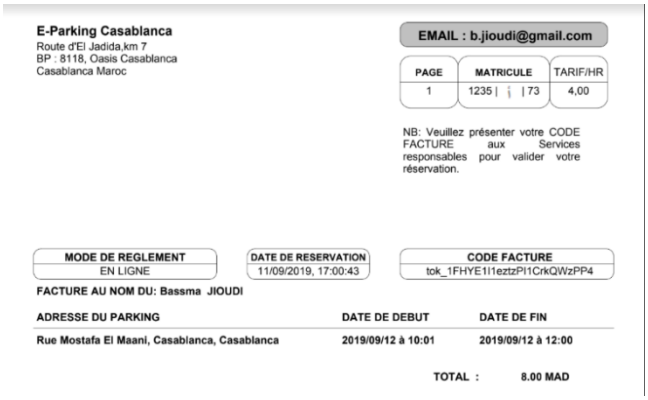


Fig. 15. Confirmation Email.

V. Conclusion

Parking is at the heart of the city, given its spatial influence and its role in urban mobility. It is also a key element in the management of its urban landscape and an essential element for effective mobility. However, parking remains a source of environmental nuisance and collapse of the quality of life. There are many challenges, such as land use, public space occupancy, economic impact, pedestrian safety and social equity. This situation becomes even more alarming in dense and ultra-dense urban environments where the availability of parking spaces significantly affects the flow of traffic. In this article, we present an analysis of smart parking concept evolution and statistical analysis of the published patent applications in this field in the ORBIT database around the world. Then, we introduce a new multi-agent architecture to manage a smart parking. Finally, we present the e-Parking, smart parking Web application developed in the Foundation of Research, Development and Innovation in Sciences and Engineering of Casablanca, Morocco (FRDISI). The proposed platform aims to improve the driver experience by combing real time pricing and reservation sharing service.

Although the main objectives of our project are achieved, the application developed could be enriched by other advanced features and improvements as the integration of multimodal systems to draw driver attention on other means of transportation available in urban environment. We also plan to communicate with parking meters to integrate on-street parking into our applications.

We have started another work with the city of Casablanca, which consists in identifying and determining the behavior of drivers and visiting drivers in the city of Casablanca. This profiling project aims to adapt the solutions proposed in the field of mobility to the reality experienced and to the actors of mobility and all stakeholders at city level.

Acknowledgment

The authors would like to thank the municipality of Casablanca-Morocco, for kindly providing the parking dataset used in this research.

References

[1] F. Galatioto and M. C. Bell, “Simulation of the illegal double parking: quantifying the traffic and pollutant impacts,” IV Int. SIIV Congr., 2007.

[2] D. C. Shoup, “Cruising for parking,” Transp. Policy, vol. 13, no. 6, pp. 479–486, Nov. 2006.

[3] R. E. Barone, T. Giuffrè, S. M. Siniscalchi, M. A. Morgano, G. Tesoriere, "Architecture for parking management in smart cities", IET Intell. Transp. Syst., vol. 8, no. 5, pp. 445-452, 2014.

[4] M. Strasser, D. Mauser, and S. Albayrak, “Mitigating traffic problems by integrating smart parking solutions into an interconnected ecosystem,” in 2016 IEEE Symposium on Computers and Communication (ISCC), 2016, pp. 32–37.

[5] I. G. B. B. Nugraha and F. R. Tanamas, "Off-street parking space allocation and reservation system using event-driven algorithm," 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), Langkawi, 2017, pp. 1-5.

[6] J. Cherian, J. Luo, H. Guo, S.-S. Ho, and R. Wisbrun, “ParkGauge: Gauging the Occupancy of Parking Garages with Crowdsensed Parking Characteristics,” in 2016 17th IEEE International Conference on Mobile Data Management (MDM), 2016, vol. 1, pp. 92–101.

[7] D. Calderone, E. Mazumdar, L. J. Ratliff and S. S. Sastry, "Understanding the impact of parking on urban mobility via routing

games on queue-flow networks," 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, 2016, pp. 7605-7610.

[8] K. C. Mouskos, J. Tvantzis, D. Bernstein, and A. Sansil, “Mathematical formulation of a deterministic parking reservation system (PRS) with fixed costs,” in 2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No.00CH37099), 2000, vol. 2, pp. 648–651 vol.2.

[9] R. C. Larson and K. Sasanuma, “Congestion Pricing: A Parking Queue Model,” Massachusetts Institute of Technology. Engineering Systems Division, Working Paper, Aug. 2007.

[10] O. Cats, C. Zhang, and A. Nissan, “Survey methodology for measuring parking occupancy: Impacts of an on-street parking pricing scheme in an urban center,” Transp. Policy, vol. 47, pp. 55–63, Apr. 2016.

[11] United nations department for economic and social affairs, united nations e-government survey 2018. S.l.: united nations, 2019.

[12] “Centre d’Etudes et de Recherches Démographiques.”. Available: https://www.hcp.ma/Centre-d-Etudes-et-de-Recherches-Demographiques_a735.html.

[13] “Démographie marocaine-Dissertation - 30062 Mots.”. Available: https://www.etudier.com/dissertations/D%C3%A9mographie-Marocaine/48607050.html.

[14] Sandra Breux and Jérémy Diaz. 2017. La ville intelligente : origine, définitions, forces et limites d'une expression polysémique. Institut national de la recherche scientifique-Centre Urbanisation Culture Société.

[15] Maria-Lluïsa Marsal-Llacuna, Mark Evan Segal, The Intelligenter Method (I) for making “smarter” city projects and plans, Cities, Volume 55, 2016, Pages 127-138, ISSN 0264-2751,

[16] Bhagya Nathali Silva, Murad Khan, Kijun Han, Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities, Sustainable Cities and Society, Volume 38, 2018, Pages 697-713, ISSN 2210-6707,

[17] Albino, V., Berardi, U., & Dangelico, R. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. Journal of Urban Technology, 22(1), 3–21. http://dx.doi. org/10.1080/10630732.2014. 942092.

[18] “Communication : Envisager sa politique de stationnement d’un point de vue « smart » ? Quelques initiatives d’utilisation de nouvelles technologies dans la gestion du stationnement (03-2016).”. Available: https://www.uvcw.be/articles/3,10,2,0,6399.htm. [Accessed: 24-Oct-2019].

[19] I. G. B. B. Nugraha and F. R. Tanamas, “Off-street parking space allocation and reservation system using event-driven algorithm,” in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017, pp. 1–5.

[20] M. Liu, J. Naoum-Sawaya, Y. Gu, F. Lecue, and R. Shorten, “A Distributed Markovian Parking Assist System,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 6, pp. 2230–2240, Jun. 2019.

[21] S. Ma, O. Wolfson, B. Xu, "UPDetector: Sensing parking/unparking activities using smartphones", Proc. ACM Int. Workshop Comput. Transp. Sci., pp. 76-85, 2014.

[22] M. Rinne, S. Torm¨ a, and D. Kratinov, “Mobile crowdsensing of parking ¨ space using geofencing and activity recognition,” in 10th ITS European Congress, Helsinki, Finland, 2014, pp. 16–19.

[23] V. Koval, V.Turchenko, V. Kochan, 2003, Smart License Plate Recognition System based on Imaging Processing using Neural Network.

[24] M. Gallo and F. Montone, "The potential effects of real-time parking (RTP) systems on cruising for parking and greenhouse gas emissions," 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Milan, 2017, pp. 1-6.

[25] A. Braga Reis, S. Sargento and O. K. Tonguz, "Smarter Cities With Parked Cars as Roadside Units," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 7, pp. 2338-2352, July 2018.

[26] K. Hassoune, W. Dachry, F. Moutaouakkil and H. Medromi, "Smart parking systems: A survey," 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), Mohammedia, 2016, pp. 1-6.

[27] P. Sotres, C. L. d. l. Torre, L. Sánchez, S. Jeong and J. Kim, "Smart City Services Over a Global Interoperable Internet-of-Things System:The Smart Parking Case," 2018 Global Internet of Things Summit (GIoTS), Bilbao, 2018, pp. 1-6.

[28] I. G. B. B. Nugraha and F. R. Tanamas, "Off-street parking space allocation and reservation system using event-driven algorithm," 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), Langkawi, 2017, pp. 1-5.

[29] Y. Geng and C. G. Cassandras, "New “Smart Parking” System Based on Resource Allocation and Reservations," in IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, pp. 1129-1139, Sept. 2013.

[30] M. Alam, B. Fernandes, J. Almeida, J. Ferreira and J. Fonseca, "Integration of smart parking in distributed ITS architecture," 2016 International Conference on Open Source Systems & Technologies (ICOSST), Lahore, 2016, pp. 84-88.

[31] P. M. d'Orey, J. Azevedo and M. Ferreira, "Exploring the solution space of self-automated parking lots: An empirical evaluation of vehicle control strategies," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, 2016, pp. 1134-1140.

[32] T. Lin, H. Rivano and F. Le Mouël, "A Survey of Smart Parking Solutions," in IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 12, pp. 3229-3253, Dec. 2017.

[33] M. Ş. Kuran, A. Carneiro Viana, L. Iannone, D. Kofman, G. Mermoud and J. P. Vasseur, "A Smart Parking Lot Management System for Scheduling the Recharging of Electric Vehicles," in IEEE Transactions on Smart Grid, vol. 6, no. 6, pp. 2942-2953, Nov. 2015.

[34] Nourinejad, M., and M. J. Roorda. 2017. Impact of hourly parking pricing on travel demand. Transportation Research Part A: Policy and Practice 98 (April):28–45.

[35] “BAnQ numérique.”. Available: http://numerique.banq.qc.ca/.

[36] L. D’Acierno, M. Gallo, and B. Montella, “Optimisation models for the urban parking pricing problem,” Transp. Policy, vol. 13, no. 1, pp. 34–48, Jan. 2006.

[37] Z. (Sean) Qian and R. Rajagopal, “Optimal occupancy-driven parking pricing under demand uncertainties and traveler heterogeneity: A stochastic control approach,” Transp. Res. Part B Methodol., vol. 67, pp. 144–165, Sep. 2014.

[38] A. Millard-Ball, R. R. Weinberger, and R. C. Hampshire, “Is the curb 80% full or 20% empty? Assessing the impacts of San Francisco’s parking pricing experiment,” Transp. Res. Part Policy Pract., vol. 63, pp. 76–92, May 2014.

[39] S. Mathur et al., “ParkNet: Drive-by Sensing of Road-side Parking Statistics,” in Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, New York, NY, USA, 2010, pp. 123–136.

[40] M. Dell’Orco and D. Teodorović, “Multi Agent Systems Approach to Parking Facilities Management,” in Applied Research in Uncertainty Modeling and Analysis, N. O. Attoh-Okine and B. M. Ayyub, Eds. Boston, MA: Springer US, 2005, pp. 321–339.

[41] F. Michel, J. Ferber, A. Drogoul, J. Ferber, and A. Drogoul, “Multi-Agent Systems and Simulation: A Survey from the Agent Community’s Perspective,” Multi-Agent Systems, 08-Oct-2018.

Semantic Micro-Services Model for Vehicle Routing using Ant Colony Optimization

Asmaa ROUDANE[*, 1], Mohamed YOUSSFI[2], Khalifa MANSOURI[3]
Lab. SSDIA, ENSET, University Hassan II Casablanca, Morocco
Bd. HassanII, Mohammedia, 28820 Morocco

***Abstract*—In this paper, we propose a new model of vehicle routing based on micro services using the principle of selection and composition of paths. In this model each micro service is responsible for a road resource, a step in the driver's journey, implementing a specific objective of the road course. The micro services are deployed in a cloud architecture in multiple instances according to a system of increase in load and fault tolerance. Drivers requests are sent to a proxy micro-service, with abstract structures of road courses represented by oriented graphs. The proxy micro-service is responsible for analyzing the request to determine the driver's profile and its context in the journey in order to select the micro-service responsible for providing information on the most appropriate road resource. The method of meta-heuristic optimization of road courses, used in the proposed approach, is based on the ant colony algorithm, or we will describe an adaptation of this optimization method to propose to the driver, at each stage of the journey, the most optimal resource to exploit.**

Keywords*—Road traffic; road traffic management systems; intelligent systems; complex systems; graphs; real-time; optimization methods; micro-services; ant colonies; vehicle routing

I. Introduction

Information Technologies have evolved in a remarkable way the Information Systems (IS) of the companies. In recent years, Web technology, virtualization of architectures and cloud computing have largely contributed to the evolution of the Information Systems dimension and have forced organizations to restructure themselves to adapt to these technological evolutions. The development of geo-localized mobile systems such as smart phones and miniaturized embedded systems such as Raspberry and Ardouino has opened a new area that allows us to interconnect the objects of our universe to scientific applications. Web 3.0 is coming to fruition and the Internet of Things (IOT) has become a reality that Information Systems must take into consideration [1]. In fact, the objects in our environment can be equipped with processing units capable of performing treatments, making decisions while being permanently connected to the Internet and thus providing applications with features that make it possible to exploit the intelligence massively collective. The great mass of data (Big Data) produced by connected objects [2] [3] and mobile applications requires the adoption of a new IS data storage and analysis strategy.

In the field of road traffic, vehicle routing must take advantage of the use of all these new information and communication technologies on the one hand, and all other road traffic management methods on the other hand. Faced with the massification of users in the road network, the automation and optimization of vehicle routing is an operation that is necessary to minimize the travel time of drivers to keep a maximum flow of traffic and avoid the phenomenon of congestion.

In front of an increasingly open environment, the selection and composition of road resources requires a more efficient strategy for better reuse of traffic data. Micro-services [3], [4], [5] are an architectural style that has emerged in recent years and are best prepared for massively distributed systems. The main idea is to break down a large application into very small, weakly coupled and independent units called "micro-service". Each micro-service must be responsible for a single feature, developed and deployed independently of each other. Based on a lightweight communication mechanism, this style seems to be an opportunity for road traffic management systems where communication between stakeholders is complex.

To help the user to cross the most optimal path to his destination we will exploit the optimization methods. Combinatorial optimization can involve several domains, always with the aim of improving the functioning by means of a sophisticated management of the resources taking into account the temporal and dynamic characteristics of the system [6][7]. In our case we will describe an adaptation of meta-heuristic optimization method by ant colonies to propose to the driver, at each stage of the journey, the most optimal resource to exploit.

The meta-heuristics of ant colony optimization (ACO) was initiated mainly by Marco Dorigo in 1992 [8], for the search of the shortest path in a graph. It was inspired by the behavior of real ants made by Deneubourg and al in 1983 [9]. The first ant colony optimization algorithm that was developed to solve the problem of the commercial traveler is: Ant System (AS) and was introduced by Dorigo and Gambardella in 1997 [10]. The goal was to optimize a route linking a set of cities. Subsequently, several applications of ACO were presented by researchers, including the problem of routing vehicles in 1999 by Bullnheimer and al [11].

II. Presentation of the Model Architecture

Consider a path represented by an oriented graph $G(V_i, E_j)$ where the nodes V_i represent the steps to reach (i = 0 to N-1, with N represents the total number of nodes) and E_j represent the arcs of the graph defining the different possible paths that the driver can take to reach his destination (j = 0 to M-1, where M represents the number of arcs of the graph). For each

*Corresponding Author

node i, we associate several possible road resource implementations R(i, k) (k = 0 to H-1, where H represents the number of possible resources associated with a node). A road resource can be either a normal step, a rest area, a service station or a control point. During his journey, the driver should explore a road resource for each node according to his personal choices or according to a recommendation system.

Over time, according to the feedback of the drivers and the results obtained by the users in the various control points of the course, each resource R(i, k) of each node i is assigned a score S_i and each arcs E_j a weight W_j reflecting the relevance of the path. The system should also be able to determine, over time, the optimal route that provides the best route for drivers. Fig. 1 shows an example of a path composed of 8 nodes and 9 arcs.

To reach a destination V(i), we define several implementations of road resources R(i, k) possible. As said previously, a road resource can be a normal step in the route, a rest area, a service station, a control point, or an aggregate of a set of resources. In this model, we propose to expose each of the road resource implementations using a micro-service MS-R(i, k) deployed in a cloud architecture. To handle the problem of increase in load and fault tolerance, multiple instances of the same micro-service MS-R(i, k) can be deployed thus defining a massively distributed architecture.

To facilitate the orchestration process, all micro-services should register and publish the description of the exposed resource in a special micro-service named MS-D representing the semantic system directory. This description includes both the functional and non-functional properties of micro-services. This is an important element in the selection and composition of services. We also propose to define a micro-service specially designed to centralize and expose the configuration of all micro-services of the platform.

In this model, all driver requests are sent to a proxy micro-service named MS-PROXY, specifically designed to handle the orchestration process by invoking the appropriate micro-service to serve the requested resource. For this, the proxy service consults the registration and discovery service to find the location of the most appropriate micro-service and will be asked to respond to the contextual request. In the case where the proxy service needs to communicate the same information to several micro-services, it should use a more efficient messaging system based on the AMQP protocol by using one of the broker implementing this protocol as Rabbit MQ. The schema of Fig. 2 shows a technical architecture showing all the micro services offered in this model inspired from a related work [12].

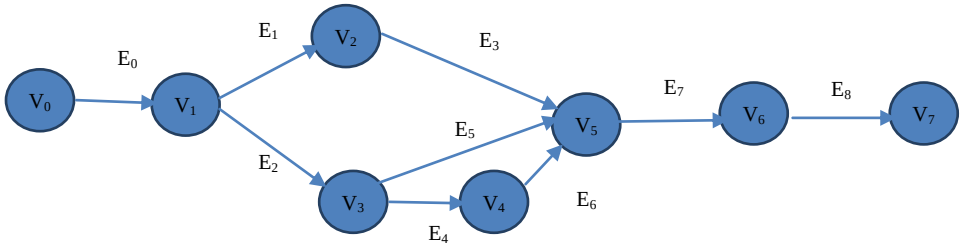


Fig. 1. Representation of a Path by an Oriented Graph.

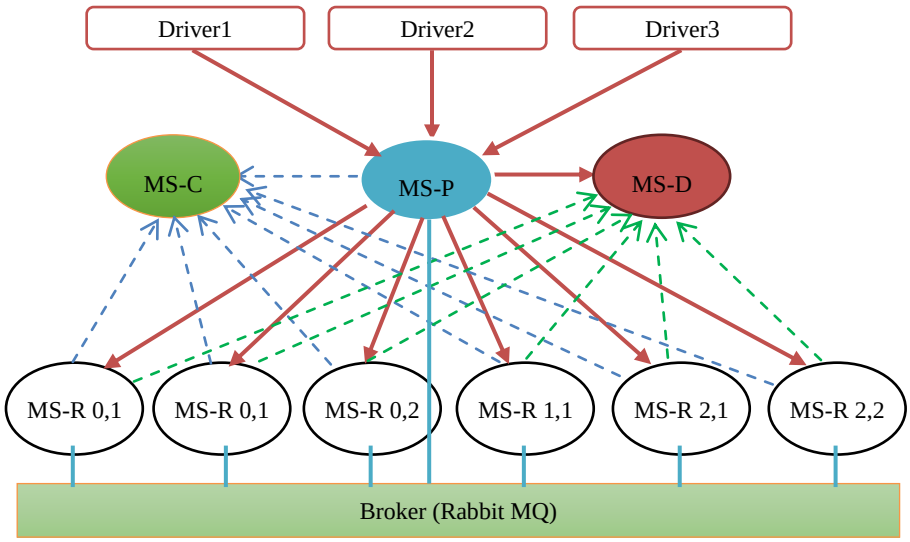


Fig. 2. Micro-Services Architecture of the Model.

III. Road Resource Model

A road resource is a structure that contains the contents of a fragment of the route. This could be a normal step in the trip, a rest area, a service station, or a control point. Generally, a road resource is an aggregate of an element of the route, service station and control point. The level of granularity of the resource depends on the context of the journey and has a considerable influence on the size and efficiency of the system. A finer level of granularity makes it possible to follow the driver more closely, thus enabling the system to effectively evaluate the drivers' journeys and to target their problems with great precision and also to facilitate their orientation. However the number of micro-services that will be deployed in the system would be too important. Hence, the need to deploy the system in a cloud environment [13]. On the other hand, a large level of granularity of road resources reduces the size of the platform, but also reduces the level of efficiency of the system. Indeed, with large basic road resources, it is more difficult to follow the driver more closely, which makes the control of the course more difficult.

To include in our model the different possibilities, we propose to use the composite pattern to structure an elementary road resource (Fig. 3). Each road resource is described by a set of functional or non-functional properties. The functional properties make it possible to describe the semantics of the resource, its coordinates, its type, its level of recommendation, etc. Non-functional properties are related to the technical aspects of the resource such as safety, cost, weight, recommended flow etc.

IV. Description of Micro-Services in the Architecture

A. Micro-Service of Road Resources

This micro-service is the central element of the model, because it's it, in collaboration with its counterparts, which exposes the state of the road network to the driver. At the moment of its start, this micro-service first connects to the configuration service to search its configuration and then to the registration service to record its location and publish its description in the dedicated directory for.

Technically, it has its own layers of data, business, and web:

- The data access layer allows the structuring and persistence of its basic content in light storage systems such as micro-DBMS or simple XML files.
- The business layer defines for the micro service the necessary processing to perform to manage the content of the supported resource. Generally, the processes of this layer evolve within the transactions which make it possible to guarantee the coherence of the data recorded during the interactions with the drivers.
- The micro service web layer exposes all business layer processes via a REST API or other remote access components such as a SOAP-based Web service. Access to this layer is restricted mainly to the Proxy micro-service that manages access to drivers in a more efficient manner.

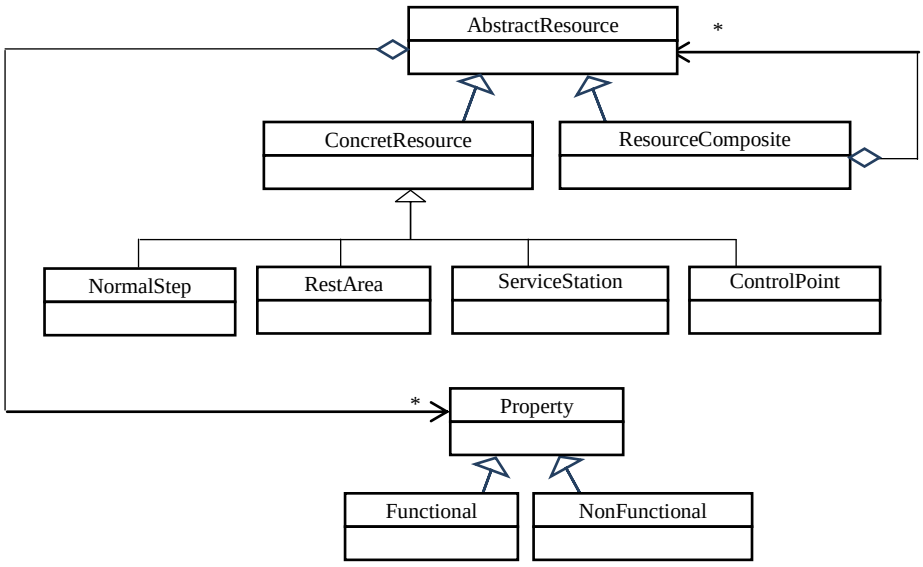


Fig. 3. Structure of a Road Resource.

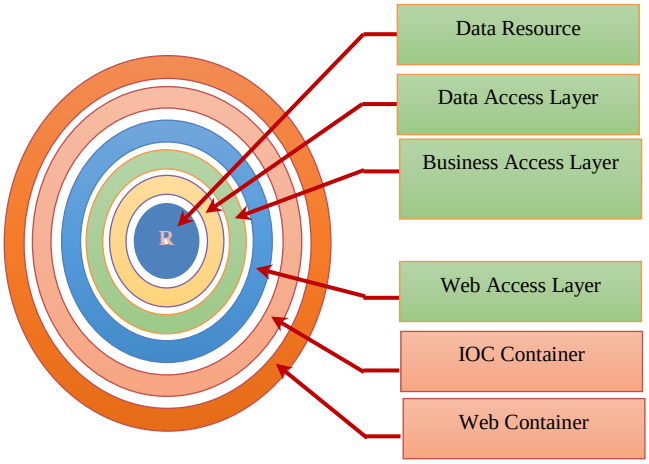


Fig. 4. Structure of a Road Resource Micro-Service.

The micro-service evolves in a container which ensures the principle of the inversion of control based on a micro-Framework IOC. The latter makes it possible to deal implicitly and to easily inject the technical aspects into the micro-service. This is the case for security aspects, transaction management, persistence management, logging, monitoring, hot configuration, cloud services, migration, and so on. To expose the micro-service features via http and SOAP protocols, a web container is also used in the architecture. Fig. 4 shows the different software layers of this micro-service inspired from a related word [14].

B. *Proxy Micro-Service*

As noted in the previous section, all driver requests are sent to the Proxy micro-service. The latter finds in the request the identifier of the route plan and that of the requested road resource. The proxy service requests the registration service to provide a list of deployed micro services that provide the content (status) of the road resource in question. Using a load balancing algorithm combined with the scoring system relating to the relevance of its resources, the proxy designates the appropriate micro-service and redirects it to retrieve the resource to be sent to the driver.

Following the feedback from the driver, the score of this resource is updated. This score will be used to calculate the arc weight linking this resource and that of the previous node in the path of the driver's journey. This step is part of an algorithm for optimizing pathways by meta-heuristics using the principle of ant colonies. This optimization system, which will be presented in the following section, provides the driver with the best path among those defined in the course plan. The communication diagram of Fig. 5 shows a diagram of interactions between the driver, the micro-service proxy, the Discovery micro-service, two instances of the micro-service exhibiting the same resource implementation relative to the contextual objective and another instance of the micro-service exposing another implementation of the resource of the same contextual objective.

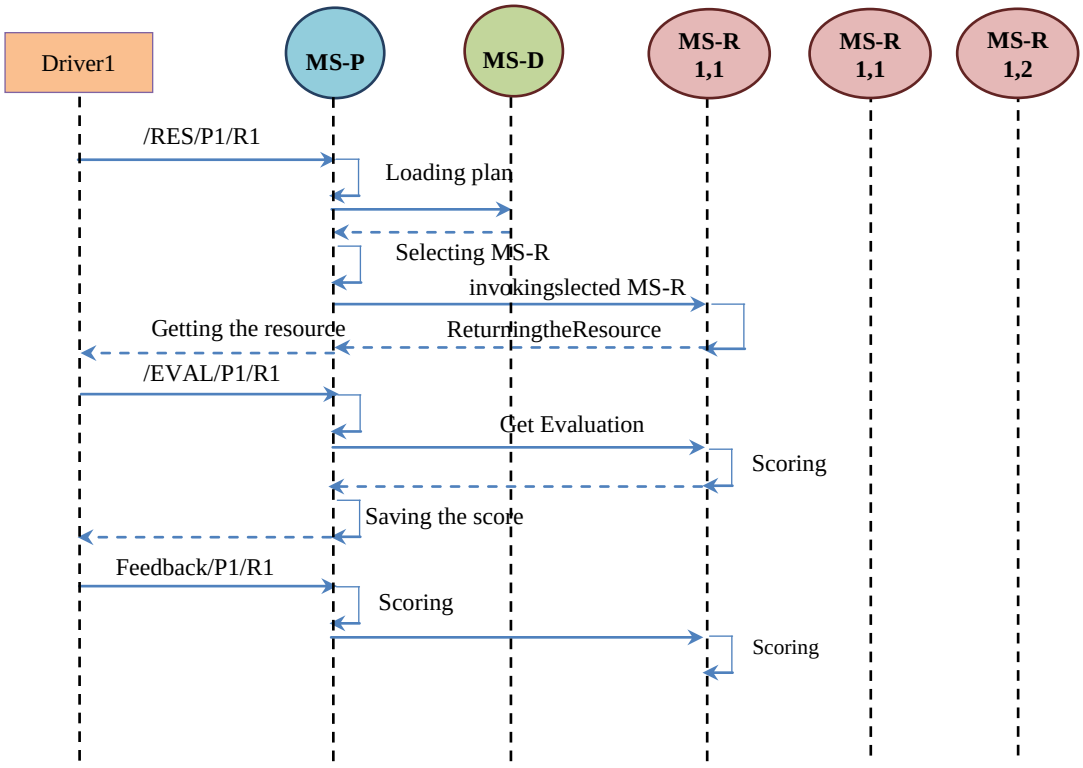


Fig. 5. Interaction Diagram Showing the Functioning of Micro-Services of the Model.

V. Algorithm for Optimizing Road Course by Ant Colonies

Ant colony algorithms are algorithms inspired by the behavior of ants, or other species forming a super organism, and which constitute a family of meta-heuristics of optimization. Initially proposed by Marco Dorigo et al. [15], in the 1990s for the search of optimal paths in a graph, the first algorithm is inspired by the behavior of ants looking for a path between their colony and a food source. The original idea has since diversified to solve a broader class of problems and several algorithms have emerged, inspired by various aspects of ant behavior. The idea comes from observing the exploitation of food resources in ants. Indeed, these, although having individually limited cognitive abilities, are able by exploiting the collective intelligence, to find the shortest way between a food source and their nest.

A model explaining this behavior is as follows:

- An ant (called "scout") runs more or less randomly around the colony.
- If it discovers a food source, it returns more or less directly to the nest, leaving on its way a trail of pheromones.
- These pheromones are attractive, ants passing nearby will tend to follow, more or less directly, this track.
- Returning to the nest, these same ants will strengthen the track.
- If two tracks are possible to reach the same food source, the one that is the shortest will be, at the same time, traveled by more ants than the long track.
- The short track will be more and more reinforced, and therefore more and more attractive.
- The long track, it will eventually disappear, pheromones are volatile by evaporation phenomenon.
- In the long term, all the ants thus determined and "choose" the shortest track.

In our case, we consider our plane of paths represented by an oriented graph of Fig. 1. Each node represents an element of the road network and each arc a possible route between two elements. Each of these arcs has its weight W initialized by the team of managers of the road traffic management system, to mark a relative importance between the relevance of several possible paths from the same origin. Each driver who travels the graph is represented by an "ant", a micro-agent, who navigates on the underlying graph, by taking a path Ej to arrive at a node V_E proposing a control of the acquired, the ant releases pheromones C whose quantity depends on the note α obtained at the end of this control.

To reflect the phenomenon of retro-propagation, pheromones are not simply released on the arc that led the ant to the current node but on the last n last arcs that the ant followed. This is done to reflect the fact that to reach its destination all the characteristics of the roads that lead to it count. Of course, this influence diminishes with time and space: the further the node is removed in the history of the path P(a) of the ant, the less it is important. For this to be taken into account, the amount of pheromone deposited decreases as retro-propagation advances. In our implementation of this algorithm, we propose that the amount of pheromone that would be added in the path E_k in the history P(a) is the quantity α divided by a number d(k) which represents the distance separating the node V_E subject destination, at the origin node of the arc E_k.

$$C_k = C_k + \frac{\alpha}{d_k} \quad (1)$$

Fig. 6 shows an example of retro-programming of pheromones for the case of a driver who has passed through a control at node 7 and obtains a note α_1. To arrive at this node, he followed the path P={E_0,E_1,E_3,E_7}. The amounts of pheromones deposited in arcs E_7, E_3, E_1 and E_0 of this course would be respectively α_1, $\alpha_1/2$, $\alpha_1/3$ et $\alpha_1/4$.

The fact that biological pheromones evaporate over time is extremely important because it allows the ant colony to rely on constantly updated information. In our artificial system, it is important to implement a form of evaporation to avoid that the system remains "stuck" in a local optimum and to open the door to the expected characteristics of dynamic adaptability:

$$C_t = \tau^x C_{t-1} \quad (2)$$

Equation 2 gives the form of evaporation for pheromones. τ, the evaporation rate, is a key setting parameter of the system. The period of evaporation, x which says at which intervals the evaporation is calculated is a system constant. Typically, τ=0.999 and x=1 day.

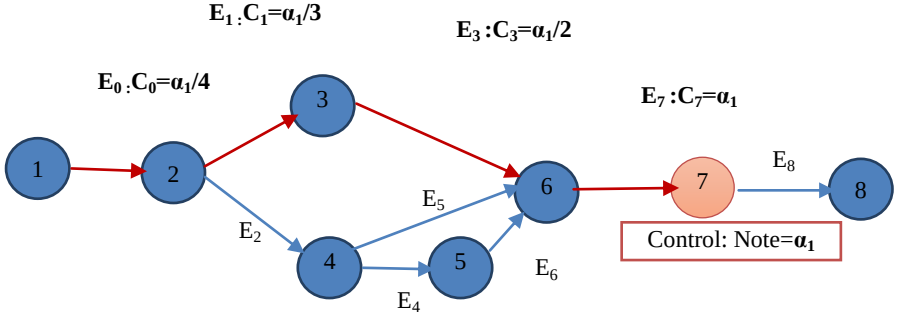


Fig. 6. Principle of Retro-Propagation of Pheromones.

In addition to the results of the controls carried out by the drivers, the road resources will be subjected to an appreciation of the users. Each road resource is therefore assigned a rating β from each driver. This score will be multiplied by a credibility factor μ associated with each user by the system and the management team. As in the previous case of the note α assigned by the system to the drivers, the note relating to the appreciation of the drivers will be translated in quantity of pheromones A placed on all the arcs which led the users towards this resource by respecting the principle of retro - propagation of equation 1:

$$A_k = A_k + \frac{\mu.\beta}{d_k} \quad (3)$$

The phenomenon of evaporation of pheromones A is established in the same way as equation 2:

$$A_t = \tau^x A_{t-1} \quad (4)$$

The values W, C and A are so-called collective factors, they concern all drivers who will use the information they contain to build their path. W is exogenous, it is brought by the team of managers to the community drivers/ants. C and A are endogenous, they are created by individuals for the service of the community: immersed intelligence, emerging culture. The system must also be able to take into account individual factors, specific to each driver, in order to achieve the compromise that is sought between the individual, the community and the environment.

The individual factors that can be taken into account are numerous (preferences, excellence, history, etc.). The idea is to allow the system to propose solutions according to the individual himself. In our case, we will consider the historical factor H.

The History Factor H carries the information on the previously visited nodes in the history of the driver (vehicle). Since this is an individual factor, there is an H value per driver per node. The default value of H is chosen to 1. When a node is visited by a driver, the corresponding value H is multiplied by a factor h inversely proportional to the score obtained. For example we can choose the following values:

h = 0.25 if the score is between 75% and 100%,

h = 0.5, if the score is between 50% and 75%,

h = 0.75 if the score is between 25% and 50%

h = 0.95 if the score is less than 25%.

The role of H is to decrease the probability that the node already visited is proposed again. Of course, if the node was the seat of a very bad score (h = 0.95) it will be more quickly proposed again than if there is a very good score (h = 0.25).

Over time, of course, drivers forget what they saw. This is why H tends to return naturally to 1. This "anti-evaporation" is described by the equation that models this phenomenon of anti-evaporation of memory pheromones.

$$H_t = H_{t-1}\left(1 + \frac{1-H_{t-1}}{H_{t-1}}\frac{1-e^{\gamma x}}{1+e^{\gamma x}}\right) \quad (5)$$

γ is a time constant that regulates the speed of the phenomenon. It must be adjusted to match the volatility of the drivers' memory. It is therefore a constant that must be the subject of a dialogue with the managers. Equations 6 and 7 allow an easy calibration of γ. Equations 4 and 5 show how to calculate the value of γ as a function of the values of H at times t and t-1 and the duration x during which a driver could possibly forget a resource already explored before.

$$\gamma = \frac{1}{x}\ln\left(\frac{1+\tau}{1-\tau}\right) \quad (6)$$

$$\tau = \frac{H_t - H_{t-1}}{1-H_{t-1}} \quad (7)$$

All the factors previously described (W, C, A and H) are unified by a function called "fitness", by analogy with the literature of genetic algorithms. Equation 8 shows the value expression of the "fitness" of the arc E_{ij} leading from the node i to the node j for the individual a. H is the memory factor, W is the weight proposed by the system, C is the control pheromone concentration and A is the pheromone concentration of the driver's feedback.The $\omega_H, \omega_w, \omega_c$ and ω_A are parameters that make it possible to give more or less importance to one factor compared to others. This function measures the excellence of a given arc, its "desirability", and will condition, through a selection procedure, the probability that the arc considered will be suggested to drivers.

$$f(a, E_{i,j}) = \omega_H H(j,a)(\omega_w W + \omega_c C + \omega_A A) \quad (8)$$

We can say that an arc is desirable when:

- He is encouraged by the team of road traffic managers (W high).
- He sees good ratings attributed by the system to drivers in control points (high C).
- He sees good appreciation ratings attributed by drivers to road resources (high A).
- The node at which it ends has never been visited by the driver or was forgotten by him (H close to 1).

When a driver validates a node, it is advisable to choose, among the arcs that come out of this node, the one that is most suitable to borrow, this is the role of the selection procedure. This is where the fitness measures are used. A selection procedure will choose an arc randomly but all the more so if the fitness of the arc is high. We will see the "effective" arcs become predominant but not exclusively, there will be room for chance and exploration. This part of chance is a crucial feature of the selection procedure, it is called selective pressure or s. The greater is s, the more the draw is guided by the fitness function and the stronger arcs will tend to dominate the weak ones.

VI. Application and Results

In this part, we propose the results of a simulation corresponding to an application implemented at the level of the proxy micro-service, which consists in randomly launching HTTP clients, each representing a driver. Each driver takes his route from one node to another in the route plan graph. We have chosen an example of a part of the course whose structure represented by Fig. 7.

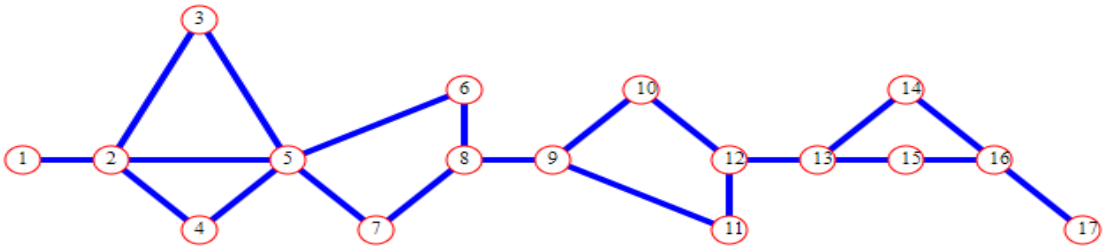


Fig. 7. Course Structure Proposed in this Application.

This course plan is an oriented graph that consists of the following nodes:

- Nodes 1, 2, 5, 9 and 13 represent normal steps in the path.
- Nodes 3, 6, 10 and 14 represent service stations.
- Nodes 4, 7, 11 and 15 represent rest areas.
- Nodes 8, 12, 16 and 17 represent control points.

In this proposal, the weight of a path proposed by managers is simply the distance between two nodes. When there are several links leaving a node, we have chosen the values of the relevance proposed by the managers: W, allowing more recommending the passage by paths which have the shortest distances. This is the case of the links (2 => 3), (2 => 4) and (2 => 5) which respectively have the values of W: 111.80, 70.71 and 100. Which means that the team managers recommend to borrow, in order of preference, a rest area (4) or a normal stage (5) or a service station (3).

When passing through a control point, a random note C is generated, representing the amount of pheromones left on its path (equation 1). Otherwise the driver leaves an appreciation of the selected road resource representing the amount of pheromones A left on its course (equation 4) by activating the evaporation procedure before proceeding to the next node.

Fig. 8, 9 and 10, respectively show the representation of the evolution curves of the parameters W, C, A and Fitness as the drivers follow their journeys.

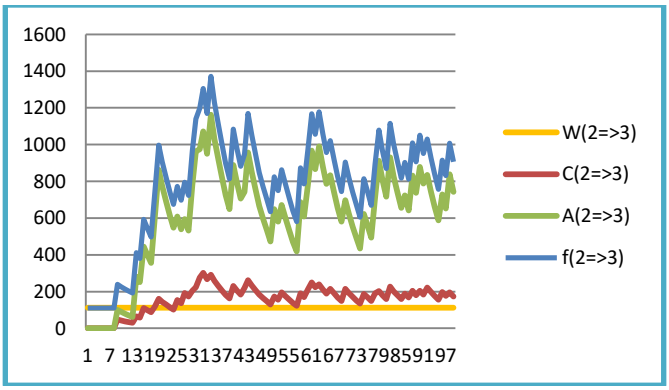


Fig. 8. Evolution of W, C, A and Fitness Values for the Link (2 => 3).

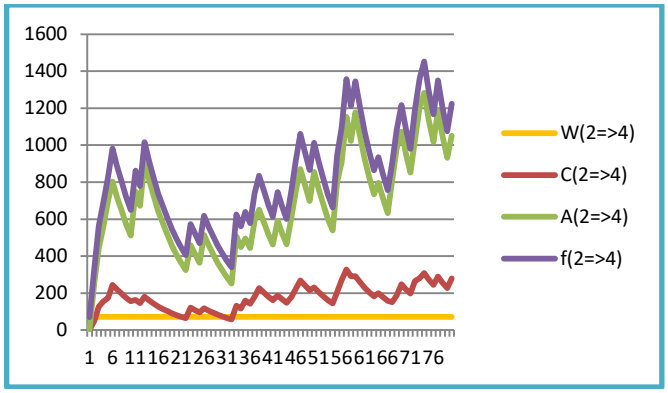


Fig. 9. Evolution of W, C, A and Fitness Values for the Link (2 => 4).

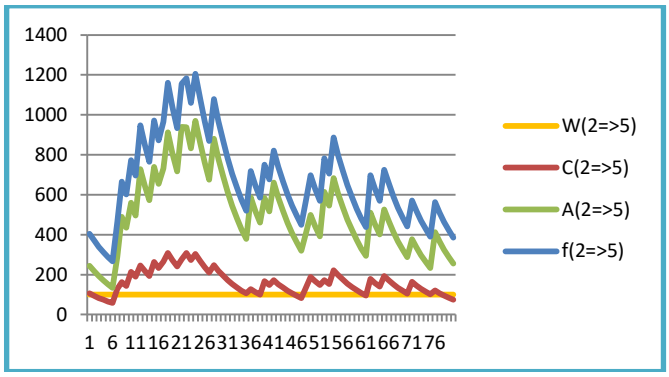


Fig. 10. Evolution of W, C, A and Fitness Values for the Link (2 => 5).

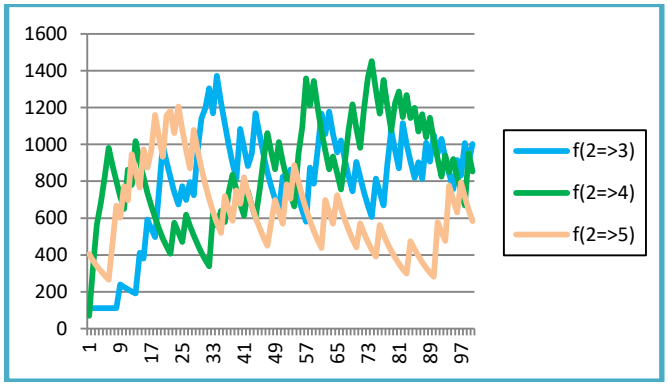


Fig. 11. Fitness Comparisons of the Three Links (2 => 3), (2 => 4) and (2 => 5).

Fig. 11 represents a comparison between the fitness of the 3 links (2 => 3), (2 => 4) and (2 => 5). We note that most of the time that the link (2 => 4) is more borrowed, reinforcing the recommendation of the team of managers by taking the path that has the shortest distance. As in the system, there is a lot of randomness, dominance can change direction. This is the case of the driver N° 34 which makes that the link (2 => 3) becomes momentarily more dominant. This is still the case of the driver 24 who leaves the fitness function of the arc (2 => 5) more dominant instantly when it was not recommended by the team of managers. This reveals other parameters that are difficult to take into consideration, such as random excellence, driver serenity and the relevance of assessments in control points. Despite all this, the model makes it possible to obtain, in the end, interesting results thus resisting the random phenomena mentioned.

VII. Conclusion

The model proposed in this contribution revolves around optimizing agents aiming to improve the vehicles routing. The approach is based on the ant colony algorithm to improve road traffic management by helping to select the best possible path for a given destination. Ant behavior is characterized by its autonomy, distributed operation and self-organizing capabilities to form systems capable of performing very complex tasks. For this reason, our model is inspired by heuristic research that has involved many criteria in the objective function on a distributed multi-agent architecture in which artificial agents are created by analogy with social insects. The proposed model can provide a new way to improve traffic conditions and optimize the use of resources to provide the driver with the best guidance to their destination. In future work we plan to propose and implement other optimization algorithms based on other parameters of the traffic environment including safety and cost of the road (not only the distance of the journey), and this to be able to select the most optimal path to a destination by adding other factors according to the preferences of the user. We also aim to combine the ACO (ant colony) algorithm and the Q-Learning reinforcement learning algorithm to improve the modeling of the optimizers agents of the vehicle routing system.

References

[1] H. Rolf, Weber, "Internet of Things - Need for a New Legal Environment?", Computer Law & Security Review, pp. 522-527, 2009.

[2] Yunchuan Sun, Houbing Song , Antonio J. Jara, Rongfang Bie, Internet of Things and Big Data Analytics for Smart and Connected Communities, IEEE Access (Vol: 4), pp. 766 - 773, February 2016.

[3] Ejaz Ahmed et al, The role of big data analytics in Internet of Things, Computer Networks Volume 129, Part 2, pp. 459-471, December 2017.

[4] D. Namiot, M. Sneps-Sneppe, "On micro-services architecture", International Journal of Open Information Technologies, vol. 2, no. 9, 2014.

[5] D. Malavalli, S. Sathappan, "Scalable microservice based architecture for enabling dmtf profiles", Network and Service Management (CNSM) 2015 11th International Conference on, pp. 428-432, Nov 2015.

[6] D. Jaramillo, D. V. Nguyen, R. Smart, "Leveraging microservices architecture by using docker technology", SoutheastCon 2016, pp. 1-5, March 2016.

[7] C.H. PAPADIMITRIOU, K. STEIGLITZ, Combinatorial optimization - algorithms and complexity. Prentice Hall, 1982.

[8] C.C. RIBEIRO, N. MACULAN (Eds.), Applications of combinatorial optimization. Annals of Operations Research 50, 1994.

[9] Dorigo, M. (1992). Optimization, learning and natural algorithms. PhD Thesis, Politecnico di Milano.

[10] Deneubourg, J. L., Pasteels, J. M., & Verhaeghe, J. C. (1983). Probabilistic behaviour in ants: a strategy of errors?. Journal of Theoretical biology, 105(2), 259--271.

[11] Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions on evolutionary computation, 1(1), 53--66.

[12] Bullnheimer, B., Hartl, R. F., & Strauss, C. (1999). An improved ant System algorithm for the vehicle Routing Problem. Annals of operations research, 89, 319--328.

[13] Ahmed ESSAYAH, Mohamed Youssfi, Omar Bouattane, Khalifa Mansouri and Elhocein Illoussamen, “QoS-based Semantic Micro Services Discovery and Composition using ACO Algorithm” International Journal of Advanced Computer Science and Applications(IJACSA), 10(6), 2019.

[14] FAREH, Mohamed El-kabir (2015) Une approche basée agents pour l'allocation des ressources dans le Cloud Computing. Masters thesis, Université Mohamed Khider - Biskra.

[15] A. Colorni, M. Dorigo et V. Maniezzo, Distributed Optimization by Ant Colonies, actes de la première conférence européenne sur la vie artificielle, Paris, France, Elsevier Publishing, 134-142, 1991.

Monitoring of Rainfall Level Ombrometer Observatory (Obs) Type using Android Sharp GP2Y0A41SK0F Sensor

Anton Yudhana[1], Yunita Dwi Andriliana[2], Son Ali Akbar[3], Sunardi[4]
Dept. of Electrical Engineering, Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Subhas Mukhopadhyay[5]
Dept. of Mechanical / Electronic Engineering
Macquaire University, NSW, Australia

Ismail Rakip Karas[6]
Dept. of Computer Engineering
Karabük University, Karabük, Turkey

***Abstract*—Measurements of rainfall carried out are generally automatic, but how many parties have carried out research using automatic rain gauge instruments. The rainfall level data obtained can be used to detect flooding so that it reduces the occurrence of natural disasters earlier. The principle works are when rain falls the water is collected and the height will be detected. The Sharp GP2Y0A41SK0F sensor will be read on Arduino Uno and pass through the signal conditioning circuit and then forwarded to the Water Pump to pump water so that the water comes out of the tube and the data will be stored. ESP8266, the Wi-Fi module will send data to Android. Measurements are made when the water is full the value of max 53.2 means heavy rain. After the data is obtained, then look for the standard error of the measurement tool for 7 days. The design of the model builds a tool using a funnel with a diameter of 14 cm and a height of 26 cm. Based on the calculation of the design that has been made the measuring instrument developed is able to measure rainfall up to 26000 mm rain height.**

Keywords*—Arduino; infrared sensor; android; observatory; rainfall

I. INTRODUCTION

Rain in Indonesia has an average rainfall as an illustration of normal rainfall conditions using historical gsmap data [1]. The resources used to calculate rainfall are rain gauge, land-based radar, and remote sensing satellites. The purpose of this study is to spread the aerial photography system that has the ability to capture panoramic images of certain areas. The season that occurs in Indonesia is based on the frequent or infrequent rainfall so that it is known as the rainy season and the dry season [2], [3]. The type of rain in the Central Java region is monsoonal, a characteristic pattern of monsoon types of rainfall that is unimodal (peak rainy season) [4]. For example, 1 mm of rainfall means that rainwater has a height of 1 mm which is accommodated in a flat place with an area of 1 m^2 [5].

An ordinary manual rain gauge used in Indonesia is a type an ordinary manual rain gauge used in Indonesia is a type Observatorium or Ombrometer. Manual rain gauge (Manual Rain Gauge) consists of two types, viz. MRG with a surface area of 100 cm2 and 200 cm2 [6]. We have reported an effective implementation for the Internet of Things used for monitoring regular domestic conditions by means of low-cost ubiquitous sensing system [7]. In the android step monitoring application to connect to the hardware is connected with Bluetooth HC-06. After being able to connect with Bluetooth, the monitoring results the measurement value will be displayed on Android [8].

The working principle of the Sharp GP2Y0A41SK0F sensor is that the sensor will emit infrared light with a frequency of 40kHz and the reflection of the sensor will be received by the infrared detector array [9]. Android (operating system) is an operating system based on Linux for cellular phones such as smartphones and tablet computers[10]. use rain sensors for water detection and light sensors for sunny or cloudy weather detectors that will send input to Arduino uno to provide buzzer output, and short message service (SMS) notifications to turn on relays to control DC motors to open and close the roof [11]. The results of the design performance of the tool can monitor the height of rainfall which is then analyzed in the form of a database and can store data 24 hours automatically to be used as a historical and graphical data [12], [13].

The results of the characterization of the sensor that shows the flow of water with the effect of the hall with the response in the form of voltage, but the characterization of response time and stability is still less responsive and less sensitive. IC chip that has the function of calculating the time starting from seconds, minutes, hours, days, dates, months, to years accurately [14]. All devices measuring the water level can work well according to the sensors used. The results of testing the data displayed on this gauge are as expected [15]. All devices measuring the water level can work well according to the sensors used. The results of testing the data displayed on this gauge are as expected [16], [17]. Use of ESP8266 Module as a WIFI module, widely used by Internet-based applications of Things because the price is cheap, thus reducing many costs and have a pretty good speed of 80 MHz [18].

II. Research Method

In this research, the method explained the results of the study and simultaneously provided an in-depth discussion. Results can be displayed with Images and others that make the reader easy to understand. Discussion can be made in several.

A. System Planning

To realize the design of rainfall measuring devices carried out in several steps to find out the stages of making the tool to completion. This flow chart shows the making of the tool consists of two stages, namely the stage of making the hardware and software design stage. To start the Arduino program (for making the program will be in accordance with what we want) we use Arduino IDE (Integrated Development Environment) [19].

In making this tool, two stages consisting of hardware and software that is in order to produce the appropriate measuring instrument. After learning and understanding the work system of the tools to be made, then make a system design by making hardware or hardware from the tool. Then testing the resulting hardware can work well or not. The hardware has been realized and is functioning properly, then making software design. The design of this software will help use hardware (hardware). If both of them have succeeded and are functioning properly, then the whole tool is tested.

Based on Fig. 1, flow chart of the system design of rainfall level detection above explains the workings of the rainfall level monitoring system. When Sharp GP2Y0A41SK0F does not detect the water level, the relay will be active at 07:00 in the morning and do the handling and sending of data on ThinkSpeak. If the water is full with a height greater than 5 cm while at less than 07:00 then it will make arrangements and sending data on ThinkSpeak. The graphic display on ThinkSpeak will run according to data obtained from the sensor.

B. Hardware Designing

The hardware in this Android-based rainfall level monitoring tool consists of a control system and data acquisition. The control system scheme in the tool is shown in the following Fig. 2, block diagram.

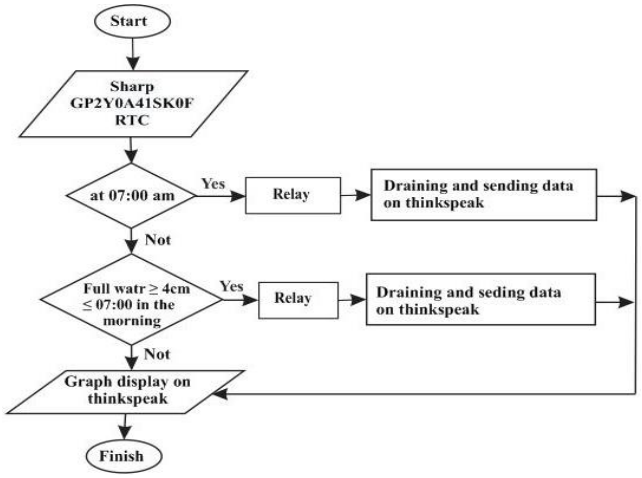


Fig. 1. System Design Flowchart.

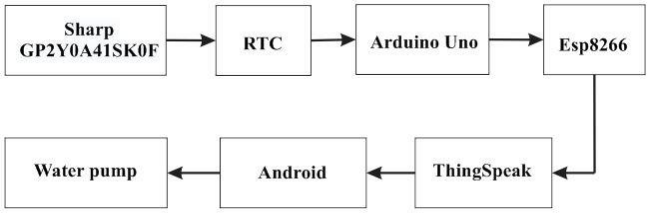


Fig. 2. Block System Design Diagram.

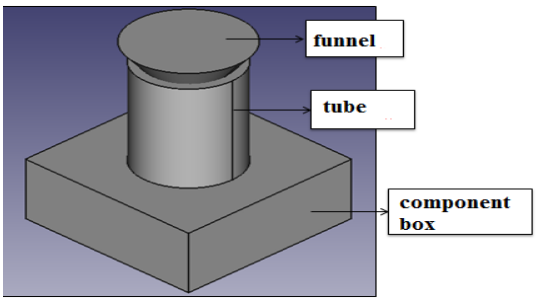


Fig. 3. Design Results of Monitoring Android-based Rainfall Levels.

The following is Fig. 3, the results of the design model of the rainfall level detection tool based on Arduino Uno.

Based on Fig. 3, the design of the data system from the design of an Android-based rainfall level monitoring tool explains that the volume of collected rainwater will provide measurable value by utilizing the Infrared Sensor. The output of the Infrared sensor will be read by Arduino Uno after passing the Arduino signal conditioning circuit. The analog value produced by the sensor will be read by Arduino and forwarded to the Water Pump to pump water so that the water comes out of the tube and the data will be stored, then ESP8266 The Wi-Fi module will send data to Android.

C. Algoritma

Calculating altitude using equations that are already widely known [20]. To calculate the height of rainwater that falls in a field with a certain area you can use the equation:

$$H = \frac{V}{L} \quad (1)$$

H = height of rainfall

V = Volume

L = area

A funnel with a diameter of 14 cm is used in measuring rainfall, meaning that the area of the funnel can be calculated by the equation:

Count the funnel radius as follows:

$$R = \frac{(\text{Funnel diameter})}{2} \quad (2)$$

Area calculation as follows:

$$L = \pi \times R^2 \quad (3)$$

Known :

$\pi = 3.14$

R = funnel radius

Volume calculation is as follows:

$$V = L \times \text{tube height} \quad (4)$$

Known :

V = Volume

L = Field area

Calculates the standard deviation as follows:

$$\text{Standard Error} = \text{Accurate Data} - \text{Measured Data} = \frac{x}{\text{Accurate Data}} \quad (5)$$

Known :

N = 4

Measured Data = 53.2 mm

Accurate Data = 76.50

III. Results and Discussion

The results and discussion of the design that has been made can be described in the following steps.

A. Infrared Sensor Testing

Infrared sensor testing is done by distance testing using a ruler as a measuring medium, the match between the distance from the meter and the results displayed on the web and android. Fig. 4, the results obtained show the water level used in sensor testing.

B. Hardware Testing

The circuit in Fig. 5 will turn on the relay and the pump will automatically turn on and suck the water inside the tube. Data transmission will be carried out when the relay and pump are on. Data will be sent to the database and android in the form of a graph. The graph will run smoothly if the internet connection used has a good network. This circuit uses a 12-volt voltage generated from the adapter.

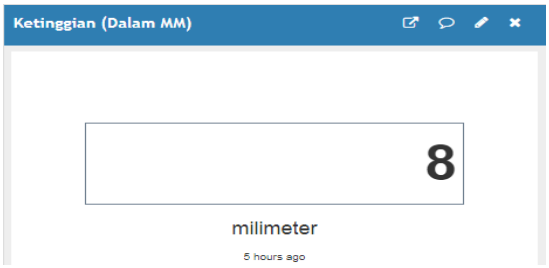


Fig. 4. Infrared Sensors.

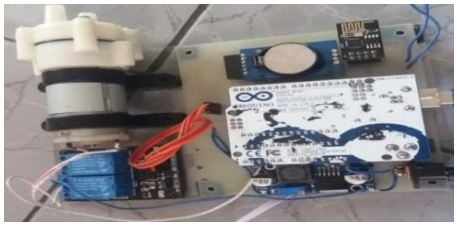


Fig. 5. Hardware Testing.

C. Software Testing

In testing Fig. 6, programs written in the Arduino IDE software will be checked for program listings whether or not the program has errors that can interfere with the performance of the system. This test is done by combining sensor listings with other component listings.

D. Testing of Rainfall Measurement Tools

The testing of rainfall testing equipment is carried out in the event of rain or heavy rain. This measurement is done by measuring the height of water entering the tube will be measured automatically with the sensor.

Calculate funnel radius using equation (2) as follows:

$$R = \frac{14}{2} = 7$$

Measurement of rainfall using a funnel, used a funnel with a diameter of 14 cm and a height of 26 cm. The area calculation below uses equation (3):

$$L = 3{,}14 \times 7 \times 7 = 153{,}86$$

Calculate volume using equation (4):

$$V = 153{,}86 \times 26 = 4.000{,}36$$

Calculate the height of rain using equation (1):

$$\text{Rainfall height (H)} = \frac{4.000{,}36}{153{,}86} = 26000 \text{ mm}$$

Table I is a table that will contain the results obtained during the measurements that have been made. After doing the calculation, the average value of the calculation will be used.

The results obtained in Table I show that the average rainfall obtained in this measurement is low. Data were taken every time it rains only 1 data and data used is data that approaches the value of manual calculation. After the measurement has been completed calculation. Table II is a table of manual calculations on the BMKG tool. In Table I and Table II, a comparison is made to determine the standard error of the measuring instrument.

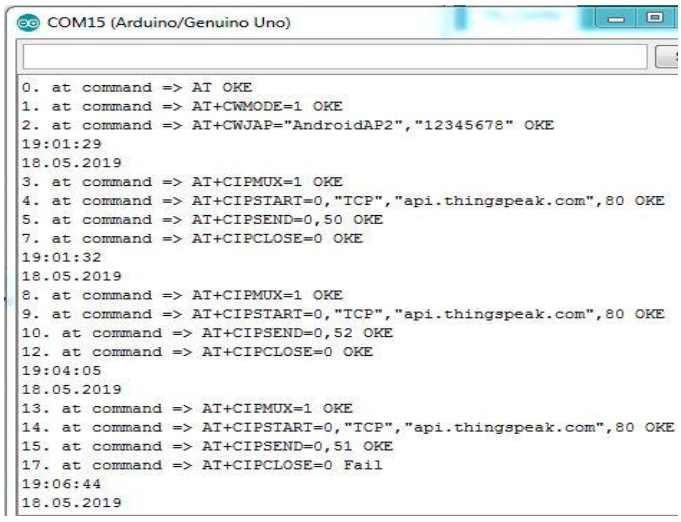


Fig. 6. Display on COM 15.

TABLE. I. RAINFALL MEASUREMENT

No.	Date/Time	Time (WIB)	Water Level (mm)	Explanation
1.	Wednesday, 15 May 2019	03:32	17,0	Low
2.	Friday, 17 May 2019	23:35	53,2	High
3.	Saturday, 18 May 2019	08:31	38,0	Low
4.	Wednesday, 22 May 2019	10:22	0	Low
5.	Thursday, 23 May 2019	14:10	5,4	Low
6.	Friday, 24 May 2019	16:07	0	Low
7.	Saturday, 25 May 2019	05:48	0	Low
n = 7			261,7	
Average			38,3	

The manual measurement carried out by BMKG in Table II has a description of low rainfall so that it does not cause flooding in an area. The importance of BMKG manual measurement is done as a comparison of measurements on the measurement tools of the current research so that researchers can find out how much error occurred in the research measuring instrument.

After the data obtained by the measurement data will then look for the standard deviation of the rainfall gauge for 7 days, the standard deviation formula is obtained using equation (5):

$$\text{Standard Error} = 76{,}50\text{ mm} - 53{,}2\text{ mm} = \frac{23{,}3\text{ mm}}{76{,}50\text{ mm}} = 0{,}30$$

$0{,}30 \times 100\% = 30\%$

TABLE. II. MANUAL BMKG MEASUREMENT OF RAINFALL

No.	Date/Time	Time (WIB)	Water Level (mm)	Explanation
1.	Wednesday, 15 May 2019	03:32	17,0	Low
2.	Friday, 17 May 2019	23:35	76,50	High
3.	Saturday, 18 May 2019	08:31	46,2	Low
4.	Wednesday, 22 May 2019	10:22	0	Low
5.	Thursday, 23 May 2019	14:10	6,4	Low
6.	Friday, 24 May 2019	16:07	0	Low
7.	Saturday, 25 May 2019	05:48	0	Low
n = 7			103,1	
Average			14,73	

TABLE. III. STANDARDS OF RAINFALL DEVIATION

No.	Date/Time	*Error*	Percentage
1.	Wednesday, 15 May 2019	0	0
2.	Friday, 17 May 2019	0,30	30%
3.	Saturday, 18 May 2019	0,18	18%
4.	Wednesday, 22 May 2019	0	0%
5.	Thursday, 23 May 2019	0,16	16%
6.	Friday, 24 May 2019	0	0%
7.	Saturday, 25 May 2019	0	0%

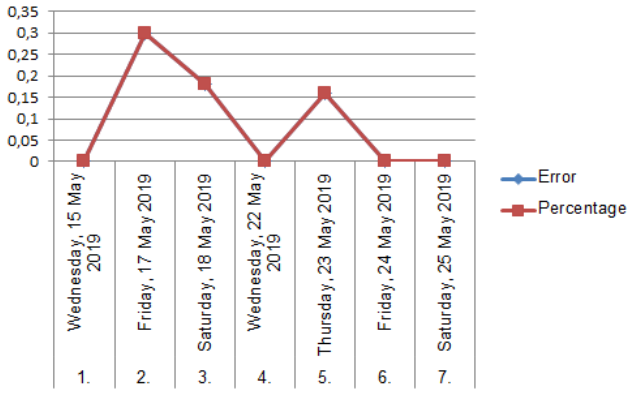


Fig. 7. Graphic Standard Deviation of Rainfall.

After calculating the standard deviation in Table III using equation (2), the result of the error value is 30%.

The results of error Table III were carried out from a comparison of current research and the BMKG manual measurements obtained were quite large. So we can find out what capacity can be measured from this research tool. Fig. 7 shows a graph of the comparison of research tools and BMKG tools.

E. Testing the Application on an Android Phone

Android testing in Fig. 8 uses the Thingview application. This test is carried out to monitor the level of rainfall at home using Android. The following is a view of Thingview that can display data sent by Thingspeak.

Data transmission from Thingspeak to Thingview obtained in Fig. 8 is a test performed on android and performs the process of sending data 26 times. Data transmission that is the process of sending it via Android results obtained is the same as the process of sending it through the website. Data transmission on Android is done every 3 minutes so that the water in the tube does not occur evaporation. This remote monitoring is very easy for us to monitor rain that occurs in an area without the need to manually check the measuring instrument.

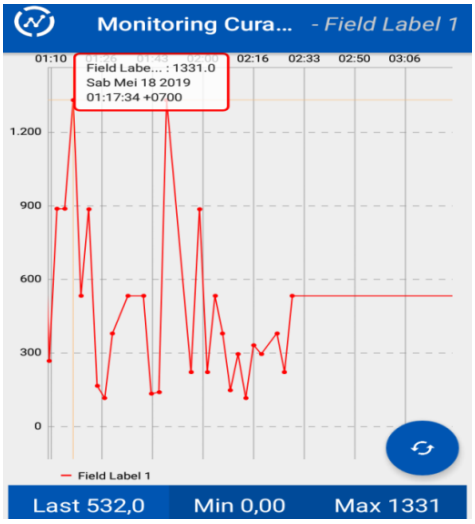


Fig. 8. Testing Android.

IV. Conclusion

The design of the tool building model in this study uses a funnel with a diameter of 14 cm and a height of 26 cm. Based on the design calculations that have been made measuring devices are designed to be able to measure rainfall up to a height of 26000 mm. The research that has been done can measure the rainfall in the measuring tube using the Sharp GP2Y0A41SK0F sensor which has been working in accordance with its functions and uses. The minimum sensor distance sensor measurement is 4 cm. The results obtained from the sensor readings will be received on the website and can be accessed via Android. All levels of rainfall monitoring based on Android can be used properly and can be used such as the use and function of sensors that have been used. Test tool in 7 days to get accurate data about the process of sending data stored on the website and the process of sending data can be received via Android. Data obtained from participation in rainfall calculation has an error of 30%. The development of tools that should be done in the future is to compare research tools with BMKG tools directly.

Acknowledgment

Acknowledgments are given by LPPM UAD for providing PKLN 2019 research funds with contract numbers PKLN-PI-001/SP3/LPPM-UAD/IV/2019.

References

[1] Athoillah, R. M. Sibarani, and D. E. Doloksaribu, “Analisis Spasial El Nino Kuat Tahun 2015 Dan La Nina Lemah Tahun 2016 (Pengaruhnya Terhadap Kelembapan , Angin dan Curah Hujan di Indonesia),” J. Sains Teknol. Modif. Cuaca, vol. 18, no. 1, pp. 33–41, 2017.

[2] D. R. Prabawadhani, B. Harsoyo, T. H. Seto, and M. B. R. Prayoga, “Karakteristik Temporal Dan Spasial Curah Hujan Penyebab Banjir Di Wilayah Dki Jakarta Dan Sekitarnya,” J. Sains Teknol. Modif. Cuaca, vol. 17, no. 1, pp. 21–25, 2016.

[3] A. Yudhana, N. S. Widodo, and Sunardi, “Implementation of Quadcopter for Capturing Panoramic Image at Sedayu Bantul,” Proceeding 1st Int. Conf. Comput. Sci. Eng., pp. 37–39, 2014.

[4] Misnawati, R. Boer, T. June, and A. Faqih, “Perbandingan metodologi koreksi bias data curah hujan chirps,” LIMNOTEK Perair. Darat Trop. di Indones., vol. 25, no. 1, pp. 18–29, 2018.

[5] D. Mulyono, J. Konstruksi, S. Tinggi, T. Garut, and C. Hujan, “Analisis karakteristik curah hujan di wilayah kabupaten garut selatan,” J. Konstr., vol. 13, pp. 1–9, 2014.

[6] A. Petonengan, J. S. F. Sumarauw, and E. M. Wuisan, “Pola distribusi hujan jam-jaman di das tondano bagian hulu,” J. Sipil Statik Vol.4, vol. 4, no. 1, pp. 21–28, 2016.

[7] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, “Towards the Implementation of IoT for Environmental Condition Monitoring in Homes,” IEEE Sens. J., no. c, 2013.

[8] A. Yudhana and A. C. Kusuma, “Water quality monitoring at paddies farming based on android,” IOP Conf. Ser. Mater. Sci. Eng., 2018.

[9] A. B. Vernandez, M. Facta, and T. Sukmadi, “Perancangan Konverter DC-DC Topologi Buck Berbasis Mikrokontroller Atmega8535 Untuk Aplikasi Heater Ekstraksi Biji Kapuk,” 2015.

[10] D. W. E. Sede, A. A. E. Sinsuw, and X. B. N. Najoan, “Rancang Bangun Aplikasi Pemesanan Tiket Online Kapal Laut Berbasis Android,” vol. 1, no. 1, 2015.

[11] A. A. Handaru et al., “Menggunakan Modul GSM Berbasis Mikrokontroler,” Sci. Electro, pp. 25–30, 2019.

[12] M. A. Novianta, “Sistem Data Logger Curah Hujan Dengan Model Tipping Bucket Berbasis Mikrokontroller,” J. Teknol., vol. 4, no. 28, 2011.

[13] R. G. Permana, E. Rahmawati, and Dzulkiflih, “Perancangan Dan Pengujian Penakar Hujan Tipe Tipping Bucket Dengan Sensor Photo – Interrupter Berbasis Arduino,” J. Inov. Fis. Indones., vol. 04, pp. 77–76, 2015.

[14] M. J. D. Suryanto and T. Rijanto, “Rancang Bangun Alat Pencatat Biaya Pemakaian Energi Listrik Pada Kamar Kos Menggunakan Modul Global System For Mobile Communications(Gsm) 800l Berbasis Arduino Uno,” J. Tek. ELEKTRO, vol. 8, pp. 47–55, 2019.

[15] S. Hidayat and Mushlihudin, “Alat Ukur Tinggi Muka Air Berbasis,” J. Ilmu Tek. Elektro Komput. dan Inform., vol. 3, no. 2, pp. 96–100, 2017.

[16] A. Yudhana, Sunardi, and A. Ikrom, “Aplikasi Android Untuk Monitoring Kualitas Lahan Pertanian,” Pros. SNST Fak. Tek., pp. 7–12, 2018.

[17] A. Yudhana and M. D. D. Putra, “Rancang Bangun Sistem Pemantauan Infus Berbasis Android,” Transmisi, vol. 20, no. 2, pp. 91–95, 2018.

[18] U. Syafiqoh and A. Yudhana, “Pengembangan Wireless Sensor Network Berbasis Internet of Things untuk Sistem Pemantauan Kualitas Air dan Tanah Pertanian,” J. Pengemb. IT, no. 02, pp. 285–289, 2018.

[19] A. Adriansyah and O. Hidyatama, “Rancang Bangun Prototipe Elevator Menggunakan Microcontroller Arduino Atmega 328p,” J. Teknol. Elektro, Univ. Mercu, vol. 4, no. 3, pp. 100–112, 2013.

[20] BMKG, “Mengenal Alat Ukur Cuaca dan Penakar Hujan Sederhana Serta Kalibrasinya,” BADAN Meteorol. Klimatologi DAN Geofis. Stasiun Klimatologi Mlati, 2015.

Deep MRI Segmentation: A Convolutional Method Applied to Alzheimer Disease Detection

Hanane Allioui[1], Mohamed Sadgal[2], Aziz Elfazziki[3]
Computer Science Department
Faculty of Sciences Semlalia
Cadi Ayyad University
Marrakech, Morocco

***Abstract*—The learning techniques have a particular need especially for the detection of invisible brain diseases. Learning-based methods rely on MRI medical images to reconstruct a solution for detecting aberrant values or areas in the human brain. In this article, we present a method that automatically performs segmentation of the brain to detect brain damage and diagnose Alzheimer's disease (AD). In order to take advantages of the benefits of 3D and reduce complexity and computational costs, we present a 2.5D method for locating brain inflammation and detecting their classes. Our proposed system is evaluated on a set of public data. Preliminary results indicate the reliability and effectiveness of our Alzheimer's Disease Detection System and demonstrate that our method is beyond current knowledge of Alzheimer's disease diagnosis.**

Keywords*—Computer-Assisted Diagnosis (CAD); Alzheimer's disease (AD); Image segmentation; Machine learning; Convolutional Neural Networks (CNN); Magnetic Resonance Imaging

I. Introduction

Alzheimer designates an incurable disease that attacks brain tissues and influences mental functions, as well as memory. This neurodegenerative disease is characterized by brain damage, including β-amyloid peptide (Aβ), neurofibrillary tangles and neuronal degeneration, which chronically damage the brain in an irreversible way [1]. Scientific works on AD detection began around 1906 [2]. However, it took years to develop the appropriate algorithms to analyze information processing in the brain. Consequently, in the last decades, several systems based on CAD were proposed [3, 4]. However, most of the proposed systems have been based on manual functionalities requiring too much precision and concentration to study MRI modalities.

The automated Computer-Aided Diagnosis of Alzheimer remains significantly a challenging task that requires advanced technical practices such as Deep Learning (DL) algorithms. Recently, deep learning has shown promising methodologies with great progress in the segmentation, identification, and classification of image patterns [5, 6, 7]. Among the most widely used DL architectures, convolutional neural networks (CNNs) perform machine learning tasks without manual functions [8, 9]. CNN has a strong ability to solve complex vision problems, such as classification [10, 11], segmentation [12] and object detection [13, 14].

The CNN network consists of different layers able to analyze and capture the quantity of the different structures present in the image. The extension of CNNs has contributed to the development of U-Net that allows the segmentation of biomedical images. The U-net is based entirely on a convolutional network, with a modified U-shaped architecture, which can perform the required tasks with fewer training images to produce more accurate segmentation swiftly using a GPU [15]. As for medical imaging, it is evident that U-Net architecture was introduced to promote precision and objects' localization in microscopic structures. U-Net combines a fully convolutional network [16] with a deconvolutional network [17]. Consequently, the resolution of the output can be ensured thanks to the number of features in the up-sampling phase, which ensures the propagation of context information to higher resolution layers.

This paper describes an improved CAD system for Alzheimer detection using 2.5D modalities. The ultimate contribution of our work is to provide a different Alzheimer detection process from several perspectives. First, the use of a robust neural network that can process the MRI scans on a 2.5D spatial context to improve the performance of segmentation and so that the whole system. Second, our method simultaneously performs brain structure segmentation for lesion detection and AD detection. Third, we tested our work on public data. Finally, we developed a novel 2.5D fully deep convoluted segmentation based on the U-Net architecture to increase the results accuracy.

This paper is ordered as follow: Section 2 states the literature review regarding Alzheimer detection techniques. Section 3 explains our proposed technique in detail. Section 4 illustrates the experimental results and Section 5 presents the conclusion of the work.

II. Literature Review

Practically, the diagnosis of AD is based on various essentials, such as genetic information, demographic factors, psychological and neuropsychological tests, indices of cerebrospinal fluid biomarkers (CSF) and brain imaging data [18]. Specifically, Numerous neuroimaging researches studied the region of interest (ROI) to inspect the brain changes caused by AD [19]. These works were based on prior knowledge to guide the selection of ROIs, which may neglect the consideration of brain changes outside of the studied areas.

Otherwise, in [20] authors presented a spherical mapping based on 3D Brain images projection into 2D using statistical characteristics of brain tissue. Anitha et al. improved the watershed algorithm and presented a method to define the diseased area using shape analysis techniques. In the same way, Kalavathi et al. [21] used FFCM [22] to segment brain images whereas, machine-learning (ML) techniques offer the possibility of using systematic methods to ensure sophisticated, automatic classification and object-based detection [23]. The ML or more properly DL methods can learn from different models and manage complex changes in neuroimaging modalities, to ensure the analysis of large data with high accuracy.

During the last decades, the need for medical image processing has known explosive growth due to the development of technologies and the large-scale detected diseases especially the Alzheimer. Accordingly, intelligent vision systems have grown so fast. So, the understanding of brain images remains an area of rising interest in AD studies. Through the literature, different classification methods have shown good evidence. Dahshan et al. [24] used an artificial neural network (ANN) and K-nearest neighbor (KNN) networks for approximation extraction and detail coefficients calculation of the 3-level discrete wavelet transform (DWT). In [25]Park used learning for classification using two distance measurements which resulted in an error rate of 18% for the classification of AD cases and 46% for normal cases. Zang et al. [26] proposed a volumetric image classification method that implements a discrete 3D wavelet transform (3D-DWT) for wavelet coefficients extraction. Zhang et al. [27] presented an early AD detection based on Eigen brains and machine learning using SVM classifiers to detect accurately AD subjects as well as the AD-related brain regions.

With the advent of Computer Assisted Diagnosis systems, especially these, relying on deep learning techniques, The CNNs [28, 29] are frequently used to solve complicated problems from both computer vision and medical imaging field. Indeed, these models learn to establish a list of image characteristics. Since CNNs are much used for MRI images studies to detect or predict AD, it is necessary to develop methods ensuring the learning from a large-scale training package [30]. The robustness of the Capsule or CapsNets [31] networks allows a fast, precise and thorough learning of data images. CapsNet requires less data for training by ensuring shorter learning curve [32]. Other methods have used different auto-encoders or CNN 3D to detect AD [33, 34].

Comparative evaluations were presented in [35], the authors presented a literature review of machine learning techniques used for AD detection. Their work illustrates the quality assessment of the previous studies and a comparison of the details. In addition, in recent years, the family of deep learning approaches has experienced considerable improvements, notably U-Net networks. Authors in [36] introduced a network that extends the u-net architecture from [37] by replacing all 2D processes with their 3D equivalent. Regarding brain segmentation, Chang et al. [38] proposed an effective 3D U-Net model to improve segmentation accuracy and feature labeling.

In contrast to all the previous solutions, the proposed method takes into consideration the incorporation of 3D spatial information in a 2.5D context using a deep convolutional network to ensure segmentation tasks. Our choice for U-Net is based on the need to boost the segmentation accuracy and the performance of image processing while studying a sensitive disease such as Alzheimer.

III. Proposed Technique

Taking into account the 2D and 3D modalities in image processing, we choose to adopt a 2.5D method which designates a set of techniques that take advantages of 3D features with fewer complexities. This involves extending the dimension of the lowest resolution of the MRI volume images into the RGB dimension providing different benefits:

- This polysemous approach enables a better representation of the 3-D features with lesser computational costs.
- The information of each slice of the MRI images can be exploited by transforming the 3D images into grayscale images with all the 3D information.
- The integration of 3D volumes can be achieved with less design and implementation requirements by optimizing memory, reducing complexity and ensuring flexibility.

A. 3D to 2.5D Transformation and Data Preprocessing

Our work aims AD by optimizing resources. For that, we have to cut the 3D image into a 2.5D image. Although the 2.5D concept was previously used in several works [39, 40], we used it to explore and characterize the emerging integration options between 2D and 3D brain images. The 2.5D is used in our work to present a sub-volume of dimensions ($x \times y \times z=3$) where x, y, and z are the MRI slice dimensions. Each 3D image contains several 2.5D scans, which cover the full studied area. To facilitate the implementation of our system, it is useful to take advantage of depth information (z-direction) in MRI images.

The point cloud C defines a set of points in a 3D volume like the example illustrated in Fig. 1, C can be written by C (xs, ys, 3), where s is the spatial sampling step of the system and (x, y) are positive integers belonging to $\mathbb{Z}2$. Therefore, the transformation of C into 2.5 projections can be performed without loss of data. Each point C (xs, ys, 3) will be converted to a pixel P (x, y). Indeed, the value of the gray level of this pixel P will be Z - Zmin, where Z min is the minimum depth of C. Next, in volume 2.5, the value of each pixel corresponds to the vertex of the point upper of Zmin. In contrast, the 2.5d transformation in 3D corresponds to a simple conversion of the acquisition parameters to assemble the adjacent images and form the 3D volume.

Our method has been tested using public data from OASIS [41] which is an Open Access database containing several Series of imaging studies. The OASIS project has been initiated to ensure the availability of neuroimaging datasets to the scientific community. This database contains cross-sectional T2 and longitudinal T1 datasets. We chose to use the

T1w [42] data which was taken at the time of the disappearance of the transverse magnetization. These datasets have been used in previous studies concerning Alzheimer disease such as [43, 27]. In fact, the OASIS database includes 489 subjects ranging in age from 42 to 96 years old. However, our selection includes 126 cases after exclusion of cases whose information is not consistent as well as the patient's cases under the age of 60 years. The MRI images from the OASIS database have been pre-processed to ensure the quality of learning, so it is essential to standardize patient data. As described in [44, 45], the effects of normal aging have been removed from MRI images as they are likely to be similar to the effects caused by Alzheimer. This could cause problems of confusion regarding the estimation of the specificities of each case.

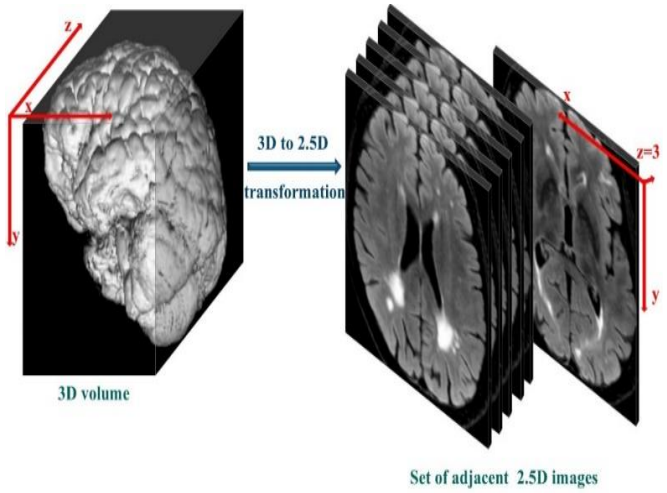


Fig. 1. Example of 3D to 2.5D Transformation.

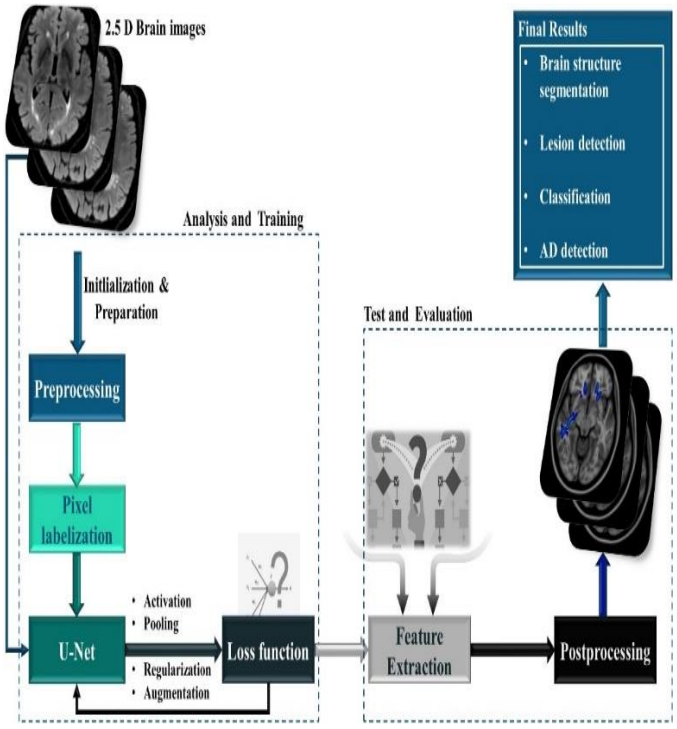


Fig. 2. Our Detailed Scheme of the Proposed CAD.

B. The Proposed CAD Architecture

In our system, the application of CNNs consists of convoluting a 2.5D image with the kernel in order to extract the maps of existing characteristics in the studied images. In fact, each map is connected to the previous layer of the network through the appropriate weights during the training to improve the input. The same kernel is converted on all the data of the image which leads to the use of several functions as will be explained in the following subsections. The proposed process is summarized in Fig. 2. We used the U-Net architecture for each view of the 2.5D brain MRI after parsing them into transversal views. The rationale behind the choice of this architecture is based on the sensitivity of the zones studied, and the limitation of memory facing the performed operations. Consequently, we have been able to optimize the formation of convolutional networks by limiting the problem of MRI images to a 2.5D domain.

Our method offers an AD diagnosis framework that extracts the characteristics of brain images, performs correct lesion detection, classifies their natures, and indicates the presence of AD. The operations performed are designed to maximize the content of the information from different views. The details of this process are related to the following concepts:

- The initialization phase: it is the phase ensuring the convergence to launch the other tasks such as the activation and the regularization.
- The activation function which transforms an input to a set of outputs by imparting nonlinearity to the network structure. Thus, devoted to deep depreciation operations, the Rectifier linear units (ReLU) has demonstrated the acceleration of training while ensuring better results comparing with sigmoid functions [46].
- The Pooling phase allows the combination of neighboring entities spatially in feature maps. The redundancy of this combination provides a denser and less invariant representation for changes that may occur in the image and reduces the computational charges.
- The regularization reduces the overfitting by forcing all the nodes of the network to learn an improved representation of data and preventing nodes from co-adapting to each other.
- The augmentation allows increasing the size of training sets and reducing the overfitting [47].
- The loss function defines the best optimal solution by determining the correct values for all weights. It's based on a learning model that minimizes the loss during the processing. For that reason, the loss function has to be minimized during the training.

In medical imaging, some tasks remain complicated such as labeling the learning images. This operation can be expensive and involve delicate and ambiguous decisions. For instance, in brain image processing, annotating the locations and scales of objects is often difficult, hence the need for a consistent way that annotates all the objects segment and detects their natures.

In our article, we deal with a learning network to segment MRI images and form lists of the contained objects. As long as we rely on the U-Net architecture, we have introduced some changes. First, we introduce a function in the pooling layer that assumes the possible location of the segmented objects. Secondly, the cost function retrieves the information from the image and signals the location of objects or their distinctive part in the studied images.

1) Our U-Net Network architecture: We present our CAD system that leverages the power of 3D image processing by adopting a deep fully convolutional neural network, trained end-to-end. U-Net allowed us to ensure a precise sense of location. The goal of our architecture is to deal with brain image processing tasks to have an accurate analysis that can find out brain lesions, their localization as well as AD diagnosis. The U-net network has revealed promising results on biomedical images [48] and natural images [49].

Fig. 3 outlines our U-Net network. Our CAD system is based on the U-Net which includes a contracting (down-sampling) path and an expanding (up-sampling) path.

The contracting path consists of 5 convolution blocks that each of them contains two convolutional layers of size (3 × 3), a stride of 1 in addition to a rectifier activation which makes it possible to increase the number of cards from 1 to 1024. Except for the last block, by the end of each block, the max-pooling (2 × 2) is applied. In the expanding path, each of the blocks begins with deconvolution of size (3 × 3) in addition to a stride of (2 × 2). Indeed, each convolutional layers of the contracting part (encoder) has a corresponding decoder in the deconvolutional layers. Finally, we opted for a SoftMax layer. Then at the end of our network a classification layer to determine the classes of the detected objects in the studied brain images.

2) Training and optimization: Once we have completed the design of our network, tested the structure we start the training process. For that, extracting global characteristics of MRI images requires large training and involves costly computation, given a large number of test parameters crossing the input/output layers. To evaluate our 2.5D approach, the Soft Dice metric was used as a network cost function during the formation phase, this metric representing a differentiable form of the DSC (DSC) [50]. For better efficiency, in the processing of MRI images, optimization based on a stochastic gradient is essential during the training phase to minimize the cost function according to its parameters. After training, the intensity is normalized by providing a linear transformation of the original intensities between two features into the corresponding learned ones. This allows the similarity of the histogram of each sequence between subjects. The time required was of the order of ten minutes.

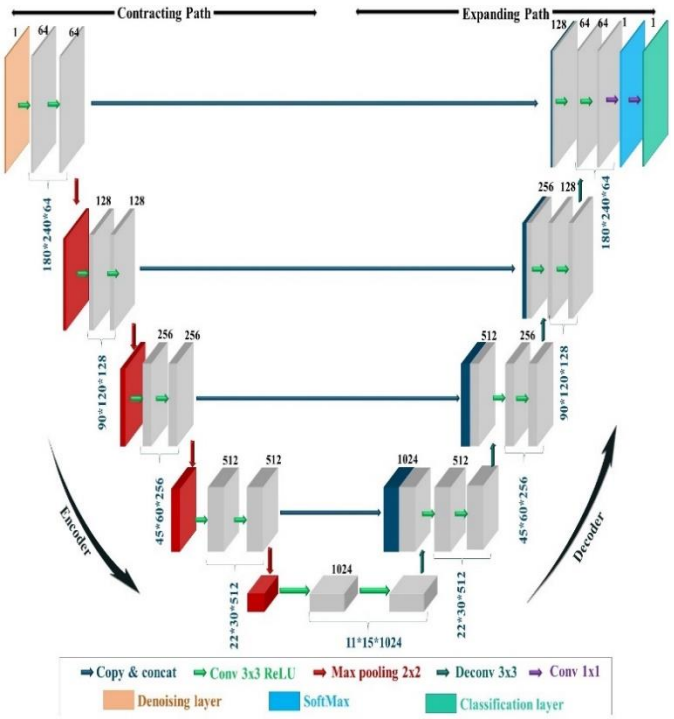


Fig. 3. Our U-Net Architecture.

IV. Experiments and Results

A. Experiments and Performance Evaluation

In this paper, we have proposed and developed a CAD system using deep convolutional networks to analyze brain damage and detect Alzheimer's disease by segmenting 2.5D images. As long as we have added a classification layer in our network, we need accurate measures to calculate the proportional values between the studied samples. For this purpose, we considered three metrics: accuracy, sensitivity, and specificity [51, 52]. These metrics are defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+2TN+FP,} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN,} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

Where:

- TP represents the value of the true positive (the number of cases having an AD correctly).
- TN is the true negative (the number of control cases)
- FP represents the value of the false positive (the number of control cases with AM)
- FN is the false negative value (the number of cases with AD considered as controls).

B. Results and Discussion

As shown in Fig. 3, our method involves of a fully CNN containing two main parts: The U-net to segment the brain images and the classification layer to classify the detected areas. The input image will first go through the contracting path to generate feature maps, then the expanding path to generate the output segmented image. After that, the detected objects in the segmented images would be classified to detect their nature and improve the network model's adaptability.

In the first experiment, height hundred images are used in order to train and validate our model. An example of the results is shown in Fig. 4 which illustrates the detection and segmentation of endometrial areas of the brain are part of the semantic segmentation. After the training phase, we carried out different sets of evaluations. This allowed us to add new examples to our initial training. In this phase, we compared the segmentation result with the classification of detected objects in cerebral images. The network can detect cases of brain damage or lesion as shown in case a. They may affect brain abilities and cause different problems. In these cases, the patient may have similar signs similar to those of AD, yet these lesions can be treated. Case b shows both brain damage and Alzheimer disease presence in the early stages. Finally, case c denotes the clear presence of AD, which shows a big difference comparing to healthy aged brains.

Currently, there is no perfect available method for image segmentation and object detection. For performance verification, the experimental evaluation is performed using a series of different previous methods. In the second experiment, we tested the results of our developed network, as well as different methods which don't need a prior training domain were known, are evaluated. All these methods are based on convolutional structures. Hosseini et al. [53] presented an adapted 3D Convolutional Network to predict the AD and proved competitive results. However, this work was dedicated to only image classification for 3D images. This indicates the need to adapt the network if we want to process 2.5D images to have high efficiency and update some functions to ensure image annotation concerning the segmented areas. Kalavathi et al. [21] proposed a method based on segmentation of the brain using contours to eliminate the area of the hug, then they applied Fast Fuzzy C Means (FFCM) to segment brain tissue to detect the existence of Alzheimer's disease. Hao Dong et al. [54] presented a reliable method for automatically segmenting brain tumors. They used deep U-Net convolution networks. They had good results for the detection of tumor regions. Yet they don't take into consideration brain damages without AD presence.

As illustrates the example in Fig. 5 the results demonstrate the efficacy of our method. Another important outcome is the number of wrong region detection without the classification stage. The methods in Fig. 5(b) and Fig. 5(d) show good results, however they can segment other objects different from the zones attacked by the Alzheimer which can create a kind of confusion during the diagnosis. In Fig. 5(c) the method ensures the segmentation of the brain by determining different objects other than lesions or AD. This phenomenon is mainly due to the fact that even the semantic segmentation is not only sufficient to diagnose AD. That's why it's preferable to combine several techniques to achieve higher accuracy.

Compared to the quantitative results of the different mentioned methods, the validation of the proposed system demonstrated good results for the complete brain segmentation. Using our method allows us to achieve an Accuracy rate of 92.71%, Sensitivity of 94.43%, and a Specificity rate of 91.59%.

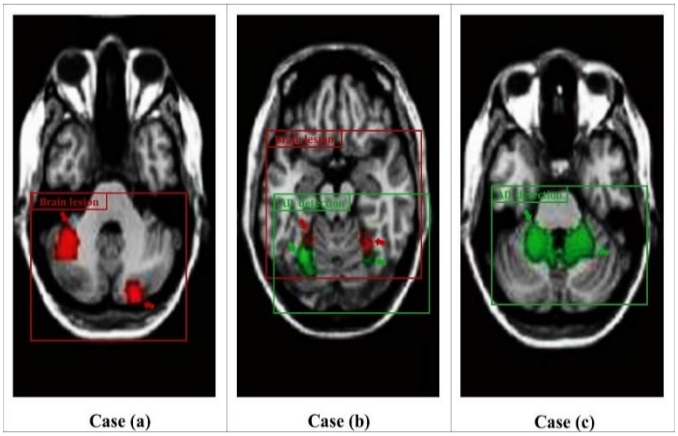


Fig. 4. Example of Possible Results [Case (a): Presence of many Lesions that may Affect Brain Function; Case (b): Significant Loss of different Brain Areas, with Lesions Present (Red Color), Plus an Onset of Alzheimer's Disease (Green Color); Case (c): Strong Presence of AD Compared to Healthy Patients.].

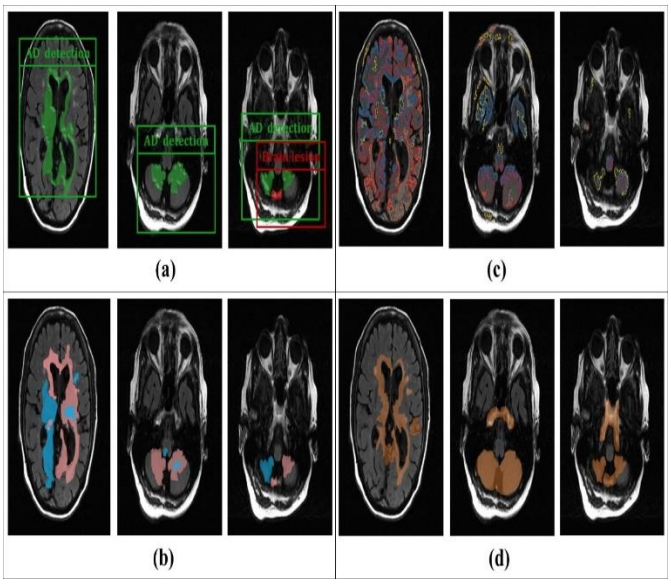


Fig. 5. Example of the different Tested Methods Results [a-Our Proposed Method, b- Hao Dong [54], Kalavathi et al. [21], Hosseini et al. [53]].

TABLE. I. QUANTITATIVE RESULTS OF OUR PROPOSED METHOD COMPARED TO THE RESULTS FROM OTHER METHODS

Tested method	Accuracy	Sensitivity	Specificity
a-Our method	**92.71 %**	**94.43 %**	**91.59 %**
b- Hao Dong [54]	89.75 %	85.85 %	87.34 %
c- Kalavathi et al. [21]	55.83 %	56.28 %	51.11 %
d- Hosseini et al. [53]	90.31 %	91.18 %	90.51 %

To quantitatively assess the effectiveness of the methods tested we present, in a random order, the results of the evaluation methods as shown in Table I. To ensure better comparison we used the same datasets from Oasis to assess them. we conclude that the perfect segmentation of cerebral images is not granted by any method. However, the proposed method was able to take first place in terms of segmentation performance taking into account the use of the same evaluation basis. According to these results, we report the high-efficiency of our proposed network.

In our process, firstly the down-sampling and training are made. Then the up-sampling outputs are classified. Visually the results indicate that our proposal can segment brain images to diagnose AD; the task that requires physician's interactions in the general clinical routine; by offering blended annotation on the studied images. Moreover, our quantitative results reveal further efficiency with the additional classification layer. This layer facilitates AD detection with less additional efforts.

The proposed network offers accurate results in less time (less than 2 hours for training and testing). These accomplishments are superiors than all the tested methods that can take more hours or even days for brain images analysis. Our method is an appropriate automatic solution for brain image segmentation and AD detection. Yet, the current work still has some limitations. First, our method was evaluated to segment 2.5D images using a T1 longitudinal dataset, but running our method on a t2 dataset can produce a more objective evaluation. Secondly, the evolution of neural networks continues to increase, so the addition of new parameters can improve the performed tasks by our network.

V. Conclusion

In this work, we proposed a developed U-Net architecture for brain images segmentation intended for Alzheimer disease and brain damage detection. We found-out that U-Net improves clearly state-of-the-art. The benefits of our method were demonstrated in comparison to different relevant methods. Our main contribution is providing an automatic and exact Alzheimer detection using an advanced full neural network on a 2.5D context to ensure the performance of segmentation whatever the used system, by reducing memory costs while processing 3D images. Moreover, our method performs brain segmentation then classification for AD detection. To our best knowledge, our proposed CAD system is the first one 2.5D MRI analysis for Alzheimer's disease detection using such learning techniques. Our developed U-Net can automatically segment a 2.5D MRI image and offer an accurate analysis of brain structures. The network to segment brain images was trained from scratch, for that we expect that it will be applicable to many brain analyses especially 2.5D segmentation tasks.

Our deep U-Net network has been able to obtain competitive results in detecting damaged regions of the brain and defining the presence of AD. The proposed method generates a model of automatic segmentation of brain lesions as well as the diagnosis of patient-specific AD, which can potentially facilitate different clinical tasks such as diagnosis, treatment planning, and patient monitoring.

References

[1] J. Sevigny et al., "The antibody aducanumab reduces Aβ plaques in Alzheimer's disease", Nature, vol. 537, pp. 50–56, 2016.

[2] NC. Berchtold and CW. Cotman, "Evolution in the conceptualization of dementia and Alzheimer's disease", Neurobiology of Aging. vol. 19, n. 3, pp. 173–189, 1998.

[3] R. Chaves et al., "SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting", Neuroscience Letters, vol. 461, pp. 293–297, 2009.

[4] E. Bron et al., "Standardized evaluation of algorithms for computer aided diagnosis of dementia based on structural MRI: the CAD Dementia challenge", Neuroimage, vol. 111, pp. 562–579, 2015.

[5] F. Li et al., "A Robust Deep Model for Improved Classification of AD/MCI Patients", IEEE Journal of Biomedical and Health Informatics, vol. 19,n. 5,pp. 1610 - 1616, 2015.

[6] N. Zeng et al., "Facial expression recognition via learning deep sparse autoencoders", Neurocomputing, vol. 273, pp. 643–649, 2018.

[7] W. Lin et al., "Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment" Frontiers in Neuroscience, vol. 12, n. 777, 2018.

[8] D. Nie et al., "Fully convolutional networks for multi-modality isointense infant brain image segmentation", Procedding of the 2016 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1342-1345, 2016.

[9] H. C. Shin et al., "Deep Convolutional Neural Networks for Computer Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning", IEEE Transactions on Medical Imaging, vol. 35, n.5, pp. 1285-1298, 2016.

[10] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[11] C. Szegedy et al., "Going deeper with convolutions", The IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[12] K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation", Medical Image Analysis, vol. 36, pp. 61-78, 2017.

[13] C. Szegedy et al., "Deep neural networks for object detection", Advances in Neural Information Processing Systems, vol. 26, pp. 2553–2561, 2013.

[14] P. Sermanet et al., "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks", Computer Vision and Pattern Recognition, 2014.

[15] M. Beyeler et al., "A GPU-accelerated cortical neural network model for visually guided robot navigation", Neural Networks, vol. 72, pp. 75-87, 2015.

[16] J. Long et al., "Fully convolutional networks for semantic segmentation", Proceedings of the IEEE 2015 Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

[17] H. Noh et al., "Learning deconvolution network for semantic segmentation", The Proceedings of the IEEE 2015 International Conference on Computer Vision, pp. 1520–1528, 2015.

[18] A. Sarica et al., "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review", Frontiers in Aging Neuroscience, vol. 9, n. 329, 2017.

[19] JP. Lerch et al., "Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls", Neurobiol Aging, vol. 29, n. 1, pp. 23-30, 2008.

[20] FJ. Martinez-Murcia et al., "A Spherical brain mapping of MR images for the dectection of Alzheimer's disease", Current Alzheimer Research, vol. 13, n. 5, pp. 575-588, 2016.

[21] P. Kalavathi et al., "Detection of Alzheimer Disease in MR Brain Images using FFCM Method", Computational Methods, Communication Techniques and Informatics, 2017.

[22] W. Cai et al., "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation", Pattern Recognition, vol. 40, n. 3, pp. 825-838, 2007.

[23] P. Sajda, "Machine learning for detection and diagnosis of disease", Annual Review of Biomedical Engineering, vol. 8, pp. 537-65, 2006.

[24] E. A. El-Dahshan et al., "Hybrid intelligent technique for MRI brain images classification", Digital Signal Processing, vol. 20, n. 2, pp. 433-441, 2010.

[25] H.J. Park, "Comparison of distance measures for manifold learning: Application to Alzheimer's brain scans", Journal of the Korean Physical Society, vol. 61, n. 7, pp. 1148–1155, 2012.

[26] Y. Zhang et al., "Detection of Alzheimer's disease and mild cognitive impairment based on structural volumetric MR images using 3D-DWT and WTA-KSVM trained by PSOTVAC", Biomedical Signal Processing and Control, vol. 21, pp. 58-73, 2015.

[27] Y. Zhang et al., "Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning", Frontiers in Computational Neuroscience, vol. 9, n.66, 2015.

[28] D. B. Sam et al., "Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3618-3626, 2018.

[29] C. Zhang et al., "VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images", IEEE Transactions on Geoscience and Remote Sensing, vol. 56, n. 8, pp. 4507 – 4521, 2018.

[30] K.R. Kruthik et al., "CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis", Informatics in Medicine Unlocked, vol. 14, pp. 59-68, 2019.

[31] S. Sabour et al., "Dynamic routing between capsules", Advances in Neural Information Processing Systems, pp. 3856-3866, 2017.

[32] D. E. Worrall et al., "Harmonic Networks: Deep Translation and Rotation Equivariance", IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, pp. 7168-7177, 2017.

[33] J. Liu et al., "Alzheimer's disease classification based on individual hierarchical networks constructed with 3-D texture features", IEEE Trans Nano Bioscience, vol. 16, n. 6, pp. 428-437, 2017.

[34] S. Liu et al., "Early diagnosis of Alzheimer's disease with deep learning", IEEE 11th International Symposium on Biomedical Imaging, pp. 1015-1018, 2014.

[35] E. Pellegrini et al., "Machine learning of neuroimaging to diagnose cognitive impairment and dementia: a systematic review and comparative analysis", arXiv:1804.01961, 2018.

[36] Ö. Çiçek et al., "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation", MICCAI 2016: Medical Image Computing and Computer Assisted Intervention, pp. 424-432, 2016.

[37] O. Ronneberger et al., "U-Net: convolutional networks for biomedical image segmentation", MICCAI 2015. LNCS, n. 9351, pp. 234–241, Springer, Heidelberg, doi: 10.1007/97, 2015.

[38] J. Chang et al., "Brain Tumor Segmentation Based on 3D Unet with Multi-Class Focal Loss", The 2018 International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), DOI: 10.1109/CISP-BMEI.2018.8633056, 2018.

[39] H. R. Roth et al., "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations", The International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 520–527, 2014.

[40] H. R. Roth et al., "A New 2.5D Representation for Lymph Node Detection using Random Sets of Deep Convolutional Neural Network Observations", Medical Image Computing and Computer-Assisted Intervention, vol. 17, n. 1, pp. 520–527, 2014.

[41] Open Access Series of Imaging Studies. [Online] 2010. https://www.oasis-brains.org.

[42] B. A. Ardekani et al., "Corpus callosum shape changes in early Alzheimer's disease: an MRI study using the OASIS brain database", Brain Structure and Function, vol. 219, pp. 343–352, 2014.

[43] A. B. Tufail et al., "Multiclass classification of initial stages of Alzheimer's disease using structural MRI phase images", IEEE International Conference on Control System, Computing and Engineering, pp. 23 – 25, 2012.

[44] J. Dukart et al., "Age correction in dementia matching to a healthy brain", Alzheimer's Disease Neuroimaging Initiative, PLoS One, vol. 6, n.7, 2011.

[45] E. Moradi et al., "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects", Neuroimage, vol. 104, pp. 398–412, 2015.

[46] K. Jarrett et al., "What is the best multi-stage architecture for object recognition", International Conference IEEE Computer Vision, pp. 2146–2153, 2009.

[47] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[48] M. Drozdzal et al., "The Importance of Skip Connections in Biomedical Image Segmentation", Deep Learning and Data Labeling for Medical Applications, pp. 179–187, 2016.

[49] J. Long et al., "Fully convolutional networks for semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440, 2015.

[50] F. Milletari et al., "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation", International Conference on 3D Vision (3DV), 2016.

[51] M. Haubro et al., "Sensitivity and specificity of CT- and MRI-scanning in evaluation of occult fracture of the proximal femur", Injury, vol. 46, n.8, pp. 1557-1561, 2015.

[52] U. A. Hashim et al., "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study", The Lancet, vol. 389, n. 10071, pp. 815-822, 2017.

[53] E. Hosseini-Asl et al., "Alzheimer's Disease Diagnostics by Adaptation of 3D Convolutional Network", IEEE International Conference on Image Processing, 2016.

[54] H. Dong et al., "Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks", MIUA: Medical Image Understanding and Analysis, pp. 506-517, 2017.

Improved Adaptive Semi-Unsupervised Weighted Oversampling using Sparsity Factor for Imbalanced Datasets

Haseeb Ali[1], Mohd Najib Mohd Salleh[2*]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

Kashif Hussain[3]
Institute of Fundamental and Frontier Sciences
University of Electronic Science and Technology of China
Chengdu, China

***Abstract*—With the incredible surge in data volumes, problems associated with data analysis have been increasingly complicated. In data mining algorithms, imbalanced data is a profound problem in machine learning paradigm. It appears due to desperate nature of data in which, one class with a large number of instances presents the majority class, while the other class with only a few instances is known as minority class. The classifier model biases towards the majority class and neglects the minority class which may happen to be the most essential class; resulting into costly misclassification error of minority class in real-world scenarios. Imbalanced data problem is significantly overcome by using re-sampling techniques, in which oversampling techniques are proven to be more effective than undersampling. This study proposes an Improved Adaptive Semi Unsupervised Weighted Oversampling (IA-SUWO) technique with sparsity factor, which efficiently solves between-the-class and within-the-class imbalances problem. Along with avoiding over-generalization, overfitting problems and removing noise from the data, this technique enhances the number of synthetic instances in the minority sub-clusters appropriately. A comprehensive experimental setup is used to evaluate the performance of the proposed approach. The comparative analysis reveals that the IA-SUWO performs better than the existing baseline oversampling techniques.**

***Keywords*—Data mining; imbalanced data; minority; majority; oversampling**

I. Introduction

At present, data mining tasks involve large amount of data which is complex and embedded with noise; hence techniques used for information extraction are needed to be efficient and effective decisions making [1][2]. In data mining, classification is the most commonly performed data analysis task in real-world applications including medical, engineering, and business [3][4][5]; i.e., cancer prediction [6], face detection [7], software fault detection [8][9], bankruptcy prediction [10][11], fraud detection [12]. Majority of classification algorithms consider that the given dataset has the proportional instances among the classes, and these algorithms are not intelligent enough to detect the inappropriate distribution of instances. However, in many practical applications, it is often found that one class known as majority class contains the number of instances to a great extent – intensively dominating the other class which has only a small number of instances; known as minority class. This phenomena is considered as imbalanced data problem [13][14]. Due to imbalanced distribution of instances among the classes, the classifiers show biased behavior to present accuracy according to the majority class while neglecting the essential minority class [15][16]. For example, phishing email dataset actually presents the imbalanced data set in which for each 1 million emails, only 30 emails present the phishing email. For this scenario, the classifier may show unconscious bias towards majority class. If the classifier neglects the minority class, it will result into higher misclassification cost [17][18].

Imbalanced data also have other aspects like desperate distribution of data in the feature space in imbalanced datasets, and these datasets usually have some problematic characteristics as the overlapping of data instances, presence of noise, small disjuncts, and small sized instances [19][20][21]. Also, another kind of data imbalances, which is present with in the class is known as within-class imbalanced problem that result in the performance loss [22][23]. Therefore, it is very difficult for trivial classifiers to predict minority class correctly. In this context, the analysis or prediction accuracy of any classifier for the minority class becomes significantly critical in real world domain which encounter the imbalanced data problem such as, affected vs. non-affected cases in several diseases predictions, non-bankruptcy vs. bankruptcy in bankruptcy predictions, and fraud detection in credit card (fraudulent vs. non fraudulent cases) [1][24][25].

Importance of imbalanced data problem is owned by worldwide researchers hence they proposed many exceptional approaches to tackle this problem. These contributions made for imbalanced data problems can be divided into three categories: data level approach, algorithmic approach, and cost sensitive approach [14][26]. Data level approaches work as the preprocessing of the data before learning process by using resampling techniques which are independent of the classifier [17]. Algorithmic approaches embrace new algorithms or modify the existing ones for imbalanced data problem [27]. Last one is the cost sensitive approach which minimizes the total cost of errors in data level or algorithmic level approach [15]. Data level approaches, also known as external level methods, are more effective in handling the class imbalance problem, as these approaches perform preprocessing of data, whereby data is modified before the learning process [28].

*Corresponding Author..

These methods manipulate the data externally by balancing the distribution of samples among the classes [17][15].

Data level approaches either use the oversampling methods in which artificial data is deliberately generated in the minority class, or the undersampling methods where data is eliminated from the majority class; in order to balance the distribution of minority class. However, removing the data from the majority class may also eliminate the potential data with it, which can be used for the learning process [14]. On the other hand, oversampling methods are preferred more by the researcher's community because there is no risk of losing any useful data. Despite this, in oversampling methods, generating exact replication of data may create overfitting in the data, where selection of incorrect instances in the data for oversampling may generate new instances that might fall in the incorrect region – giving rise to overlapping with other instances that belong to other class, causing overfitting and over-lapping of samples and deteriorating the performance of classifiers [20]. To overcome this, various methods have been proposed in the related literature. The effective approaches are clustering-based approaches in which the input space is partitioned into the clusters then the oversampling technique is applied [29]. In some clustering-based approaches, majority and minority classes are clustered separately; possibly when minority class instances shows small disjoints, so it needs to make several minority sub-clusters [30]. The number of instances vary differently in minority sub-clusters, and it raises within-class imbalance problem because it is necessary to oversample all minority sub-clusters; otherwise the classifier biases towards the oversampled class [23].

In this paper, an improved oversampling technique is proposed, so-called Improved Adaptive Semi-Unsupervised Weighted Oversampling (IA-SUWO) with sparsity factor for class imbalance problem. IA-SUWO clusters the minority class instances and assigns higher weights to the minority instances which are closer to majority instances, in order to manage hard-to-learn minority instances. Secondly, it also considers every minority sub-cluster for oversampling along with small concepts which are far from the majority clusters and ignored by several other techniques (Fig. 1). The proposed approach avoids over-lapping between the synthetic minority instances and majority instances by using the semi-unsupervised clustering approach that significantly avoids majority class clusters to come in-between two minority clusters that need to be merged. It assigns weights to minority instances for oversampling according to average Euclidean distance of minority instances from the instances of majority class, in addition to decrease more chances of over-generalization. Moreover, it also assigns weights according to the sparsity factor of the minority instances in each sub-cluster to enhance the learnability of the classifier for the minority class instances which are sparse apart. The IA-SUWO technique identifies the sub-clusters misclassification error and assigns sizes of sub-cluster appropriately based on their complexity in being misclassified. In order to validate the proposed IA-SUWO technique, a comprehensive experimental setup is performed. Three publicly available datasets are used to evaluate the proposed method after the classification process on the two classifiers. Precision, F-measure, and ROC are used as the performance measures. Outcome of this whole experimental setup is compared to state of the art with four other existing techniques.

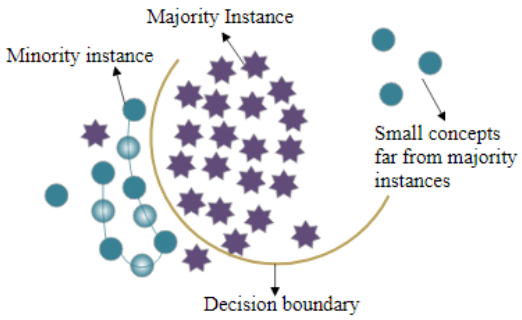


Fig. 1. Small Concepts of the Minority Class that are Far from the Majority Class are neglected and not oversampled.

The remaining paper is organized as follows: Section II presents the related works from recent literature. The detail of the proposed technique is given in Section III, while the research methodology adopted for this study is explained in Section IV. Section V provides the experimental results for analysis and discussion. The study is duly concluded in Section VI which also highlights potential future directions pertaining to relative research line.

II. Related Work

As discussed earlier in this paper, the data level approaches are effective in balancing the data distributions. Resampling methods used for the preprocessing of data can be categorized into two major types: oversampling and undersampling. This current research is intended to focus on the oversampling techniques (Fig. 2) for imbalanced data problem, therefore a brief overview of existing oversampling techniques, proposed in recent literature, is presented.

Oversampling methods are further categorized into random and informed methods [14]. Random Oversampling (ROS) is the pioneer and simplest technique used for oversampling. This technique randomly generates synthetic instances until the desired ratio. However, in spite of its ease in implementation and its simplicity, it encounters a major drawback that it generates exact replication of the original minority instances, which often results in over fitting [26] [31]. Addressing this problem, Nitesh Chawla proposed first informative method proposed for imbalanced data problem in 2002, named as Synthetic Minority Oversampling Technique (SMOTE) [26]. In which synthetic minority instances are generated by linear interpolation of two neighboring instances. SMOTE generates minority instances between randomly selected minority instance and its nearest neighbor. SMOTE produces new minority instances with the same primary number of original instances. However, some of these generated new instances fall into the incorrect region and overlapped with the instances of the other class result into over generalization [12] (Fig. 2(i)).

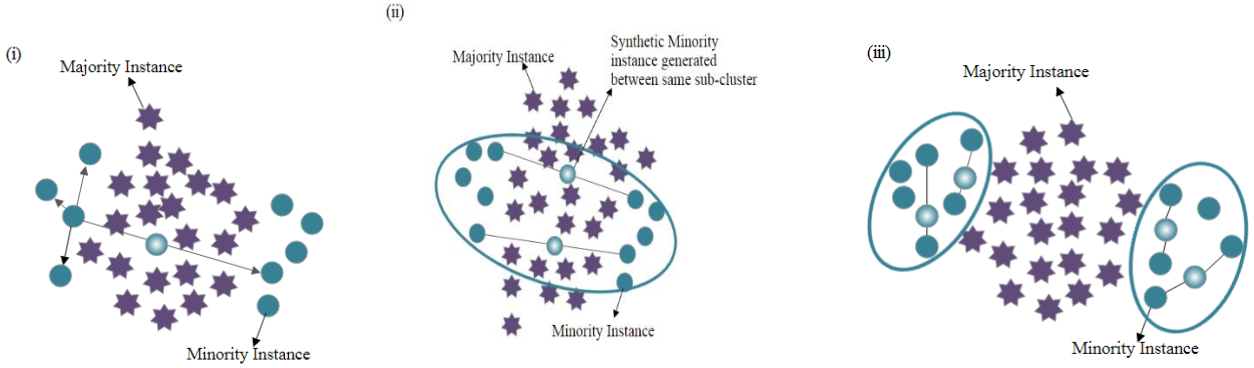


Fig. 2. Presenting the different Oversampling Scenarios in which, (i) Synthetic Instances Generated between Selected Instance and its 4 Nearest Neighbors; (ii) Synthetic Instances are generated between Selected Instance and its Nearest Neighbors from the Same Cluster; (iii) Synthetic Instances are generated between Selected Instance and its 4 Nearest Neighbors that belong to the Same Cluster, which Avoids Overlapping.

Informed methods identify the most effective areas for the oversampling like, Safe-level SMOTE [32] is a modified SMOTE algorithm which identifies the safe level value for each minority instance and applies weight degree. It oversamples in only safe areas by differentiating between noisy and safe instances. There are some other approaches which focus on the class regions for oversampling, as Borderline-SMOTE [33] which determines borderline among the two classes then generates synthetic minority instances near the decision boundary. This technique chooses the targeted instances which are close to the borderline. Considering class regions, Cluster-SMOTE [34] is another approach which focuses on the certain regions for oversampling. In this regard, it partitions the data into clusters. This technique first uses the *k*-means to cluster minority class then it employs SMOTE within those naturally occurring minority clusters. Moreover, unlike SMOTE which produces new instances equal in number to original instances. Techniques like ADASYN [35] and its variation KernalADASYN [36] are proposed with objective to choose hard-to-learn instances. By employing the weight assigning approach, it assigns more weights to these instances for oversampling. The basic reason behind this approach is to priorities the minority instances to avoid their misclassification and to enhance the number of synthetic instances to be generated [37].

Majority Weighted Minority Oversampling technique (MWMOTE) [37] uses clustering approach, which partitions the majority and minority clusters separately and then assign weights to minority instances according to the Euclidean distance of the minority instances from majority instances. However, it neglects the small disjoints which are far from the majority instances (Fig. 1). Presence of small disjoints of minority class results in within-class imbalance problem. It is necessary to oversample all the minority sub-cluster; otherwise, the classifier biases towards the oversampled ones. Adaptive Semi Unsupervised Weighted Oversampling (A-SUWO) [23] which uses semi-unsupervised clustering approach to cluster the minority class instances to avoid over-generalization. Secondly, in order to oversample all the minority sub-cluster including those small concepts that are far from the majority class, this technique measures the misclassification error rate to determine the complexity of each minority sub-cluster in being misclassified. Later, it assigns larger size to those sub-clusters which have higher misclassification error rate. Enhanced Minority Oversampling Technique (EMOTE) [38] enhances the minority class distribution by generating new instances in their neighborhood in order to improve the classifier performance. It effectively improves the classification results by tuning the wrongly classified instances into correctly classified instances by using its proposed oversampling approach. For imbalance learning, an Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) method [39] creates an ensemble by combining an evolutionary algorithm (EA) with a new clustering based synthetic data generation method. In this method, regions for oversampling of minority instances are identified by using the clustering approach based on the modern ideas. The EA benefits in lowering the overall computational cost and in optimization of parameters for data generation method. Self-Organizing Map-based Oversampling (SOMO) [29] is another technique that uses self-organizing map to convert input data into two dimensional space. It generates synthetic instances into effective areas. It also uses SMOTE to generate synthetic minority instances into clusters which are found in the lower dimensional space. SOMO alleviates between class and within-class imbalances problem. *k*-means SMOTE [17] is another technique which uses density factor for the data generation. After clustering the input data and finding sparsity factor among all clusters, it generates synthetic samples by using SMOTE according to the weight assigned based on the density of the clusters.

A Radial-Based Oversampling (RBO) [19] tackles the noisy imbalanced data classification problem. This method generates minority data instances into the rightful regions according to their imbalance ratio calculated by radial based functions, also removes noise from the data effectively. A robust oversampling technique, proposed in [40], for imbalanced data learning which uses Gaussian Mixture Model (GMM) to balance the distribution of instances in both classes. This method considers high dimensional feature space for generating synthetic instances and GMM determines and filters out outlier instances from the minority class. An exclusive technique for ordinal regression imbalanced problem also uses oversampling approach for data generation in minority class based on the weights assigned. This synthetic minority oversampling for ordinal regression (SMOR) considers

generation direction for each candidate [41]. Another oversampling technique for imbalances problem in ordinal regression is proposed which is adaptive structure based. It enhances the data generation process in minority class after exploration of immature and complex minority instances [42].

III. Improved Adaptive Semi-Unsupervised Weighted Oversampling (IA-SUWO)

This research proposes an improved oversampling technique, so-called IA-SUWO. The proposed technique significantly enhances the learnability of the classifier and improves its accuracy by using sparsity factor for assigning weights to overcome the limitation of the conventional A-SUWO. Standard A-SUWO assigns weights to instances for oversampling according to their Euclidean distance from majority samples (Fig. 3). Here, weights are assigned to those instances which are closer to majority instances or in front of majority clusters. However, it neglects the instances that are far from majority instances or located behind in the minority sub-clusters. Minority instances contain potential data about minority class or could be neglected by the classifier. Therefore, if weights are assigned according to sparsity of each minority cluster, it can assign weights appropriately. Sparsity factor finds the sparse minority instances in each cluster by measuring its density. Synthetic instances are generated according to both approaches of assigning weights to significantly alleviate the between-the class and within-class imbalance problem.

A. Sparsity Factor for Oversampling in IA-SUWO

Sparsity factor in a cluster can be defined as the measure of sparse instances that are found to be scattered in a particular cluster or where density of instances in the cluster is low. This research improves the conventional A-SUWO by assigning weights to minority instances according to sparsity factor for oversampling. Those minority instances present in the minority clusters which are sparse apart and far from the majority instances assumed as incompetent but certainly they are not. Hence, after assigning weights based on their Euclidean distance from majority instances, the proposed model looks forward for the density of each cluster to measure the sparsity factor. Density of each cluster is measured by dividing the number of minority instances present in each sub-cluster by the average distance between them raised to the power of features count *m*. Density is inversely proportional to the sparsity; therefore, sparsity factor is measured as the inverse of density. After measuring sparsity for each filtered minority cluster, sum of all the sparsity measures is taken, which is then transformed into the sampling weights. These sampling weights are assigned to instances of all minority filtered cluster which determine the expedient distribution of instances to be generated in each cluster. In this way, instances which present lower density measure attain more weights for oversampling. It is noted that these sampling weights are based on the comparative analysis of density measure of each cluster, which determines the cluster density as compared to others on average. It considers only the distance between the minority instances while measuring the density of minority clusters.

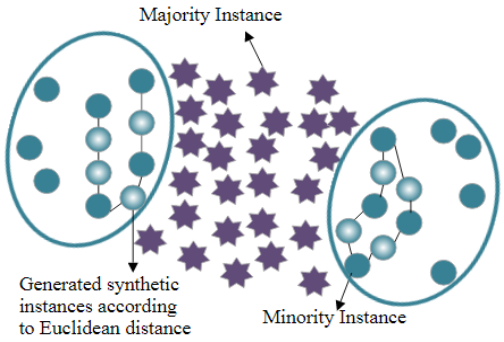


Fig. 3. Synthetic Instances are Generated According to the Euclidean Distance from Majority Instances.

B. Semi-unsupervised Clustering

Before clustering, this technique finds the noisy instances and removes them from the dataset based on the method proposed by [33]. This is done in order to identify if any instance is noisy, and also to determine its *NN* nearest neighbors. If its neighbors belong to another class, this instance is declared to be noise hence removed from the dataset. After removing the noise from dataset, it uses hierarchical clustering to cluster the majority class first, that results into *m* majority sub-clusters $Cmaj_{i=1,\dots,m}$. Later, all the minority sub-clusters are allocated to minority sub-clusters $Cmin_{j=1,\dots,n}$ based on the semi unsupervised hierarchical clustering proposed by [23] to avoid overlapped generated minority instances. This algorithm is based on the complete-linkage agglomerative hierarchical clustering, but it checks overlapping in each iteration between the two nominated minority sub-clusters say $Cmin_a$ and $Cmin_b$ selected for merging [43]; because this algorithm is modified as it uses information about majority class sub-clusters. It is therefore, the proposed approach is not fully unsupervised but semi-unsupervised. Here, the two closest minority sub-clusters are not allowed to be merged if any majority class sub-cluster exists between them. While in the case when there is not any majority sub-cluster between these two minority sub-clusters, and their distance is also less than the pre-defined threshold, they are allowed to be merged.

This approach avoids overlapping of synthetic minority instances with majority instances significantly by avoiding majority cluster to come in between. The semi-unsupervised hierarchical clustering algorithm has the following steps:

- Each minority instance is assigned to an individual sub-cluster which results into *n* sub-clusters of minority class of size B = $Cmin_{j=1,\dots,n}$.
- Two minority sub-clusters that are identified can be $Cmin_a$ and $Cmin_b$ for merging with lowest Euclidean distance л.
- Determine the majority sub-clusters $C_{maj_i} \in A$, (where A belongs to a set of majority class) which have Euclidean distance between $Cmin_a$ and $Cmin_b$ lesser than π.

- Majority sub-cluster exists between $Cmin_a$ and $Cmin_b$ if this $A \neq \emptyset$; in this case these minority clusters should not be merged. In order to avoid these sub-clusters to be considered for merging again, distance between $Cmin_a$ and $Cmin_b$ is set to a large number.
- Otherwise, these minority sub-clusters are merged and become a new sub-cluster $Cmin_c$.
- These steps are again repeated for every newly created sub-cluster $Cmin_c$, and these steps are repeated unless the Euclidean distance between any two closest sub-clusters of minority class becomes lesser than a threshold *T*. Finally, in this way *n* Minority sub-clusters are formed.

In order to find a better estimated value for *T*, this algorithm measures Euclidean distance d_{med} among all the *h* instances of majority and minority class by using Eq. (1).

$$T = d_{avg} \times c_{thres} \quad (1)$$

where c_{thres} is a user-defined constant parameter and its optimum value depends on the dataset. The greater value of c_{thres} results into larger cluster size which can result in increasing the chance of over-lapping, while the smaller value for c_{thres} may result in small cluster sizes which can cause overfitting or less diverse synthetic samples generation. Optimum value for c_{thres}, is in the range of [0.7, 2.0]. It depends on the dataset, for examples, for wine dataset the best c_{thres} value is 1.0.

C. Adaptively Sub-Cluster Sizing

Mostly, all sub-clusters present in dataset have similar sizes after the oversampling by using the existing clustering-based techniques. However, the sub-clusters that have higher chances of misclassification like those small concepts that are far from the majority class, need larger sizes and more oversampling. Though, some sub-clusters have lower chance of misclassification and do not need much oversampling. For this scenario, A-SUWO resets the sizes of all sub-clusters according to their misclassification rate, which means the misclassification of the related instances results into two achievements: firstly, larger size is assigned to sub-cluster which is prone to more miss-classified, and secondly, it balances the ratio of samples in both classes. This is done by using cross validation in Linear Discriminant Analysis (LDA), which calculates the complexity or misclassification rate of every sub-cluster. For the classification, each of the *n* minority sub-cluster is partitioned into *k* similar sized divisions Fig. 4.

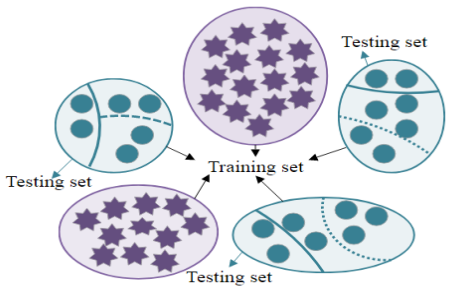


Fig. 4. Split Minority Sub-Clusters into Testing and Training Partitions.

All majority instances from all sub-clusters and *k-1* divisions from each minority sub-cluster is used for training purpose, whereas all majority sub-clusters and one division that is remained in each minority sub-cluster are used for testing purpose. LDA is a simple classifier and it needs no parameters to be tuned, it runs *k* times for cross validation and measures the misclassification error rate. The misclassification error ε_{jk} is calculated in each fold *f*, for each minority sub-cluster *j* for the number of instances that are wrongly classified as the majority instance while testing. This misclassification error ε_{jk} is divided by total number of instances present in each sub-cluster R_j to calculate error rate ε_{jk}. By taking average of all error rates, this result into average error rate $\bar{\varepsilon}_{jk}$ of all minority sub-clusters for all folds.

Using the following equation the standardized average error rate $\hat{\varepsilon}_j$ is calculated by standardizing $\bar{\varepsilon}_{jk}$.

$$\hat{\varepsilon}_j = \frac{\bar{\varepsilon}_j}{\sum_{j=1}^{n} \bar{\varepsilon}_j} \quad (2)$$

In order to get the final sizes of any two minority sub-clusters say L_1 and L_2 should have equivalent ratio to their average error rates $\hat{\varepsilon}_{L_1}$ and $\hat{\varepsilon}_{L_2}$ as Eq. (3):

$$\frac{S_{L_1}}{S_{L_2}} = \frac{\hat{\varepsilon}_{L_1}}{\hat{\varepsilon}_{L_2}} \forall\, L_1, L_2 \in \{1, \dots, n\} \quad (3)$$

Here, for the final sizes of minority sub-clusters L_1 and L_2, S_{L_1} and S_{L_2} are the final sizes after oversampling respectively. While, for the L_1 and L_2, the standardized average error rates are $\hat{\varepsilon}_{L_1}$ and $\hat{\varepsilon}_{L_2}$, respectively.

D. Synthetic Instance Generation

Before A-SUWO, there were many existing techniques which generate synthetic instances in safe areas, borderline, and in those clusters that are near the majority class. Likewise, in the MWMOTE, there might be some sub-clusters that are distant from the instances of majority class and neglected completely, which means they are not oversampled. This causes within-class imbalance problem and classifier becomes biased towards oversampled ones. Some of the techniques that generate synthetic instances between the candidate instance and its nearest neighbor which belongs to another sub-cluster, produce the overlapping of generated instance with majority instances (Fig. 2(i)). Similarly, generating the instances between candidate instance and its nearest neighbor from the same sub-cluster but far from it, also results into overlapping (Fig. 2(ii)). A-SUWO overcomes these problems by generating the synthetic instances between the primary instance and its nearest neighbor within the same sub-cluster to avoid overlapping of instances (Fig. 2(iii)). Secondly, it oversamples all the minority sub-clusters (smaller ones) that are far from the majority instance which reduce the within-class imbalance problem. However, A-SUWO generates instance using Euclidean distance of minority instances from the majority instances and assigns weights to those instances of minority sub-clusters that are nearer to majority instances. This causes the vacant space and ignorance of those instances that are far away from the majority instances or at the back within the same sub-cluster, even if they carry the important information about the minority class. IA-SUWO improves this method by

using the sparsity factor and assigns weights adaptively to all instances within each individual minority sub-cluster that results into the appropriate number of instances generated within each sub-cluster and oversamples every sub-cluster to its required extent (Fig. 5).

For oversampling, probability distribution of instances is derived from the weight assign to instances of the minority class. This weight is assigned according to two approaches: according to the Euclidean distance from majority samples and sparsity factor. At first, weights are assigned according to the Euclidean distance from the majority instances. Purpose of assigning weights according to the Euclidean distance is that, the minority instances which are found to be more nearer to decision boundary or majority instances have higher chances of being misclassified. Therefore, in order to assign weight to minority instances in the minority sub-cluster C_{min}, find the k nearest neighbor for the h^{th} minority instance x_{jh}, according to its Euclidean distance from the majority instance $y_{jh(v)}$, and measure this distance $d(x_{jh}, y_{jh(v)})$, where this $v = 1 \dots, k$ which implies the indices of the nearest neighbors. This distance $d(x_{jh}, y_{jh(v)})$ is normalized by dividing it by the total number of features D to make it robust to the datasets used with several number of features, see Eq. (4):

$$\hat{d}(x_{jh}, y_{jh(v)}) = \frac{d(x_{jh}, y_{jh(v)})}{D} \quad (4)$$

Then, define the closeness factor as $\Gamma(x_{jh}, y_{jh(v)})$ between x_{jh} and $y_{jh(v)}$ using Eq. (5):

$$\Gamma(x_{jh}, y_{jh(v)}) = f_i\left(\frac{1}{\hat{d}(x_{jh}, y_{jh(v)})}\right) \quad (5)$$

where f_j is defined as a cutoff function, which is used for avoiding $\frac{1}{\hat{d}(x_{jh}, y_{jh(v)})}$ to become extremely large in the situation when two instances x_{jh} and $y_{jh(v)}$ in a sub-cluster C_j become quietly close to each other. Hence, f_j can be expressed as Eq. (6):

$$f_j(x) = \begin{cases} x \; if, x \le TH_i \\ TH_j, otherwise \end{cases} \quad (6)$$

Here, the largest value f_j (x) can achieve TH_j, whereas TH_j is automatically determined for every sub-cluster C_j. This is gained by measuring the Euclidean distance of all minority instances x_{jh} in every sub-cluster to their nearest majority instance $y_{jh(1)}$ and after this identify $f\left(\frac{1}{\hat{d}(x_{jh}, y_{jh(1)})}\right)$, TH_j is set as average of $f\left(\frac{1}{\hat{d}(x_{jh}, y_{jh(1)})}\right)$, using Eq. (7):

$$TH_j = \sum_{j=1}^{R_j} f\left(\frac{1}{\hat{d}(x_{jh}, y_{jh(1)})}\right) \quad (7)$$

where R_j present the number of instances in sub-cluster C_j. As aforementioned TH_j is determined automatically, but this is really critical, because the weighting algorithm runs separately for each sub-cluster and a specific threshold is required to each sub-cluster.

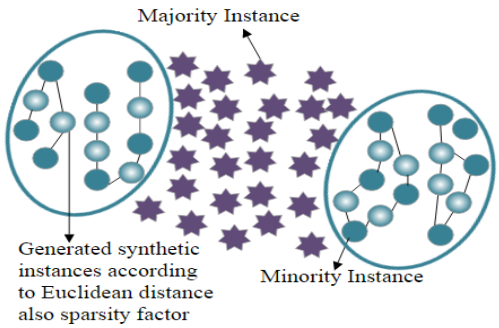


Fig. 5. Synthetic Instances are generated According to Euclidean Distance and Sparsity Factor.

In order to assign higher weights to the instances that are closer to the majority instances, Eq. (5) present the reciprocal of $\Gamma(x_{jh}, y_{jh(v)})$. At last, the weights $W(x_{jh})$ are decided according to the Euclidean distance of minority instance (x_{ij}) from the all nearest neighbors, using Eq. (8):

$$W(x_{jh}) = \sum_{v=1}^{k} \Gamma(x_{jh}, y_{jh(v)}) \quad (8)$$

This weight is transformed into the probability distribution for oversampling along with the weights assigned according to the proposed approach that is based on the sparsity factor. As the sparsity factor is proposed in this paper for assigning weights to the minority instances. For determining the sparsity factor, the proposed approach measures the density of each minority sub-cluster, because inverting the density results in the sparsity measure. Therefore, density is measured as we first determine the mean Euclidean distance among all the minority instances, which is measures by finding the distance between each minority instance present in that sub-cluster. This distance is used for measuring the density as Eq. (9):

$$density(f) = \frac{minority\ count(f)}{average\ minoirty\ distance(f)^m} \quad (9)$$

where density is inversely proportional to the sparsity hence, inverse of density will be equal to the sparsity factor that can be measured as Eq. (9):

$$sparsity(f) = \frac{1}{density(f)} \quad (10)$$

After measuring sparsity for each filtered minority cluster, we can take sum of all the sparsity measures and this sparsity sum is then transformed into the sampling weights as Eq. (11):

$$Sparsity\ sum = \sum_{f \in filteredcluster} sparisty\ factor(f) \quad (11)$$

As the sparsity sum is calculated which is then used in weight formula as Eq. (12):

$$W(f) = \frac{sparsity\ factor(f)}{sparsity\ sum} \quad (12)$$

Consequently, these sampling weights are assigned to instances of all minority filtered cluster which determine the expedient distribution of instances to be generated in each cluster. In this way, instances which present lower density measure attain more weights for oversampling. It is noted that these sampling weights are based on the comparative analysis of density measure of each cluster, which determines the

density of a cluster as compared to others on average. It considers only the distance between the minority instances while measures the density of minority clusters.

To get the probability distribution $P(X_{jh})$, we sum that the both weights are as Eq. (13):

$$W(X) = W(f) + W(jh) \tag{13}$$

At last, these weights are transformed in the probability distribution $P(X_{jh})$ of synthetic samples by using Eq. (14):

$$P(x_{jh}) = \frac{W(X)}{\sum_{h=1}^{r_j} W(x_{jh})} \tag{14}$$

Finally, in order to obtain the each cluster size up to S_j, each sub-cluster C_j, $j = 1, \dots, n$ is oversampled according to the probability distribution of weights assigned by both approaches to minority instances. For this purpose, an instance 'a' is selected randomly from the probability distribution in any particular sub-cluster, then instance 'b' from its nearest neighbor is selected which belongs to that same sub-cluster and a new instance 'c' is generated between these selected instances *a*, and *b* like as Eq. (15):

$$c = \beta a + (1 - \beta)b \tag{15}$$

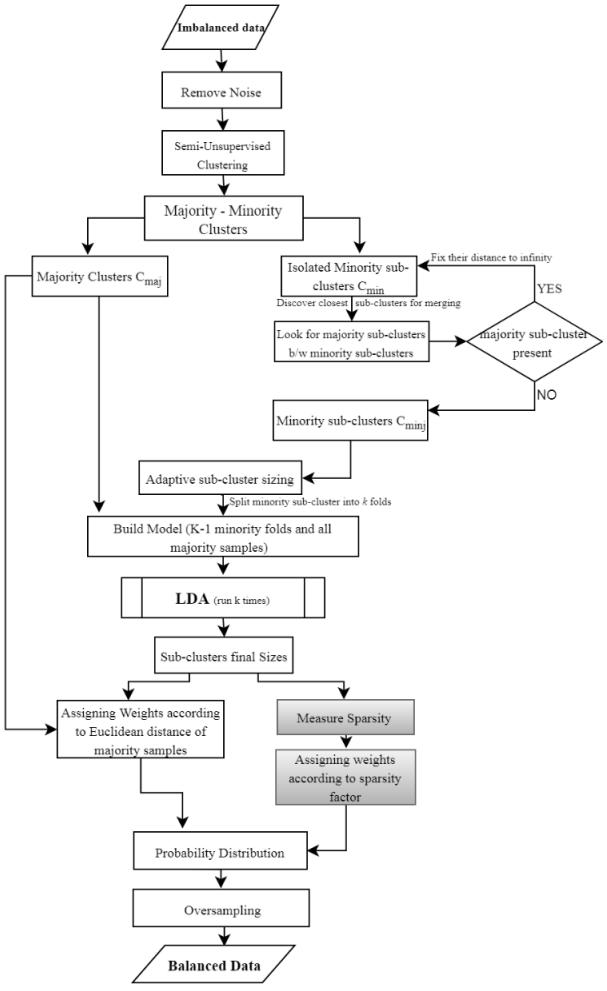


Fig. 6. Flow Diagram of the Proposed Method.

where, β is a random number between [0, 1]. Synthetic instances are generated in each minority sub-cluster, these instances are generated among the instances which are closer to the majority instances only. Proposed technique is summarized in the flow diagram shown by Fig. 6.

IV. Research Methodology

At first, for developing the proposed model, datasets are used for the experimental purpose to evaluate the performance of the proposed model. For performance evaluation of the proposed oversampling technique, some standard benchmark datasets are used that contain binary or multi-classes in which, number of samples in the classes varied exceptionally. Those datasets which have more than two classes are converted into binary-class dataset, labeling the smallest class as minority class and a bigger one as majority class. These datasets are examined for selection on the basis of the number of features, the total number of samples in both classes, number of minority samples, number of majority samples and their imbalance ratio.

IA-SUWO Algorithm

1. Semi-unsupervised clustering
   i. Noisy samples eliminated from the dataset.
   ii. Evaluate *T* by using Eq. (1).
   iii. Make majority class *m* sub-clusters $Cmaj_{i=1,\dots,m}$.
   iv. Each minority sample is assigned to an isolated sub-cluster.
   v. Discover the closest sub-clusters $Cmin_a$ and $Cmin_b$.
   vi. Look for any majority sub-cluster if overlapping between $Cmin_a$ and $Cmin_b$.
   vii. If any majority sub-clusters exist, set the distance equal to infinity and go to step (1.v). Otherwise, combine $Cmin_a$ and $Cmin_b$ and form one new sub-cluster $Cmin_c$.
   viii. Repeat steps 1.v to 1.vii until Euclidean distance among the closest sub-cluster is less than *T* threshold.
2. Adaptive sub-cluster sizing
   i. Split randomly every sub-cluster of minority class into *k* divisions.
   ii. Use *k*-1 divisions from each minority sub-cluster as a training set, and construct a model in addition to all majority instances.
   iii. Use one portion which is remained in each minority sub-cluster for testing of the model.
   iv. Evaluate Average Minority Standardized Error Rate $\hat{\varepsilon}_j$.
   v. Repeat *k* times, steps (2.ii) to (2.iv).
   vi. Identify the final sizes S_j for all sub-clusters $Cmin_{j=1,\dots n}$ by using Eq's. (2) and (3).
3. Assigning weights for synthetic samples generation within each minority sub-cluster.
   (a) By determining the Euclidean distance from majority samples for each sub-cluster $j = 1,2, \dots, n$:
      i. Find nearest neighbors *NN* surrounded by majority samples, for all the minority samples X_{jh} in sub-cluster $Cmin_j$.
      ii. Evaluate $W(X_{jh})$ by estimating TH_j, for each minority sample in $Cmin_j$ by using Eqs. (4) to (8).
   (b) By using the sparsity factor, for each cluster, compute sampling weights based on each sub-clusters density.
      i. For each filtered cluster *f*, find average distance (f) ← mean (Euclidean distances (*f*)) between the minority samples.
      ii. Determine the density measure by using Eq. (9).
      iii. Get the measure of sparsity by inverting the density measure using Eq. (10).
      iv. Determine the sampling weight of each sub-cluster as the sparsity factor of sub-cluster divided by the sum of all sub-cluster sparsity measures using Eqs. (11) and (12).
      v. Convert these weights for probability distribution $P(X_{jh})$ by Eqs. (13) and (14).
4. Oversample minority instances: First initialize the $O = I$, then for each sub-cluster $j = 1,2, \dots, n$:
   i. Select a minority instance 'a' in sub-cluster j from $P(X_{jh})$ for oversampling.
   ii. Randomly select one of its nearest neighbor *NN* 'b' from the same sub-cluster.
   iii. Produce a new synthetic instance *'c'* using Eq. (15) between *'a'* and *'b'*, and add 'c' to set O.
   iv. Repeat steps from (4.i) to (4.iv) until the sub-cluster size reaches S_j.

A. Datasets

This research used three main datasets for the experimental setup namely IRIS, WINE, and GLASS. These three datasets are commonly used by every researcher for the evaluation of their proposed approaches for imbalanced data problem. There are many benchmark datasets present in UCI directory, but these datasets are more simple and convenient to evaluate the performance of proposed model, as well as, better judgement of their impact on the dataset and learning process. Brief description of these datasets is as follows.

IRIS: This dataset is created by R. A. Fisher, a famous dataset in pattern recognition and classification. There are 150 instances with 4 attributes and with 3 classifications in this dataset. The classification of Iris dataset involves the classification of data like petal length, petal width, sepal length, and sepal width into three classes of species: Iris Versicolor, Iris Sentosa, and Iris Verginica.

This dataset is transformed into the binary class dataset in which one class contains 50 instances that is minority class, while the second class contains 100 instances which present the majority class. For this experiment, total 100 instances are randomly taken from minority and majority instances in the whole dataset for training purpose and 50 random instances for testing. After applying the proposed technique and other standard oversampling techniques on this dataset, the performance of these techniques are evaluated.

WINE: This dataset is created by C. Blake used for examining the classifier performance. It is used for comparison of classifiers with high dimensional settings. It contains a total of 178 instances and 13 attributes which have been classified into three main classes, where each attribute is continuous. In class distribution, the number of instances in the three classes are 59, 71, and 48, respectively. For this experimental setup, this dataset is transformed into the binary class dataset in which one class (minority) has 71 instances and the second class (majority) contains 107 instances. The proposed technique and other oversampling techniques are applied to this dataset then the performance of these techniques are evaluated.

GLASS: This dataset is collected by B. German on the fragments of glass confronted in the forensic work. The study of the classification of glass types drives in the criminological investigation. At the scene of the crime, the glass left can be used as evidence. This dataset is consisted of 9 attributes and 6 classes. There are 163 instances of window glass including 4 subclasses, and 51 instances of Non-window glass with 3 subclasses. Thus, there are a total of 7 classes of a glass. The input attributes of this dataset include Refractive Index, Aluminum, Barium, Calcium, Iron, Magnesium, Potassium, Silicon and Sodium. For this experimental setup, this dataset is transformed into the binary class dataset in which one class (minority) has 51 instances and the other class (majority) contains 163 instances. The proposed technique and other oversampling techniques are applied to this dataset. The description of these datasets is summarized in Table I.

B. Experimental Setup

In this experimental process, the proposed IA-SUWO technique and four other oversampling techniques are applied to these datasets which output the new oversampled datasets. This research is implemented on MATLAB r2015a, on a workstation with 64 bit operating system, 4 GB RAM and 2.6 GHz CPU. Since, every technique generates different number of synthetic minority samples in each dataset, therefore 4-fold stratified cross-validation is used for determining the mean and standard deviation of the oversampling methods performance measures. To reduce the randomness effect of the results, the average of results taken by repeating each experiment three times. The selected datasets are remedied for the imbalanced data problem after being oversampled.

The outcome of the proposed technique and contribution of this research is validated by comparing its oversampling results with the standard technique. It is evaluated by measuring the extent of synthetic samples generated and the reduced imbalanced ratio achieved by both techniques on this dataset. However, to evaluate the performance of each oversampling technique and the correctness of the synthetic samples generated by oversampling techniques, it is necessary to find the classification accuracy of these oversampled datasets. For the classification of these datasets, this research used four classifiers known as; Naïve Bayer, K-Nearest Neighbor (KNN), Logistic regression, and Neural Network. These classifiers are trained and tested on these datasets using 70-30% and 50-50% training and testing ratios, and also classified by 10-Fold cross-validation. These results are measured by the performance evaluations metrics, Precision, F-measure, and ROC; which are compared with standard A-SUWO and other oversampling techniques namely, SMOTE, Borderline-SMOTE, and Safe Level-SMOTE.

C. Performance Measures

Precision, F-measure, and ROC are the performance measures chosen by this research for the evaluation of proposed IA-SUWO. As these are the common measures used by research community for the imbalanced data domain. Precision measures the exactness of the classifier that the number of samples of minority class, which are labeled as positive are actually positive.

TABLE. I. DATASET DESCRIPTION

No.	Dataset	Minority Class	Majority Class	No. of feature	No. of Instances	No. of Minority samples	No. of Majority Samples	Imbalanced Ratio
1	Iris	1	All others	4	150	50	100	1:2.0
2	Wine	-1	All others	13	178	71	107	1:1.83
3	Glass	-1	All others	9	214	51	163	1:3.20

F-measure is calculated by precision and recall, where recall measure the completeness of the classifier. And, the relative importance between precision and recall is adjusted by F-measure. Receiver Operating Characteristics (ROC) is an important factor for evaluating the classification model performance, as it determines that how much the model is capable of distinguishing between the classes.

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

$$F_{measure} = \frac{(1+\beta^2)*Recall*Precision}{\beta^2*Recall*Precision} \quad (18)$$

Receiver Operating Characteristics (ROC) is also an important evaluation matrix for determining the classification model performance, as it tells that how much the model is capable of distinguishing between the classes. ROC is acquired by scheming the true positive rate over false positive rate.

$$TPR = \frac{TP}{TP+FN} \; FPR = \frac{FP}{FP+TN} \quad (19)$$

V. Results and Discussion

Performance of proposed technique for the validation of contributed part is evaluated by measuring the extent of oversampling, which means the number of synthetic samples generated and the reduced imbalances ratio achieved by the standard and improved technique on the selected datasets. These results are shown in Table II and discussed as below.

Table II shows the results before and after the oversampling technique applied on all datasets. These results demonstrate that the IA-SUWO generated more synthetic samples than the standard technique using the sparsity factor. The difference in the number of samples generated after oversampling by both techniques indicates that these samples are correctly generated in minority clusters, where the sample are sparse apart and the density was low, as these clusters required more oversampling. The exactness of the generated samples can be evaluated by the classification of the datasets. IA-SUWO generated optimum number of synthetic samples among the minority sub-cluster, as the model did not generate samples randomly or incredibly large number of samples which can cause overfitting which significantly deteriorates the performance of the classifier. The imbalanced ratio given in Table II, after the oversampling, is more reduced by the purposed model than the standard technique. The ideal and best value for imbalanced ratio is 1:1 for both classes, and the value closer to this ratio respectively and the maximum value of imbalanced ratio should not be more than standard technique.

Table III presents the mean results of classification obtained from two classifiers according to 70-30%, 50-50% testing and training ratios, and 10-fold cross-validation for the proposed methods and other oversampling techniques. The best measure is presented in bold. IA-SUWO obtained best results according to at least one measure in all the datasets. Table IV Neural Network classifier shows the best results most of all the measures for both ratios of testing and training; also 10-fold cross validation for proposed IA-SUWO. For Iris and Wine datasets, IA-SUWO presented the best results in all measures by using 50-50% testing and training ratios. For Glass, IA-SUWO achieved best results in all measure using 70-30% testing and training ratios. KNN classifier gave the best results for IA-SUWO using 70-30% testing and training ratio on the Iris dataset in all measures and 10-fold cross validation on the glass dataset.

IA-SUWO presented overall good results on all the datasets. These results declare that the IA-SUWO generated useful synthetic instances by using the sparsity factor which determined the space where the oversampling is needed the most that can be helpful for the learnability of the minority class. In this way, it also enhanced the reliability on the classification results by using any classifier on the datasets which are oversampled by using IA-SUWO.

TABLE. II. Comparative Oversampling Results for Improved and Standard A-SUWO

Dataset	Technique	#. of majority instances	#. of minority instances before oversampling	Imbalanced Ratio	#. of minority instances after oversampling	Imbalanced Ratio
IRIS	IA-SUWO A-SUWO	100 100	50 50	1:2.0 1:2.0	**85** 80	**1:1.17** 1:1.25
WINE	IA-SUWO A-SUWO	107 107	71 71	1:1.50 1:1.50	**96** 90	**1:1.11** 1:1.18
GLASS	IA-SUWO A-SUWO	163 163	51 51	1:3.2 1:3.2	**125** 120	**1:1.30** 1:1.35

TABLE. III. CLASSIFICATION RESULTS FOR OVERSAMPLING TECHNIQUES ON 3 DATASETS USING KNN

Dataset	Testing-Training	Measure	SMOTE	Borderline-SMOTE	Safe-level SMOTE	A-SUWO	IA-SUWO
IRIS	70-30%	Precision	0.898	0.738	0.832	0.911	**0.934**
		F-measure	0.877	0.668	0.824	0.908	**0.923**
		ROC	0.879	0.895	0.898	0.929	**0.941**
	50-50%	Precision	0.891	0.549	0.882	**0.958**	0.923
		F-measure	0.954	0.545	0.877	**0.967**	0.908
		ROC	**0.976**	0.623	0.893	0.974	0.934
	10-Fold	Precision	0.901	0.708	0.921	**0.971**	0.963
		F-measure	0.898	0.706	0.932	**0.969**	0.962
		ROC	0.976	0.740	0.957	0.970	**0.986**
WINE	70-30%	Precision	**0.986**	0.978	0.958	0.977	**0.987**
		F-measure	0.986	0.977	0.954	0.976	0.974
		ROC	0.982	0.980	0.960	0.975	**0.998**
	50-50%	Precision	0.972	0.944	0.930	**0.973**	0.957
		F-measure	**0.972**	0.944	0.930	0.971	0.915
		ROC	0.973	0.945	0.932	0.970	**0.989**
	10-Fold	Precision	0.978	0.974	**0.986**	0.979	0.979
		F-measure	0.977	0.972	0.986	0.978	**0.989**
		ROC	0.980	0.971	**0.989**	0.975	0.978
GLASS	70-30%	Precision	0.971	0.966	0.965	0.975	**0.983**
		F-measure	0.969	0.945	**0.985**	0.984	0.963
		ROC	0.969	0.962	0.985	0.984	**0.988**
	50-50%	Precision	0.966	0.959	0.966	**0.981**	0.937
		F-measure	0.963	0.959	0.963	0.981	**0.984**
		ROC	0.965	**0.991**	0.966	0.971	0.977
	10-Fold	Precision	0.954	0.959	0.959	0.924	**0.986**
		F-measure	0.954	0.959	0.959	0.956	**0.982**
		ROC	0.954	0.971	0.966	0.967	**0.985**

TABLE. IV. CLASSIFICATION RESULTS FOR OVERSAMPLING TECHNIQUES ON 3 DATASETS USING NEURAL NETWORK

Dataset	**Testing-Training**	**Measure**	**SMOTE**	**Borderline-SMOTE**	**Safe-level SMOTE**	**A-SUWO**	**IA-SUWO**
IRIS	70-30%	Precision	0.970	0.901	0.910	0.934	**0.989**
		F-measure	0.965	0.899	0.905	0.923	**0.981**
		ROC	0.985	0.909	0.921	0.992	0.991
	50-50%	Precision	0.971	0.904	0.965	0.971	**0.986**
		F-measure	0.970	0.901	0.966	0.969	**0.979**
		ROC	0.995	0.926	0.971	0.991	**0.998**
	10-Fold	Precision	0.955	0.945	0.947	0.954	**0.956**
		F-measure	**0.955**	0.940	0.947	0.954	0.946
		ROC	0.977	0.963	0.980	**0.996**	**0.996**
WINE	70-30%	Precision	0.977	**0.982**	0.958	0.974	0.977
		F-measure	0.977	0.975	0.954	0.975	**0.978**
		ROC	**0.998**	0.989	0.996	0.991	0.993
	50-50%	Precision	0.973	0.973	0.958	0.942	**0.986**
		F-measure	0.972	0.972	0.958	0.942	**0.986**
		ROC	0.997	0.994	0.992	0.997	**0.998**
	10-Fold	Precision	0.986	0.966	0.986	0.978	**0.987**
		F-measure	0.986	0.965	0.968	0.978	**0.988**
		ROC	**0.998**	0.995	0.995	0.997	0.991
GLASS	70-30%	Precision	0.969	0.969	0.940	0.975	**0.988**
		F-measure	0.969	0.969	0.938	0.974	**0.978**
		ROC	0.972	0.978	0.981	0.987	**0.995**
	50-50%	Precision	0.943	**0.988**	0.924	0.971	0.987
		F-measure	0.935	0.980	0.917	0.973	**0.994**
		ROC	0.956	0.983	0.971	**0.994**	0.990
	10-Fold	Precision	0.943	0.965	0.949	0.977	**0.979**
		F-measure	0.935	0.963	0.969	0.967	0.967
		ROC	0.966	0.966	0.968	0.969	**0.973**

VI. Conclusion

In preprocessing of data by using any resampling technique, oversampling techniques are being preferred mostly by the authors because it does not encounter the risk of losing any potential data. However, by oversampling of data, the synthetic samples over-fitted or can cause over-generalization, but this problem is avoided by IA-SUWO. The standard A-SUWO is lacking the ability to generate more appropriate strength of synthetic samples, while the results were taken after the whole experimental process showed that IA-SUWO outperformed standard technique by using sparsity factor.

Sparsity factor actually measures the density of all clusters, and any of cluster having low density results in high sparsity of the samples, which means that more synthetic samples need to be generated in that cluster. In this way, it increases the ability of any classifier to classify more accurately the minority class instances. The results of oversampling on these datasets clarify the oversampling rates and amount of imbalanced ratio reduced by improved and standard A-SUWO. The results show that the IA-SUWO gave better results in terms of reducing imbalances ratio by using sparsity factor, which appropriately generated more samples than standard A-SUWO; within the minority clusters where the samples are sparse apart and required more oversampling, these generated synthetic instances employed a positive role in the classification process.

Acknowledgment

The authors would like to thank Research Management Center (RMC), Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research, Vote No. H334, and research fund, E15501, UTHM.

References

[1] F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, “Under-sampling class imbalanced datasets by combining clustering analysis and instance selection,” Inf. Sci. (Ny)., vol. 477, pp. 47–54, 2019.

[2] S. Wang and X. Yao, “Multiclass Imbalance Problems : Analysis and Potential Solutions,” IEEE Trans. Syst. Man. Cybern., vol. 42, no. 4, pp. 1119–1130, 2012.

[3] H. Yu, J. Ni, and J. Zhao, “ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data,” Neurocomputing, vol. 101, pp. 309–318, 2013.

[4] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, “nDNA-prot: Identification of DNA-binding proteins based on unbalanced classification,” BMC Bioinformatics, vol. 15, no. 1, pp. 1–10, 2014.

[5] M. Zulqarnain, R. Ghazali, S. H. Khaleefah, and A. Rehan, “An Improved the Performance of GRU Model based on Batch Normalization for Sentence Classification,” Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 9, pp. 176–185, 2019.

[6] 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.

[7] S. Zafeiriou, C. Zhang, and Z. Zhang, “A survey on face detection in the wild: Past, present and future,” Comput. Vis. Image Underst., vol. 138, pp. 1–24, 2015.

[8] R. Malhotra, “A systematic review of machine learning techniques for software fault prediction,” Appl. Soft Comput. J., vol. 27, pp. 504–518, 2015.

[9] M. Reza, S. Miri, and R. Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 20–25, 2016.

[10] A. Abu-Srhan, B. Alhammad, S. Al, and R. Al-Sayyed, “Visualization and Analysis in Bank Direct Marketing Prediction,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, pp. 651–657, 2019.

[11] W. Y. Lin, Y. H. Hu, and C. F. Tsai, “Machine learning in financial crisis prediction: A survey,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 42, no. 4, pp. 421–436, 2012.

[12] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[13] R. Longadge, S. S. Dongre, and L. Malik, “Class imbalance problem in data mining: review,” Int. J. Comput. Sci. Netw., vol. 2, no. 1, pp. 83–87, 2013.

[14] N. V Chawla, N. Japkowicz, and P. Drive, “Editorial : Special Issue on Learning from Imbalanced Data Sets,” Sigkdd Explor., vol. 6, no. 1, pp. 2000–2004, 2004.

[15] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “A comparative study of data sampling and cost sensitive learning,” Proc. - IEEE Int. Conf. Data Min. Work. ICDM Work. 2008, pp. 46–52, 2008.

[16] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, “Finding the Best Classification Threshold in Imbalanced Classification,” Big Data Res., vol. 5, pp. 2–8, 2016.

[17] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” Inf. Sci. (Ny)., vol. 465, pp. 1–20, 2018.

[18] B. Krawczyk, M. Woźniak, and G. Schaefer, “Cost-sensitive decision tree ensembles for effective imbalanced classification,” Appl. Soft Comput. J., vol. 14, no. PART C, pp. 554–562, 2014.

[19] M. Koziarski, B. Krawczyk, and M. Woźniak, “Radial-Based oversampling for noisy imbalanced data classification,” Neurocomputing, no. 2019, 2019.

[20] M. Galar, A. Fern, E. Barrenechea, and H. Bustince, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches,” IEEE Trans. Syst. MAN, Cybern. C Appl. Rev., vol. 42, no. 4, pp. 463–484, 2012.

[21] Q. Kang, X. S. Chen, S. S. Li, and M. C. Zhou, “A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification,” IEEE Trans. Cybern., vol. 47, no. 12, pp. 4263–4274, 2017.

[22] S. Alshomrani, A. Bawakid, S. O. Shim, A. Fernández, and F. Herrera, “A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets,” Knowledge-Based Syst., vol. 73, pp. 1–17, 2015.

[23] I. Nekooeimehr and S. K. Lai-Yuen, “Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets,” Expert Syst. Appl., vol. 46, pp. 405–416, 2016.

[24] A. Agrawal, H. L. Viktor, and E. Paquet, “SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling,” 2015 7th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag., vol. 01, pp. 226–234, 2015.

[25] A. Majid, S. Ali, M. Iqbal, and N. Kausar, “Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines,” Comput. Methods Programs Biomed., vol. 113, no. 3, pp. 792–808, 2014.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[27] B. Zadrozny and C. Elkan, “Learning and making decisions when costs and probabilities are both unknown,” Proc. seventh ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’01, pp. 204–213, 2001.

[28] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, “Clustering-based undersampling in class-imbalanced data,” Inf. Sci. (Ny)., vol. 409–410, pp. 17–26, 2017.

[29] G. Douzas and F. Bacao, “Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning,” Expert Syst. Appl., vol. 82, pp. 40–52, 2017.

[30] S. Ite and S. Ite, “Class imbalances versus small disjuncts,” Sigkdd Explor., vol. 6, no. 1, pp. 40–49, 2004.

[31] J. Liu, X. Zhou, D. Li, X. Li, Z. Dong, and S. Wang, Advanced Data Mining and Applications. 2005.

[32] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE : Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," Adv. Knowl. Discov. Data Mining, PAKDD 2009, pp. 475–482, 2009.

[33] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE : A New Over-Sampling Method in," Springer-Verlag Berlin Heidelb., pp. 878–879, 2005.

[34] A. S. Nickerson, N. Japkowicz, and E. Milios, "Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets," Proc. Eighth Int. Work. AI Statitsics, no. 2001, p. 5, 2001.

[35] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proc. Int. Jt. Conf. Neural Networks, pp. 1322–1328, 2008.

[36] B. Tang and H. He, "KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning," 2015 IEEE Congr. Evol. Comput. CEC 2015 - Proc., pp. 664–671, 2015.

[37] S. Barua, M. Islam, X. Yao, and K. Murase, "MWMOTE — Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, 2014.

[38] S. Babu and N. R. Ananthanarayanan, "EMOTE: Enhanced Minority Oversampling TEchnique," J. Intell. Fuzzy Syst., vol. 33, no. 1, pp. 67–78, 2017.

[39] P. Lim, C. K. Goh, and K. C. Tan, "Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning," IEEE Trans. Cybern., vol. 47, no. 9, pp. 2850–2861, 2017.

[40] Z. Tianlun and Y. Xi, "G-SMOTE: A GMM-BASED Synthetic Minority Oversampling Technique for Imbalanced Learning," arxiv1810.10363v1 , a Prepr., 2018.

[41] T. Zhu, Y. Lin, Y. Liu, W. Zhang, and J. Zhang, "Minority oversampling for imbalanced ordinal regression," Knowledge-Based Syst., vol. 166, pp. 140–155, 2019.

[42] C. Science, D. Dhanalakshmi, A. S. Vijendran, and A. Info, "Adaptive Data Structure Based Oversampling Algorithm for Ordinal Classification," Indones. J. Electr. Eng. Comput. Sci. Vol., vol. 12, no. 3, pp. 1063–1070, 2018.

[43] E. M. Voorhees, "Implementing agglomerative hierarchic clustering algorithms for use in document retrieval," Inf. Process. Manag., vol. 22, no. 6, pp. 465–476, 1986.

Factors Contributing to the Success of Information Security Management Implementation

Mazlina Zammani[1], Rozilawati Razali[2], Dalbir Singh[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, 43600
Bangi, Selangor, Malaysia

***Abstract*—Information Security Management (ISM) concerns shielding the integrity, confidentiality, availability, authenticity, reliability and accountability of the organisation's information from unauthorised access in order to ensure business continuity and customers' confidence. The importance of information security (IS) in today's situation should be given due attention. Recognising its importance, organisations nowadays have devoted wide efforts in protecting their information. They establish information security policy, processes, and procedures as well as reengineer their organisational structures to align with ISM principles. Regardless of the efforts, security incidents continue to occur in many organisations. This phenomenon shows that the current implementation of ISM is still ineffective due to unaware of the factors contributing to the success of ISM. Thus, the objective of this paper is to identify ISM success factors and their elements through a large-scale survey. The survey involves 243 practitioners from statutory bodies, public and private organisations in Malaysia. The results of the survey indicate that top management, IS coordinator team, ISM team, IS audit team, employees, third parties, IS policy, IS procedures, resource planning, competency development and awareness, risk management, business continuity management, IS audit and IT infrastructure are the factors that contribute to the success of ISM implementation. These factors shall guide practitioners in planning and refining ISM implementation in their organisations.**

Keywords*—*Information security; information security management; success factors; key factors

I. Introduction

Information security management (ISM) is a systematic preservation approach to protect the integrity, confidentiality, availability, authenticity, reliability and accountability of information [1],[2],[3]. Every single day, organisations and their information are exposed to security threats and incidents such as malware, virus, malicious spyware, spam, phishing and sabotage from an extensive range of sources [4],[5],[6],[7]. Researches have indicated that security incidents increase over the years [8]. For example, from the year 2010 to 2016, there were 60,000 security incidents occurred at large organisations in the United States [4]. In the Netherlands, 18% of all small and medium organisations are hit by cyber-attacks each year and in Malaysia, a total of 3280 security incidents have been reported in the half-year 2018 [9],[10].

The increasing number of security incidents has led organisations to enhance their ISM plans in order to shield their critical information [11]. Thus, organisations have begun taking systematic approaches in managing their information. However, there are still weaknesses in the implementation, which causes security incidents continue to occur [1], [12]. Organisations in general are seen as fail to manage their information security (IS) appropriately. One of the main reasons is that the organisations are not aware of the factors contributing to the success of ISM implementation [13]. In essence, these success factors should be given serious attention in order to ensure that ISM is effectively implemented [14]. Therefore, this paper aims to explore the success factors of ISM implementation. The factors shall be used as a guide to organisations in bettering their ISM practices.

This paper is organised as follows. Section 2 provides a brief review of the ISM factors that were gathered from the literature. Section 3 describes the methodology used to collect and analyse the empirical data. Section 4 presents the findings of the analysis and finally, Section 5 summarises the findings.

II. Background

An ISM is a method or approach to managing information securely and effectively. It involves various aspects such as people, process, organisational documents, and technology [15], [16]. The people aspect contains the key players of ISM, who shall own certain characteristics for implementing ISM process. The implementation of the process should be guided by organisational documents and supported by the latest technology [17].

A. People

In the people aspect, the key players involved are Top management, ISM team, IS coordinator team, IS audit team, employees and third parties. Top management is a key pillar for the ISM's accomplishment. Top management should portray high leadership qualities and have good knowledge of the security objectives and governance to ensure the goal of ISM is achieved [15]. In addition, the commitment of top management is required to make decisions, provide feedback and giving support to the whole security activities undertaken [14].

The ISM team is responsible for implementing security operations and activities. Hence, the team's knowledge in the security domain is required. Skills, commitment, and cooperation of team members are necessary to ensure that security operations run smoothly. Furthermore, the willingness of the whole team members to accept changes at any time is also expected [18].

Meanwhile, the IS coordinator team is responsible for managing and coordinating the main documents and security activities such as awareness and training programs. The team acts as a liaison between top management, ISM team, IS audit team and employees. Therefore, the members of the team need to be knowledgeable about the overall ISM scope and activities. They also need to give a full commitment in coordinating the ISM activities and have good communication skills [18].

On the other hand, employees and third parties should be aware of the latest security policy, threats and issues that occur in the organisation. In order to reduce the security incidents, the employees and third parties must comply with security policy, laws and agreements [19]. In addition, employees' motivation towards the implementation of IS controls is also needed.

Apart from top management, ISM team, IS coordinator team, employees and third parties, the IS audit team equally plays a role in the success of ISM implementation [20]. The team should be committed in ensuring the security controls, processes, and activities are implemented properly [21]. The IS audit team must have knowledge of IS and the matters to be audited [22]. They need to possess the necessary auditing skills by applying appropriate auditing techniques. Additionally, communication skills are required to obtain useful information from auditees. Furthermore, the team's commitment and cooperation are also needed to ensure the effectiveness of the auditing process.

B. Organisational Documents

Organisational documents refer to the internal documents that need to be well-established and complied during the implementation of ISM. Two factors recognised under the organisational documents are IS policy and IS procedures.

IS Policy is a set of rules enacted by an organisation to ensure that all employees and third parties are complying with the IS prescription [23]. The policy should be comprehensive in covering the controls proposed by the international standards and must be in line with IS requirements and ISM scope. It must be clear in describing IS objectives and the responsibilities of the parties involved [19], [24]. In addition, the policy should be communicated and disseminated to the employees, third parties and stakeholders. It should also be periodically reviewed to ensure it is appropriate to the current needs [25]. Beside IS Policy, IS procedures also contribute to the success of ISM implementation. IS procedures are the operating guidelines that comprise a series of activities that explain how to execute IS processes. The procedure must be clear and complete in describing the work steps to be carried out. [25]. It must also be regularly reviewed and communicated among the individuals or teams involved. These features should be present in each procedure to ensure that the ISM processes can be implemented effectively.

C. Process

In the aspect of process, there are five identified factors involved in ISM namely Resource Planning, Competency Development and Awareness, Risk Management, Business Continuity Management, and IS Auditing. Resource planning contains financial and human resources to support and execute ISM processes and activities. The financial and human resources are necessary to guarantee the ISM processes and activities work efficiently [26]. In addition, competency development and awareness are also important for the success of ISM [19]. The competency development and awareness consist of training programs and awareness programs [25]. The goal of the training programs is to ensure that the individuals and teams involved in ISM operations acquire the knowledge and expertise for handling the tasks. The objective of the awareness programs is to ensure the employees, third parties and stakeholders are aware of IS policy, IS issues and IS threats as well as their responsibilities in protecting the organisation's information [25]. Risk management is the crucial process in ISM [19]. Risk management focuses on assessing, analysing, mitigating, and controlling the risks [27]. Risk assessment and risk treatment are two main activities in risk management. Risk assessment is an activity of measuring and analysing the risk levels while risk treatment is an activity of implementing the suitable actions to control the risks [28].

Business continuity management (BCM) is another vital process in ISM [29]. It is a holistic management process to ensure the continuity of critical processes whenever disasters or unintended events happen [30],[31]. BCM requires a business continuity plan (BCP) that highlights the resources, activities, and responsibilities for managing the unintended events. In order to affirm the plan is feasible and effective, the organisation shall perform periodic tests on the plan. Beside risk management and BCM, IS audit is also another key process in ISM [25],[32]. Through the IS audit process, the compliance of IS policy, procedures and controls can be checked and evaluated [33]. The major items in the audit process are audit program, audit findings and reporting as well as follow-up audit to check the corrective and preventive actions that have been taken [20].

D. Technology

Another factor that contributes to the effectiveness of ISM implementation is the IT infrastructure that is listed under the technology aspect [34]. IT infrastructure includes the use of the latest software and hardware that support the implementation and monitoring of security operations. The review above indicates that there are various success factors that need to be considered when implementing ISM. The comprehensive explanation about the factors and their elements can be found in [18].

III. METHODOLOGY

This study adopts a mixed-method approach, which comprises qualitative and quantitative approaches. The purpose of combining these two approaches is to gain a better understanding of the phenomenon [35]. Fig. 1 illustrates the research design. The qualitative part consists of analysing the existing literature together with a series of interviews with experienced ISM practitioners. The results of the qualitative part have been reported in [18]. The study then further adopts a quantitative approach to confirm and refine the qualitative findings through a large-scale survey. The results are discussed in this paper.

TABLE. I. DETAILS OF THE QUESTIONNAIRE

Aspect	CSFs	Number of Items	Items
People	Top Management	5	Knowledge of top management regarding ISM objectives and governance.
			Knowledge of top management on ISM issues and problems
			The leadership of top management in leading the ISM.
			The commitment of top management in supporting the financial and human resources.
			The commitment of top management in giving feedback on any ISM issues and problems
	IS Coordinator Team	3	Knowledge of coordinator team on overall ISM scope and activities.
			The commitment of coordinator team in coordinating ISM activities
			Good communication skills.
	ISM Team	5	Knowledge of IS team in the information security domain.
			Skills of ISM team in handling and operating procedures of information security.
			The commitment of IS team towards information security operations that are being carried out.
			The IS team's willingness to accept additional assignments based on current needs.
			Cooperation between IS team members in carrying out security operations.
	IS Audit Team	5	Knowledge of IS Audit team regarding the ISM scope that needs to be audited.
			Skills of IS Audit team in applying auditing techniques.
			The commitment of IS Audit team throughout the auditing process.
			Cooperation between IS Audit team members.
			Good communication skills with the auditee.
	Employees	3	Awareness of employees regarding the importance of information security.
			Employees' compliance with the requirements outlined in the information security policy.
			Employees' motivation towards the implementation of information security controls
	Third Parties	3	Awareness of third parties regarding the importance of information security.
			Third parties' compliance with the requirements outlined in the information security policy.
			Third parties' compliance on information security agreement that has been signed.
Organisational Documents	IS Policy	4	Clear in defining IS objectives, the roles, and responsibilities of the employees, third parties, and stakeholders.
			Comprehensive which covers the requirements and controls set by the ISM standards and aligns with the ISM scope.
			Communicated to all employees, third parties, and other stakeholders.
			Reviewed/revised periodically or according to current needs.
	IS Procedures	4	Clearly explain the objectives and responsibilities of the individuals and team involved in the execution of procedures.
			Complete in describing the work steps to be carried out.
			Reviewed/revised periodically or according to current needs.
			Communicated to individuals or team involved.
Process	Resource Planning	2	Financial Resources for purchasing new assets and maintaining existing assets, the cost of manpower and the cost to perform IS activities.
			Human Resource for implementing IS processes and operation.
	Competency Development & Awareness	2	Awareness programs for all employees, third parties, and stakeholders.
			Training programs for individuals and teams involved in ISM processes and operation.
	Risk Management	2	Risk Assessment
			Risk Treatment
	Business Continuity Management	2	Business continuity plan that outlines the resources, procedures, activities, and responsibilities of the individuals and teams involved.
			Simulation (testing) on the business continuity plan
	IS Audit	3	Audit programs which consists of audit planning, audit training, and audit execution.
			Audit findings and reporting.
			Follow-up audit to check the corrective and preventive actions that have been done.
Technology	IT Infrastructure	2	Software to support the implementation and monitoring of information security.
			Hardware to support the implementation and monitoring of information security.

A. Sampling

The sampling method used for the survey was a stratified random sampling. A stratified random sampling is one obtained by separating the population into non-overlapping groups (e.g., ISM Team, IS Audit Team, ISMS Coordinator Team) and then applying a simple random sample from each stratum. This technique was chosen to ensure the presence of the key subgroups within each sample. The samples used in the study were ISM experts and practitioners from statutory bodies, public and private organisations in Malaysia.

B. Instruments

A set of questionnaire was developed based on the findings obtained from the earlier qualitative study. The contents of the questionnaire were reviewed by two experts from ISM field. A pilot study involving 30 ISM practitioners was conducted to ensure the reliability of the questionnaires. The questionnaire was then amended based on the feedback received from the experts and pilot study.

The questionnaire was divided into two sections. The first section consisted of the respondent's demographic profile. Meanwhile, the second section comprised questions regarding ISM factors and their items, as outlined in Table I. Respondents were required to respond to the questions on a 5-point Likert scale (1-Strongly Disagree, 2-Disagree, 3-Somewhat Agree, 4-Agree, 5-Strongly agree). In total, there were fourteen ISM factors and forty-five items all together to be confirmed.

C. Protocol

For distributing the questionnaires, the study hired fifteen enumerators and used online forms in order to ensure they were efficiently collected. A total of 400 questionnaires were disseminated to potential respondents among the statutory bodies, public and private agencies in Malaysia, based on the predetermined criteria set by the researchers. A two-week period was given to the enumerators to distribute and follow up with the respondents.

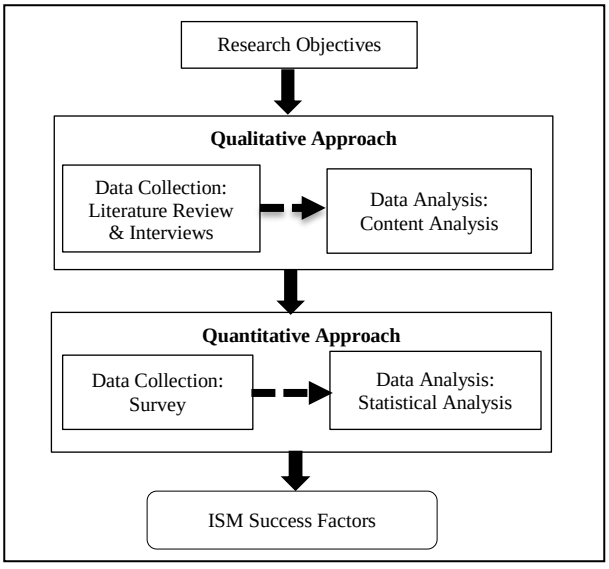


Fig. 1. Research Design.

D. Analysis

From 400 questionnaires that were distributed, only 255 questionnaires were returned, which makes the response rate is 64%. However, only 243 questionnaires were considered for the data analysis, as the remaining were incomplete. The collected data was then analysed by using statistical analysis, as described in the later section.

IV. Result and Discussion

A test of scale reliability is executed to guarantee the measurement scales have consistently and accurately captured the meaning of the constructs. There are several common methods used in measuring reliability. The method used in this study is the internal consistency by determining the Cronbach alpha coefficient (α). Cronbach alpha values must be above 0.6 to confirm the consistency and reliability of the constructs [36]. The higher the value of Cronbach's alpha coefficient (α), the greater the reliability of the internal consistency. Table II shows the results of the Cronbach alpha coefficient (α) analysis. The values of the Cronbach's alpha coefficient of all the constructs range from 0.798 to 0.939, suggesting that the entire scale has a good level of internal consistency.

Table III shows the demographic characteristics of the respondents. Most respondents participated in this study are from government agencies (92.6%), followed by private agencies (4.1%) and statutory bodies (3.3%). Majority of the respondents have been in the industry for 5-15 years. 44.9% of the respondents have less than 3 years of experience in ISM while 51% have over 3 years of experience in ISM. In terms of ISM specialisation, 46.1% of respondents are in the ISM team, 9.9% from the coordinator team, 4.1% are in the audit team, and 10.7% are involved in multiple categories.

TABLE. II. Internal Consistency

Factors	**Number of Items**	**Value of Cronbach's Alpha (α)**
Top Management	5	.884
IS Coordinator Team	3	.843
ISM Team	5	.906
IS Audit Team	5	.939
Employees	3	.897
Third Parties	3	.903
IS Policy	4	.905
IS Procedures	4	.926
Resource Planning	2	.798
Competency Development & Awareness	2	.855
Risk Management	2	.923
Business Continuity Management	2	.884
IS Audit	3	.922
IT Infrastructure	2	.875

TABLE. III. RESPONDENTS' DEMOGRAPHIC CHARACTERISTICS

Respondent characteristics	Frequency	Percentage (%)
Agency Type		
Government	225	92.6
Statutory Body	8	3.3
Public Agency	10	4.1
Length of Service		
Not stated	2	.8
Less than 5 years	26	10.7
5 – 10 years	92	37.9
11-15 years	77	31.7
Over 15 years	46	18.9
ISM experience		
Not stated	10	4.1
Less than 3 years	109	44.9
3 - 6 years	69	28.4
Over 6 years	55	22.6
ISM specialization		
Not specified	5	2.1
Top management	5	2.1
Coordinator team	24	9.9
ISM team	112	46.1
Audit team	10	4.1
Employees	52	21.4
Others	9	3.7
Combined categories	26	10.7

The distribution of respondents' feedback on the ISM success factors are shown in Table IV. The findings indicate that most respondents agreed that the fourteen factors contribute to the success of ISM implementation. This is demonstrated by the number of respondents (N) who chose "Agree" and "Strongly Agree" values are higher than the number of respondents who chose "Strongly disagree" and "Disagree". This indicates positive agreements towards all factors. These findings confirm the previously gathered qualitative data.

Table IV shows that Top Management is the most highly agreed factor concerning the success of ISM implementation, followed by IS Policy, IS Procedures, ISM Team, IT Infrastructure, Risk Management and IS Coordinator Team (N for "Strongly Agree" > 50%).

Fig. 2 shows the median values for the whole factors. Seven factors, namely Top Management, IS Coordinator Team, ISM Team, IS Policy, IS Procedures, Risk Management, and IT Infrastructure have a median value of 5 while other factors have a median value of 4.

TABLE. IV. DISTRIBUTION OF RESPONDENTS' FEEDBACK ON ISM SUCCESS FACTORS

ISM Success Factors	Strongly disagree	Disagree	Somewhat agree	Agree	Strongly agree
	N	N	N	N	N
Top Management	0 (0.0%)	3 (1.2%)	17 (7.0%)	71 (29.2%)	152 (62.6%)
IS Policy	0 (0.0%)	0 (0.0%)	15 (6.2%)	79 (32.5%)	149 (61.3%)
IS Procedures	0 (0.0%)	0 (0.0%)	13 (5.3%)	85 (35.0%)	145 (59.7%)
ISM Team	0 (0.0%)	0 (0.0%)	10 (4.1%)	89 (36.6%)	144 (59.3%)
IT Infrastructure	0 (0.0%)	0 (0.0%)	17 (7.0%)	90 (37.0%)	136 (56.0%)
Risk Management	0 (0.0%)	1 (0.4%)	14 (5.8%)	101 (41.6%)	127 (52.3%)
IS Coordinator Team	0 (0.0%)	2 (0.8%)	11 (4.5%)	104 (42.8%)	126 (51.9%)
Competency Development & Awareness	0 (0.0%)	0 (0.0%)	22 (9.1%)	104 (42.8%)	117 (48.1%)
Employees	0 (0.0%)	2 (0.8%)	24 (9.9%)	105 (43.2%)	112 (46.1%)
Resource Planning	0 (0.0%)	3 (1.2%)	21 (8.6%)	108 (44.4%)	111 (45.7%)
Business Continuity Management	0 (0.0%)	1 (0.4%)	20 (8.2%)	115 (47.3%)	107 (44.0%)
IS Audit	0 (0.0%)	2 (0.8%)	24 (9.9%)	111 (45.7%)	106 (43.6%)
IS Audit Team	0 (0.0%)	4 (1.6%)	26 (10.7%)	110 (45.3%)	103 (42.4%)
Third Parties	0 (0.0%)	5 (2.1%)	54 (22.2%)	106 (43.6%)	78 (32.1%)

Fig. 3 shows mode value for the whole factors. Ten factors have the mode value of 5 and four factors have the mode value 4. Based on the mode and medium values obtained, it clearly verifies that all the factors affecting ISM's success.

Meanwhile for the items analysis, the minimum, maximum, median and mode values for each item are tabulated as in Table V. Generally, 45 items were analysed from a total of 243 respondents.

The findings demonstrate that all the items have a maximum value of 5. For the minimum value, there are three items that have a minimum value of 1, which are commitment-feedback of top management, communication skills of audit team, and awareness programmes, while the rest of the items have the minimum value of 2 and 3. In addition, all the items have a median and mode value of 4 and 5. In short, these findings indicate that the respondents agreed that all the factors and their items contribute to the success of ISM implementation.

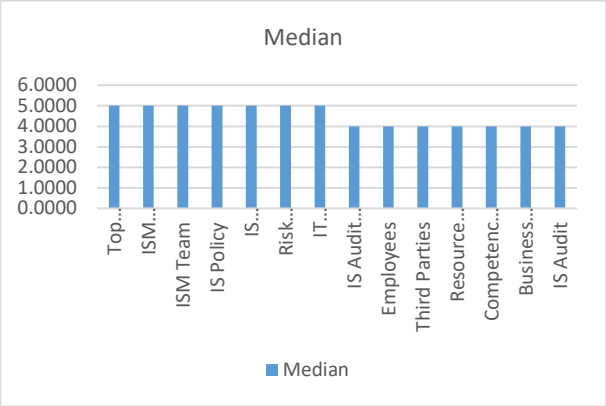


Fig. 2. Median Values for ISM Implementation Success Factors.

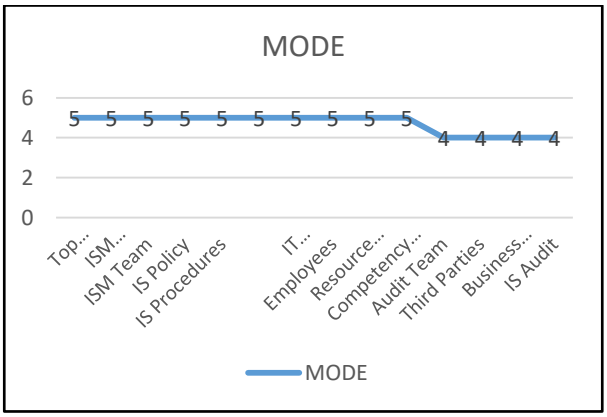


Fig. 3. Mode Values for ISM Implementation Success Factors.

A. Factor Analysis

Factor analysis is used to confirm the contributing factors of ISM implementation. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were first conducted to determine the sampling adequacy. Fig. 4 shows that the score of Bartlett's Test is 91 (significance 0.000), which means there is a real inter-variable correlation (significance < 0.05). The score of KMO is 0.936, which means all variables in this study could be used for further analysis.

Table VI shows the Measure of Sampling Accuracy (MSA) value obtained by each factor. Any factor that has the score of MSA > 0.5 will be seen as valid. The findings show that the MSA scores for all factors are greater than 0.5.

Table VII displays the result of the factors loading. It shows that the result of all factors were > 0.3. The factors were grouped into two components; component 1 and component 2. Component 1 includes the factors in the aspect of process, organisational documents and technology which are Business Continuity Management, Risk Management, Competency Development & Awareness, Resource Planning, IS Audit, IS Policy, IS Procedures and IT Infrastructure. Meanwhile components 2 represent the factors in the aspect of people which are IS Coordinator Team, Top Management, ISM Team, Employees, IS Audit Team and Third Parties. Based on the analysis, it can be concluded that all fourteen factors contribute to the success of ISM implementation.

TABLE. V. ITEMS ANALYSIS OF ISM SUCCESS FACTORS

Factors	Items	Min.	Max.	Median	Mode
Top management	knowledge-objectives	3	5	5	5
	knowledge-issues	3	5	5	5
	leadership	3	5	5	5
	commitment-resources	2	5	5	5
	commitment-feedback	1	5	5	5
IS Coordinator team	knowledge	3	5	5	5
	commitment	2	5	5	5
	communication skill	3	5	5	5
ISM team	knowledge	3	5	5	5
	skill	3	5	5	5
	commitment	2	5	5	5
	readiness	2	5	4	4
	cooperation	3	5	5	5
IS Audit team	knowledge	3	5	5	5
	auditing skill	2	5	5	5
	commitment	2	5	5	5
	cooperation	3	5	4	5
	communication skills	1	5	4	4
Employees	awareness	2	5	5	5
	compliance	2	5	5	5
	motivation	2	5	4	4
Third Parties	awareness	2	5	4	4
	compliance – IS policy	2	5	4	5
	compliance – agreement	2	5	4	5
IS Policy	clear	3	5	5	5
	comprehensive	2	5	5	5
	communicated	3	5	5	5
	reviewed	3	5	5	5
IS Procedures	clear	2	5	5	5
	complete	2	5	5	5
	reviewed	2	5	4	5
	communicated	2	5	5	5
Resource Planning	financial resources	2	5	4	5
	human resources	3	5	5	5
Competency Development & Awareness	awareness program	1	5	4	4
	training	2	5	5	5
Risk Management	risk assessment	2	5	5	5
	risk treatment	2	5	5	5
Business Continuity Management	business continuity plan	3	5	5	5
	simulation (testing)	3	5	4	5
IS Audit	audit programs	3	5	4	4
	audit findings and reporting	3	5	4	5
	follow-up audit	2	5	4	4
IT Infrastructure	software	3	5	4	4
	hardware	3	5	4	5

B. Kruskal-Wallis H test

A non-parametric test Kruskal-Wallis H is carried out to investigate any significant differences in the respondents' views among various sectors (public, private and statutory bodies). Fig. 5 shows the results of Kruskal-Wallis H test analysis. The results indicate that there is no significant difference in the view of the ISMS success factors among public, private and statutory bodies respondents [x2 (2, N = 243) = .154, p> 0.05]. In other words, respondents of the public, private and statutory bodies have similar views on the factors that contribute to the success of ISM implementation.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.936
Bartlett's Test of Sphericity	Approx. Chi-Square	3093.826
	df	91
	Sig.	.000

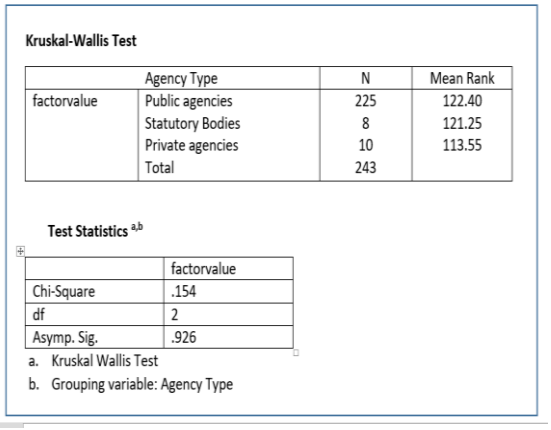
Fig. 4. KMO and Bartlett's Test.

TABLE. VI. ANTI-IMAGE CORRELATION MATRIX OF ITEMS

Factors	MSA Score
Top Management	.930
IS Coordinator Team	.918
ISM Team	.942
IS Audit Team	.917
Employees	.952
Third Parties	.964
IS Policy	.892
IS Procedures	.904
Resource Planning	.969
Competency Development & Awareness	.960
Risk Management	.937
Business Continuity Management	.949
IS Audit	.930
IT Infrastructure	.963

TABLE. VII. FACTORS LOADING

Rotated Component Matrix[a]		
	Component 1	Component 2
Business Continuity Management	.828	
Risk Management	.814	
IS Procedures	.792	
Competency Development & Awareness	.783	
Resource Planning	.763	
IT Infrastructure	.746	
IS Policy	.743	
IS Audit	.711	
IS Coordinator Team		.843
Top Management		.777
ISM Team		.773
Employees		.723
IS Audit Team		.715
Third Parties		.688



Kruskal-Wallis Test

	Agency Type	N	Mean Rank
factorvalue	Public agencies	225	122.40
	Statutory Bodies	8	121.25
	Private agencies	10	113.55
	Total	243	

Test Statistics [a,b]

	factorvalue
Chi-Square	.154
df	2
Asymp. Sig.	.926

a. Kruskal Wallis Test
b. Grouping variable: Agency Type

Fig. 5. Kruskal-Wallis H Test.

V. CONCLUSION AND FUTURE WORK

This study has confirmed fourteen factors and forty-five items that contribute to the success of ISM implementation. The factors and items were confirmed quantitatively through a survey. The fourteen factors are Top Management, IS Coordinator Team, ISM Team, IS Audit Team, Employees, Third Parties, IS Policy, IS Procedures, Resource Planning, Competency Development and Awareness, Risk Management, Business Continuity Management, IS Internal Audit, and IT infrastructure. All these factors and items are classified into four aspects, namely, Human, Organisational Documents, Process and Technology that should be taken into account in order to ensure the effectiveness of ISM implementation. These findings shall be used by practitioners to strategise ISM initiatives in their respective organisations. As ISM initiatives are indeed continuous, future work may need to look into ways on how to measure ISM implementation level based on these factors.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia for supporting the research. The authors also thank the respondents who participated in the study.

REFERENCES

[1] J. D. E. Lange, R. V. O. N. Solms, and M. Gerber, "Information Security Management in Local Government," pp. 1–11, 2016.

[2] M. Dhingra, "Review on Information Security Management," Int. Conf. Futur. Trends Eng. Sci. Humanit. Technol., pp. 1–4, 2016.

[3] A. Kurnianto, R. Isnanto, and A. Puji Widodo, "Assessment of Information Security Management System based on ISO/IEC 27001:2013 On Subdirectorate of Data Center and Data Recovery Center in Ministry of Internal Affairs," in E3S Web of Conferences, 2018, vol. 31, p. 11013.

[4] M. a Kuypers, T. Maillart, and E. Paté-cornell, "An Empirical Analysis of Cyber Security Incidents at a Large Organization," pp. 1–22, 2016.

[5] K. Hajdarevic and P. Allen, "A new method for the identification of proactive information security management system metrics," in Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on, 2013, pp. 1121–1126.

[6] R. A. Abbas, M. R. Mokhtar, R. Sulaiman, Z. A. Othman, and A. M. Zin, "Impact of disasters in Southeast Asia on Malaysian computer networks," J. Theor. Appl. Inf. Technol., vol. 37, no. 2, pp. 188–198, 2012.

[7] M. I. Alshar'e, R. Sulaiman, M. R. Mukhtar, and A. M. Zin, "A user protection model for the trusted computing environment," J. Comput. Sci., vol. 10, no. 10, pp. 1692–1702, 2014.

[8] Y. Bobbert and H. Mulder, "Governance Practices and Critical Success factors suitable for Business Information Security," in International Conference on Computational Intelligence and Communication Networks Governance, 2015, pp. 1097–1104.

[9] F. Mijnhardt, T. Baars, and M. R. Spruit, "Organizational Characteristics Influencing Information Security Maturity," J. Comput. Inf. Sci., vol. 56, no. 2, pp. 106-115, 2016.

[10] MyCERT, "Reported Incidents based on General Incident Classification Statistics 018," 2018. Available at https://www.mycert.org.my/portal/statistics-content?menu=b75e037d-6ee3-4d11-8169-66677d694932&id=e49c91c1-4b04-4748-8152-294764f9c8dc

[11] H. Kong, J. Woo, T. Kim, and H. Im, "Will the Certification System for Information Security Management Help to Improve Organizations' Information Security Performance ? The Case of," vol. 9, no. June, pp. 1–12, 2016.

[12] M.S. Mohd Asri and R. Rozilawati, "An assessment model of information security implementation levels," in Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, 2011, pp. 1–6.

[13] M. Zammani and R. Razali, "Information security management success factors," Adv. Sci. Lett., vol. 22, no. 8, pp. 904–913, 2016.

[14] N. Maarop, N. Mustapha, R. Yusoff, R. Ibrahim, and N. M. M. Zainuddin, "Understanding Success Factors of an Information Security Management System Plan Phase Self-Implementation," Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng., vol. 9, no. 3, pp. 884–889, 2015.

[15] Z. A. Soomro, M. H. Shah, and J. Ahmed, "Information security management needs more holistic approach: A literature review," Int. J. Inf. Manage., vol. 36, no. 2, pp. 215–225, 2016.

[16] M. M. Edwards, "Identifying Factors Contributing Towards Information Security Maturity in an Organization," Nova Southeastern University, 2018.

[17] S. Yulianto, C. Lim, and B. Soewito, "Information Security Maturity Model: A Best Practice Driven Approach to PCI DSS Compliance," in In Region 10 Symposium (TENSYMP), 2016, pp. 65–70.

[18] M. Zammani and R. Razali, "An Empirical Study of Information Security Management Success Factors," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 904–913, 2016.

[19] M. A. Alnatheer, "Information Security Culture Critical Success Factors," in 2015 12th International Conference on Information Technology - New Generations, 2015, pp. 731–735.

[20] R. Z. W. S. M.S, "CyberSecurity Malaysia: Towards Becoming a National Certification Body for Information Security Management Systems Internal Auditor," vol. 10, no. 8, pp. 2907–2910, 2016.

[21] M. Suomu, "Automated ISMS control auditability," no. May, 2015.

[22] L. Yang, "Study on the Improvement of the Internal Audit Work in IT Environment," in 2011 Fourth International Symposium on Knowledge Acquisition and Modeling, 2011, pp. 233–236.

[23] M. Razilan, A. Kadir, S. Norwahidah, S. Norman, S. A. Rahman, and A. Bunawan, "Information Security Policies Compliance among Employees in Cybersecurity Khalid S . Soliman International Business Information Management Association (IBIMA)," no. November 2016, 2017.

[24] W. Sung and S. Kang, "An Empirical Study on the Effect of Information Security Activities: Focusing on Technology, Institution, and Awareness," Proc. 18th Annu. Int. Conf. Digit. Gov. Res., pp. 84–93, 2017.

[25] A. N. Singh, M. P. Gupta, and A. Ojha, "Identifying factors of 'organizational information security management,'" J. Enterp. Inf. Manag., vol. 27, no. 5, p. 8, 2014.

[26] K. Haufe, "Maturity based approach for ISMS," Universidad Carlos III de Madrid, 2017.

[27] A. Singh, "Improving information security risk management," ProQuest Diss. Theses, p. 121, 2009.

[28] S. Fenz, J. Heurix, T. Neubauer, and F. Pechstein, "Current challenges in information security risk management," Inf. Manag. Comput. Secur., vol. 22, no. 5, pp. 410–430, 2014.

[29] N. H. Mansol, N. Hayaati, M. Alwi, and W. Ismail, "Success Factors towards Implementation of Business Continuity Management in Organizations," Int. J. Digit. Soc., vol. 5, no. 1/2, pp. 869–871, 2014.

[30] N. H. Mansol, N. Hayaati, M. Alwi, and W. Ismail, "Embedding Organizational Culture Values towards Successful Business Continuity Management (BCM) Implementation," in 2014 International Conference on Information Technology and Multimedia (ICIMU), 2014, pp. 31–37.

[31] N. Aisyah, S. Abdullah, N. L. Noor, E. Nuraihan, and M. Ibrahim, "Contributing Factor To Business Continuity Management (Bcm) Failure – a Case of Malaysia Public Sector," in Proceedings of the 5th International Conference on Computing and Informatics, 2015, no. 077, pp. 530–538.

[32] A. Tsohou, S. Kokolakis, C. Lambrinoudakis, and S. Gritzalis, "A security standards' framework to facilitate best practices' awareness and conformity," Inf. Manag. Comput. Secur., vol. 18, no. 5, pp. 350–365, 2010.

[33] R. Valverde, M. Wolden, R. Valverde, and M. Talla, "The effectiveness of COBIT 5 Information Security Framework for reducing Cyber Attacks on Supply Change Management System," IFAC-PapersOnLine, vol. 48, no. 3, pp. 1846–1852, 2015.

[34] A. A. Norman and N. M. Yasin, "Information Systems Security Management (ISSM) Success Factor: Retrospection From the Scholars," Proceedings of the 11th European Conference on Information Warfare and Security, no. July 2012. pp. 339–344, 2012.

[35] J. W. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 2013.

[36] H. D. Ghazali, "Kesahan dan kebolehpercayaan dalam kajian kualitatif dan kuantitatif," J. Pendidik. Maktab Perguru. Islam, no. 1985, pp. 61–82, 2005.

Improving Long Short-Term Memory Predictions with Local Average of Nearest Neighbors

Anibal Flores[1]
Grupo de Investigación en Ciencia de Datos, Universidad Nacional de Moquegua, Moquegua, Perú

Hugo Tito[2]
E.P. Ingeniería de Sistemas e Informática, Universidad Nacional de Moquegua, Moquegua, Perú

Deymor Centty[3]
E.P. Ingeniería Ambiental Universidad Nacional de Moquegua Moquegua, Perú

***Abstract*—The study presented in this paper aims to improve the accuracy of meteorological time series predictions made with the recurrent neural network known as Long Short-Term Memory (LSTM). To reach this, instead of just making adjustments to the architecture of LSTM as seen in different related works, it is proposed to adjust the LSTM results using the univariate time series imputation algorithm known as Local Average of Nearest Neighbors (LANN) and LANNc which is a variation of LANN, that allows to avoid the bias towards the left of the synthetic data generated by LANN. The results obtained show that both LANN and LANNc allow to improve the accuracy of the predictions generated by LSTM, with LANN being superior to LANNc. Likewise, on average the best LANN and LANNc configurations make it possible to outperform the predictions reached by another recurrent neural network known as Gated Recurrent Unit (GRU).**

***Keywords*—Long Short-Term Memory; Local Average of Nearest Neighbors; univariate time series prediction; LANN; LANNc; Gated Recurrent Unit; GRU**

I. Introduction

Forecasting is one of the most exciting subfields in the field of time series. Since the beginning, forecasting techniques have evolved greatly from simple linear regressions, passing for moving averages, autoregressive models, machine learning models, until reach Deep Learning [1] techniques.

Within Deep learning, for forecasting activities, recurrent neural networks are very common, and within them Long Short-Term Memory [2] and the Gated Recurrent Unit [3].

Long Short-Term Memory in many forecasting works has been used successfully, and the changes implemented to improve or reduce the error rate mainly includes input adjustments, tunning of parameters, number of layers, training epochs, etc.

After analyzing and evaluate the prediction results of an LSTM model in a 4-year meteorological time series corresponding to maximum temperatures, it was observed that various synthetic values could better approximate their real values through imputation processes. Fig. 1 shows in most cases how the imputation of a single value can improve the error rate of an estimated value with LSTM.

Thus, in this paper, it is proposed to improve the error rate of LSTM predictions through univariate time series imputation algorithms, such as: Local Average of Nearest Neighbors (LANN) [4] and a variation of this that will be called LANNc. LANNc tries to solve the problem of bias to the left of the synthetic data generated by LANN, just as CBRm [5] does with CBRi "in press" [6].

For the experimentation of the proposals, gap-sizes in the range of 1 to 11 consecutive NA values are evaluated, also different amounts of predicted values 15, 30, 60, 90, 120 and 150 days are considered. The results achieved show that the algorithms used allow to improve the error rate of LSTM.

It is important to highlight that the study is limited to the prediction of highly seasonal meteorological time series as it is the case of maximum temperature time series.

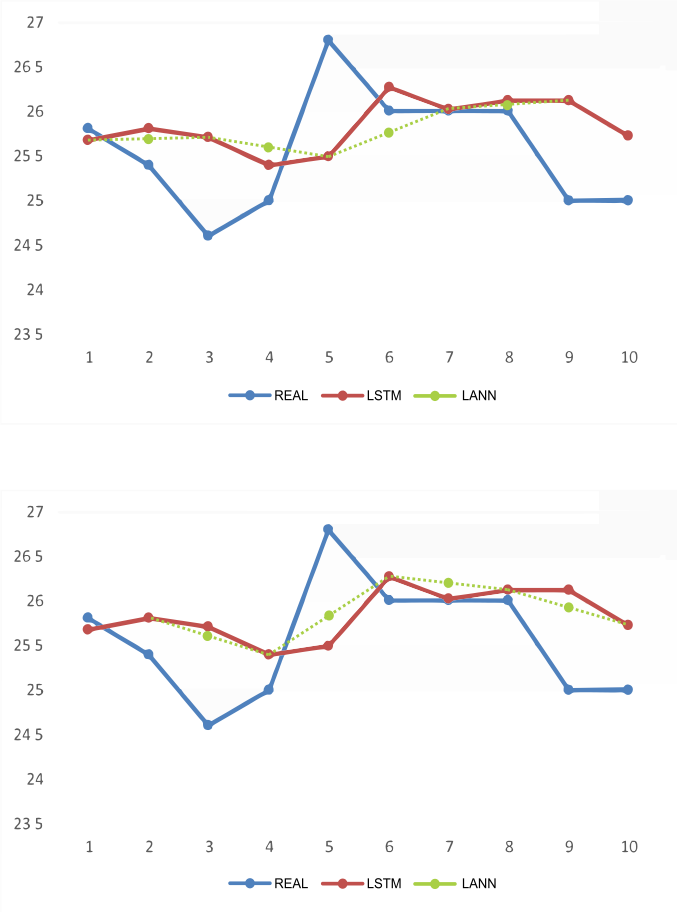


Fig. 1. How LANN can Improve LSTM Predictions.

The content of this paper is organized as follows: In Section II, a brief review of the state of the art in relation to forecasting techniques is shown; in Section III, some concepts and definitions are included that will allow a better understanding of the paper content; in Section IV, the implementation of the proposal is described; then, in Section V, the results achieved are described and analyzed; in Section VI, the results achieved are compared with other proposals of the state of the art; in Section VII, the conclusions reached at the end of this work are described; and finally it shows the future work that can be implemented to improve the results achieved.

II. Related Work

A. Simple Linear Regression

In the field of regression, one of the first models used for forecasting is known as linear regression. It consists in a statistical analysis to identify the relationship between the dependent and independent variables [7]. Simple Linear Regression is given by equation (1)

$$y(t) = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (1)$$

B. ARIMA

ARIMA [8] (Autoregressive Integrated Moving Average) is a statistical model that uses variations and regressions of statistical data in order to find patterns for a prediction into the future. It is a dynamic time series model, that is, future estimates are explained by past data and not by independent variables.

Some works that implement ARIMA for forecasting are briefly described below:

In [9] the authors implement the ARIMA model and the NNT Back Propagation Neural Network to forecast wind speed time series. The results show a slight superiority of the ARIMA model relative to the model based on neural networks.

In [10] the authors of the paper propose an ARIMA model to predict the number of epidemic disease for a center of disease control and prevention. The results achieved are compared with those of a model based on Simple Moving Average (SMA) showing the superiority of ARIMA over SMA.

In [11] the authors implement ARIMA and Support Vector Machine models to forecast load time series. The results show that ARIMA is better for linear type of load, while SVM is better for non-linear type of load time series.

C. Prophet

Prophet [12] is a forecasting decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in equation (2).

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2)$$

Where: g(t) is the trend function, s(t) represents periodic changes and h(t) represents the holidays, ϵ_t represents the error.

D. Gated Recurrent Unit (GRU)

GRUs are a gating mechanism in recurrent neural networks [3]. The GRU is like a long-term memory with a forget gate and with fewer parameters than LSTM, since it lacks an output gate. GRUs have been shown to exhibit even better performance in certain smaller datasets. However, the LSTM is "strictly stronger" than the GRU, since it can easily perform an unlimited count, while the GRU cannot [13]. Fig. 2 shows the architecture of GRU.

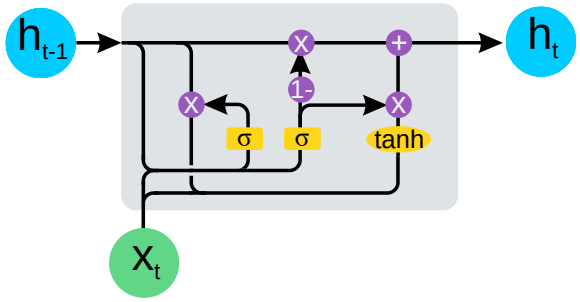


Fig. 2. Architecture of GRU.

Some works that implement GRU are briefly described below:

In [14], the authors propose the use of LSTM and GRU recurrent neural networks for electric load time series forecasting. The results achieved show the superiority of the GRU results over the results achieved by LSTM.

In [15] the authors propose the use of LSTM and GRU for traffic flow prediction comparing the results achieved with ARIMA model, demonstrating the superiority of the recurrent neural networks in this type of time series.

In [16] the authors propose the power load forecasting of residential community using Gated Recurrent Unit (GRU). The GRU results are compared with LSTM results in different configurations. GRU proved to be more efficient than LSTM for this type of time series.

In [17] the authors propose the use of a multilayer recurrent neural network called MS-GRU to forecast load electricity time series. The results are compared with the traditional recurrent neural networks such as LSTM and GRU, showing greater precision in the proposal MS-GRU.

III. Background

A. Recurrent Neural Networks (RNN)

A recurrent neural network is a neural network model for modeling time series [2]. The structure of this type of network is very similar to that of a standard multilayer perceptron (MLP), with the difference that it allows connections between hidden units associated with a time delay. Through these connections, the model can retain information from the past [18], allowing it to discover temporal correlations between events that may be very far from each other. Fig. 3 shows the architecture of a recurrent neural network.

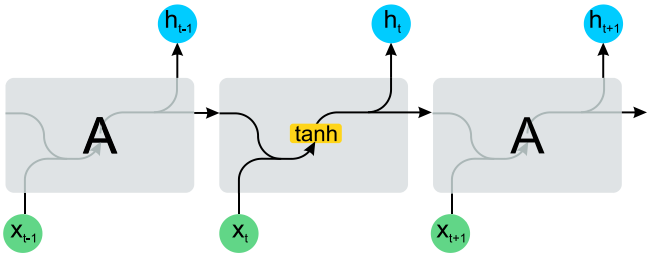


Fig. 3. Architecture of Recurrent Neural Network.

Recurrent neural networks are difficult to train [2] due to the problems of vanishing and exploding gradients, these problems resulted in the creation of LSTM networks.

B. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) network was created with the goal of addressing the vanishing gradients problem, this is due to the unfold process of an RNN. LSTM networks use special hidden units, whose task is to remember entries for a long time [1]. LSTM networks have subsequently proved to be more effective than conventional RNNs [1], especially when they have several layers for each time step. Fig. 4 shows the LSTM architecture.

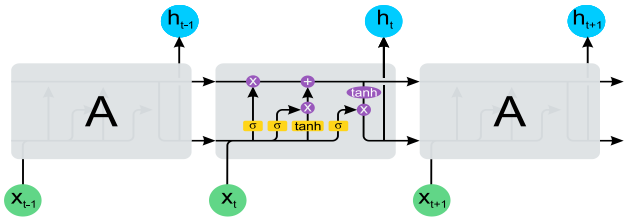


Fig. 4. Architecture of LSTM Network.

C. Local Average of Nearest Neighbors (LANN)

Local Average of Nearest Neighbors (LANN) [4] is a very simple algorithm for univariate time series imputation that uses the mean of the prior and next value of a block of NAs to replace the missing or NA values according equation (3).

$$NA=(\pi\rho\iota o\rho + \nu\varepsilon\xi\tau)/2 \quad (3)$$

LANN produces very good results in the imputation processes because, according to the analysis carried out in [4] the values closest to the missing values correspond to prior and next.

IV. Process

A. Selection of Time Series

The selected time series corresponds to maximum daily temperatures at the SENAMHI[1] meteorological station known as Punta de Coles[2] in the city of Ilo-Peru. The data that will be used for the training correspond to 4 years (2012-2015) and the data that will be used for testing correspond to the year 2016.

B. LSTM Model

The architecture that will be used to implement LSTM is shown in Fig. 5.

[1] https://www.senamhi.gob.pe/
[2] Lat.: 17°41'55.2"S Long.:71°22'25"W Alt.: 25 msnm

```
lstm_model = Sequential()

lstm_model.add(LSTM(units=50, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
lstm_model.add(Dropout(0.2))

lstm_model.add(LSTM(units=50, return_sequences=True))
lstm_model.add(Dropout(0.2))

lstm_model.add(LSTM(units=50, return_sequences=True))
lstm_model.add(Dropout(0.2))

lstm_model.add(LSTM(units=50))
lstm_model.add(Dropout(0.2))
lstm_model.add(Dense(units = 1))

lstm_model.compile(optimizer = 'adam', loss = 'mean_squared_error')
lstm_model.fit(lstm_ftrs, labels, epochs = 100, batch_size = 32)
```

Fig. 5. Architecture for LSTM Model in Python.

C. Inserting NAs

Once the LSTM model is built and compiled predicting a certain number of days, NA values are inserted, which will be calculated in the next stage using the LANN and LANNc algorithms. The NA values blocks were inserted uniformly in the predicted time series as shown in Fig. 6.

D. Applying LANN/LANNc

LANN is the first algorithm that is used to impute the predicted time series adapted from [4] for gap-sizes over 2 NA values.

LANNc: is an adaptation of LANN algorithm that solves the problem of bias to the left in imputation processes with gap-sizes over 1 NA value.

Fig. 7 shows the difference between the LANN and LANNc imputation.

The adapted LANN algorithm is shown in Table I.

E. Evaluation

The evaluation of the results is performed through Root Mean Squared Error (RMSE) that is calculated with the equation (4).

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1}(Pi-Ri)^2}{n}} \quad (4)$$

The results achieved are shown in the Results section.

20.7	NA	19.5	NA	20.3	NA	21.2	...
20.7	NA	NA	19.2	NA	NA	21.2	...
20.7	NA	NA	NA	20.3	NA	NA	...
20.7	NA	NA	NA	NA	19.1	NA	...
20.7	NA	NA	NA	NA	NA	21.2	...

...

Fig. 6. Distribution of NA Values in Predicted Time Series.

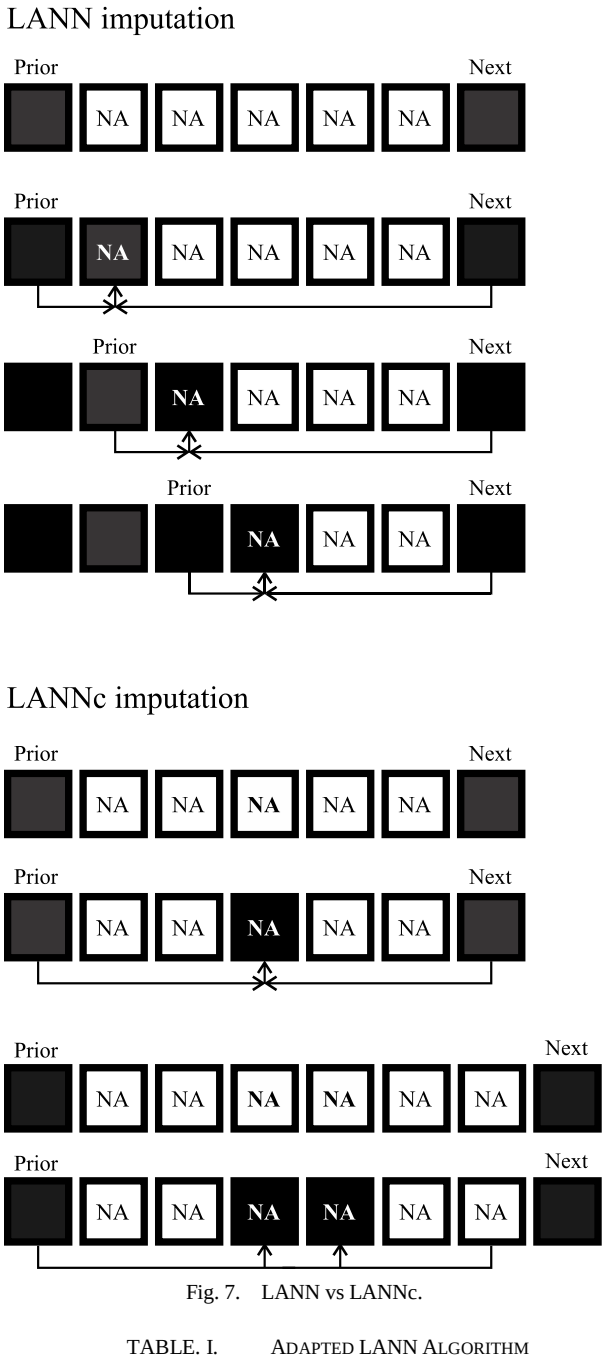


Fig. 7. LANN vs LANNc.

TABLE. I. ADAPTED LANN ALGORITHM

```
function lann(tsna,pos)
{   npos=pos.length;
    for(i=0;i<npos;i++)
    {   res=getPriorNext();
        prior=res[0];
        next=res[1];
        base=(prior+next)/2;
        tsna[pos[i]]=base.toFixed(2);
    }
    return tsna;
}
```

V. RESULTS

This section shows the results achieved after experimentation. The LANN and LANNc algorithms were implemented with different configurations of NA values between 1 and 11 as it shown in Table II and Table III with the respective RMSE values.

According to Table II and Fig. 8, it is appreciated that on average, from the eleven (11) configurations experienced for LANN, all of them improve the RMSE of LSTM. The best configuration corresponds to six consecutive NA values (RMSE 0.6577).

According to Table III and Fig. 9, it is appreciated that on average, from the eleven (11) configurations experienced for LANNc seven (7) allow to improve the RMSE of LSTM, so the best configuration corresponds to two consecutive NA values (RMSE 0.6606).

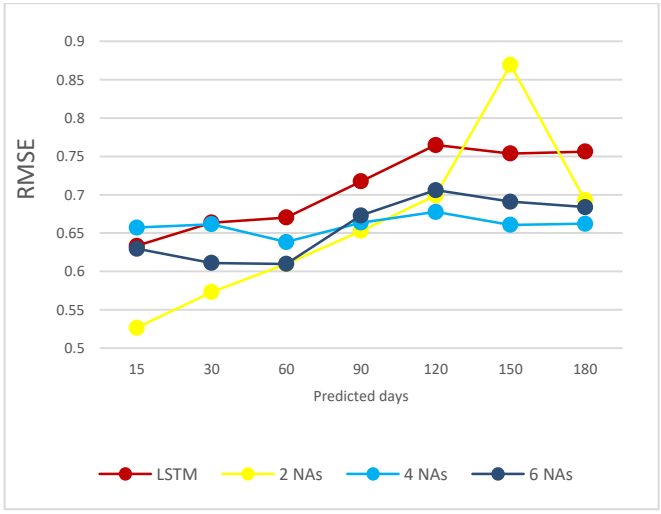


Fig. 8. LSTM vs Top 3 LANN.

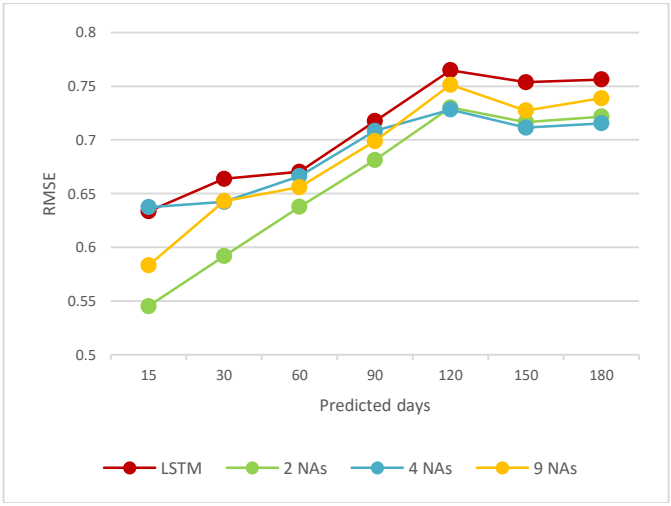


Fig. 9. LSTM vs Top 3 LANN.

TABLE. II. LSTM VS LANN

Technique		RMSE of Predicted Days							Avg
		15	30	60	90	120	150	180	
LSTM		0.6334	0.6637	0.6702	0.7175	0.7649	0.7537	0.7562	0.7085
LANN	1	0.6515	0.6523	0.6428	0.7003	0.7575	0.7421	0.7384	0.6978
	2	**0.5265**	**0.5730**	**0.6096**	**0.6527**	0.6990	0.8695	0.6929	0.6604
	3	0.6377	0.6352	0.6328	0.6686	0.6961	0.6858	0.6877	0.6634
	4	0.6572	0.6616	0.6383	0.6636	0.6776	0.**6606**	0.6622	0.6601
	5	0.6389	0.6550	0.6636	0.6691	0.**6880**	0.6697	**0.6553**	0.6628
	6	0.6296	0.6111	0.6097	0.6730	0.7059	0.6910	0.6838	**0.6577**
	7	0.6649	0.7564	0.7054	0.7045	0.7258	0.7011	0.6954	0.7076
	8	0.5896	0.6610	0.6706	0.7122	0.7059	0.6813	0.6723	0.6704
	9	0.5831	0.6612	0.7201	0.7373	0.7294	0.7046	0.6749	0.6872
	10	0.6396	0.6810	0.7189	0.7131	0.7457	0.7234	0.7149	0.7052
	11	0.6185	0.6484	0.7180	0.7050	0.6992	0.6837	0.6735	0.6780

TABLE. III. LSTM VS LANNC

Technique		RMSE of Predicted Days							Avg
		15	30	60	90	120	150	180	
LSTM		0.6334	0.6637	0.6702	0.7175	0.7649	0.7537	0.7562	0.7085
LANNc	1	0.6515	0.6523	0.6428	0.7003	0.7575	0.7421	0.7384	0.6978
	2	**0.5452**	**0.5918**	**0.6377**	0.**6813**	0.7302	0.7166	0.7216	**0.6606**
	3	0.6586	0.6360	0.6753	0.7143	0.7435	0.7294	0.7357	0.6989
	4	0.6374	0.6422	0.6663	0.7084	0.**7282**	0.**7113**	0.**7154**	0.6870
	5	0.6186	0.6430	0.7073	0.7164	0.7673	0.7475	0.7362	0.7051
	6	0.6388	0.6414	0.7101	0.7352	0.8015	0.7833	0.7681	0.7254
	7	0.6197	0.7018	0.6638	0.7424	0.7949	0.7815	0.7840	0.7268
	8	0.5876	0.6081	0.7546	0.7708	0.7935	0.7621	0.7552	0.7188
	9	0.5831	0.6430	0.6559	0.6989	0.7513	0.7273	0.7388	0.6854
	10	0.6090	0.6617	0.7010	0.7351	0.7623	0.7322	0.7226	0.7034
	11	0.5926	0.6531	0.7653	0.7616	0.8357	0.7914	0.7681	0.7382

VI. DISCUSSION

According to Table IV and Fig. 10, it can be seen that, on average, the best configuration of the proposed LANN and LANNc algorithms allowed to overcome the techniques mentioned in the Related Work section. It should be noted that GRU in forecasting of time series, in general, in most cases outperforms LSTM, it can be seen in [15] [14] [19] [20]. However, by applying any of the two imputation techniques mentioned in this work, on average it was possible to improve LSTM predictions and overcome the results achieved by GRU.

TABLE. IV. COMPARISON WITH ANOTHER TECHNIQUES

Technique	RMSE of Predicted Days							Avg
	15	30	60	90	120	150	180	
LSTM	0.6334	0.6637	0.6702	0.7175	0.7649	0.7537	0.7562	0.7085
LANN*	0.6296	0.6111	0.6097	0.6730	0.7059	0.6910	0.6838	**0.6577**
LANNc**	0.5452	0.5918	0.6377	0.6813	0.7302	0.7166	0.7216	0.6606
PROPHET	0.5512	0.7054	1.0516	1.1637	1.1274	1.1274	1.0403	1.0279
GRU	0.5953	0.6917	0.6678	0.6689	0.7076	0.6751	0.6727	0.6684
ARIMA	0.6134	1.2988	2.2932	2.5240	2.2320	2.2320	2.0440	2.1639

* 6 NAs ** 2 NAs

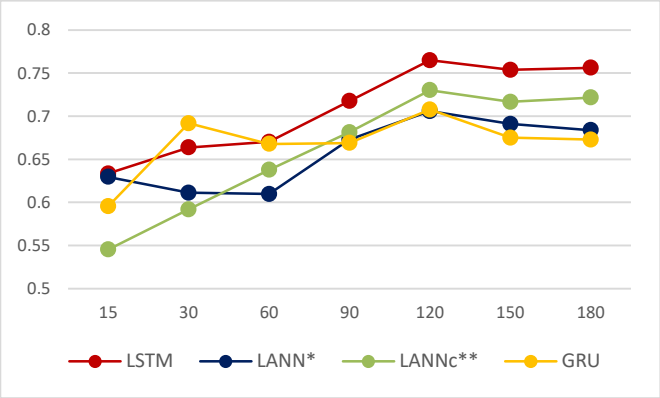


Fig. 10. Comparison of LSTM, LANN, LANNc and GRU.

VII. Conclusions

The use of Imputation techniques based on Local Average of Nearest Neighbors allowed to improve the prediction results of LSTM by exceeding on average the prediction results of GRU and other state of the art techniques.

Despite the risk of inserting bias to the left of the gap of NA values, LANN obtained a better performance than LANNc and in all NA cases it outperformed or improved LSTM predictions.

VIII. Future Work

The use of imputation techniques on the results of prediction techniques to improve their accuracy, opens a new research line that will improve current techniques. At this point, it is possible to experiment with other imputation techniques such as SMA [21], LWMA [21], EWMA [21], ARIMA-Kalman [22] [23], CBRi [6], CBRm [5], HSV [24], LANNf [24], etc.

In this work the results of LSTM were improved, it would be important to analyze how much the results of GRU or other prediction techniques can be improved through imputation techniques.

References

[1] Y. LeCun, Y. Bengio & G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.

[2] R. Pascanu, T. Mikolov, Y. Bengio, "On the dificulty of training recurrent neural networks," de 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.

[3] C. Kyunghyun, V. Bart, G. Caglar, B. Dzmitry, B. Fethi, S. Holger & B. Yoshua, "Learning phrase representations using RNN enconder-decoder for statistical machine traslation," arxiv.org, pp. 1-15, 2014.

[4] A. Flores, H. Tito and C. Silva, "Local average of nearest neighbors: univariate time series imputation," International Journal of Advanced Computer Science and Applications, vol. 10, nº 8, pp. 45-50, 2019.

[5] A. Flores, H. Tito & C. Silva, "CBRm: case based reasoning approach for imputation of medium gaps," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 10, nº 9, pp. 376-382, 2019.

[6] A. Flores, H. Tito & C. Silva, "CBRi: a case based reasoning-inspired approach for univariate time series imputation. In Press," de 6th IEEE Latin American Conference on Computational Intelligence LA-CCI, Guayaquil, Ecuador, 2019.

[7] S. Kavitha, S. Varuna and R. Ramya, "A comparative analysis on linear regression and support vector regression," de Online International Conference on Green Engineering and Technologies, Coimbatore, India, 2016.

[8] R. Hyndman & G. Athanasopoulos, Forecasting: principles and practice, Melbourne, Australia: OTexts, 2018.

[9] J. Palomares, J. de la Rosa, J. Ramiro, J. Melgar, A. Agüera & A. Moreno, "ARIMA vs neural networks for wind speed forecasting," de International Conference of Computational Intelligence for Measurement Systems and Applications, Hong Kong, China, 2009.

[10] Y. Pan, M. Zhang, Z. Chen, M. Zhou, Z. Zhang, "An ARIMA based model for forecasting the patient number of epidemic disease," de 13th International Conference on Service Systems and Service Management, Kunming, China, 2016.

[11] M. Abdullah, A. Hoque, "Comparison of ARIMA and SVM for short-term load forecasting," de Annual Information Technology, Electromechanical Engineering and Microelectronics Conference, Jaipur, India, 2019.

[12] S. Taylor & B. Letham, "Forecasting at scale," PeerJPreprints, pp. 1-25, 2017.

[13] W. Gail, G. Yoav & Y. Eran, "On the practical computational power of finite precision RNNs for language recognition," arxiv.org, pp. 1-9, 2018.

[14] S. Kumar, L. Hussain, S. Banarjee & M. Reza, "Energy load forecasting using deep learning approach-LSTM and GRU in spark cluster," de Fifth International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, 2018.

[15] R. Fu, Z. Zhang & L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," de 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 2016.

[16] J. Zheng, X. Chen, K. Yu, L. Gan, Y. Wang & K. Wang, "Short-Term power load forecasting of residential community based on GRU neural network," de International Conference on Power System Technology, Guangzhou, China, 2018.

[17] L. Kuan, Z. Yan, W. Xin, S. Wenxue, J. Zhe, Z. Young, X. Nan, Z. Xing, "Short-term electricity load forecasting method based on multilayered self-organizing GRU network," de IEEE Conference on Energy Internet and Energy System Integration, Beijing, China, 2017.

[18] M. Paco, C. López Del Alamo & R. Alfonte, "Forecasting of meteorological weather time series through a feature vector based on correlation," de 18th International Conference Computer Analysis of Images and Patterns CAIP 2019, Salerno, Italy, 2019.

[19] B. Wang, W. Kong, H. Guan & N. Xiong, "Air quality forecasting based on gated recurrent long short term memory model in internet of things," IEEE Access, vol. 7, pp. 69524 - 69534, 2019.

[20] A. Tokgoz & G. Unal, "A RNN based time series approach for forecasting turkish electricity load," de 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018.

[21] S. Moritz, "Package ImputeTS," cran.r-project.org, 2019.

[22] P. Zarchan, H. Musoff, Fundamentals of kalman filtering, American Institute of Aeronautics and astronautics, 2000.

[23] S. Moritz, T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation in R," The R Journal, vol. 9, nº 1, pp. 207-2018, 2017.

[24] A. Flores, H. Tito & D. Centty, "Model for time series imputation based on average of historical vectors, fitting and smoothing," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 10, nº 10, pp. 346-352, 2019.

Data Sanitization Framework for Computer Hard Disk Drive: A Case Study in Malaysia

Nooreen Ashilla Binti Yusof[1], Siti Norul Huda Binti Sheikh Abdullah[2], Monaliza Binti Sahri[5]
Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia

Mohamad Firham Efendy bin Md Senan[3]
Nor Zarina binti Zainal Abidin[4]
Digital Forensic Department, CyberSecurity Malaysia
Level 5, Tower 1, Menara Cyber Axis, Jalan Impact
63000 Cyberjaya, Selangor, Malaysia

***Abstract*—In digital forensics field, data wiping is considered one of the anti-forensics' technique. On the other perspective, data wiping or data sanitization is the technique used to ensure that the deleted data are unable to be accessed by any unauthorized person. This paper introduces a process for data sanitization from computer hard disk drive. The process was proposed and tested using commercial data sanitization tools. Multiple testing has been conducted at accredited digital forensic laboratory of CyberSecurity Malaysia. The data sanitization process was performed using overwritten method provided by state-of-the-art data sanitization tools. For each sanitization tool, there are options for the wiping technique (overwritten) process. The options are either to wipe using single pass write or multi pass write. Logical data checking in the hard disk sector was performed during pre and post data disposal process for a proper verification. This is to ensure that the entire sector has been replaced by data sanitization bit pattern in correspondence to the selected wiping technique. In conclusion, through the verification of data sanitization it will improve the process of ICT asset disposal.**

Keywords*—*Data sanitization; anti-forensics; digital forensics; wiping technique

I. Introduction

Data security is nation major concern to curb data breach recklessly [1]. Several approach to handle security breach such as the introduction of steganography and cryptography [2] were tested. These techniques help to improve quality of the data and security [2] during data acquisition for forensic analysis. However, to protect data on media storage such as hard disk drive and magnetic media is depends on two recovery methods [3] [4], are: (i) hardware and (ii) software. The challenges faced on data protection techniques is data privacy and manipulation [1] of data which may lead to privacy violations such as electronic crime [5] for hard disk drive users.

The delete and format method are commonly used by users to dispose their data [6] from hard disk drive. Both methods of data disposal are provided by the computer operating system [6], [7]. Previous researchers [7], [8], [9], [10] suggested the data disposal by using data sanitization methods instead of the delete and format method before the hard disk drive been discarded or sold. Currently, the data sanitization follows the NIST Special Publication 800-88 (Revision 1) Guidelines for Media Sanitization but there is no process of data sanitization from hard disk drive that can be referred online. Hence, this paper introduces the data sanitization process including verification process to dispose the data from the hard disk drive.

The number of overwrite pass becomes an issue to researchers as some of them said single pass overwrite is sufficient [6] to delete the data in the hard disk drive. In fact, the overwriting method proposed by NIST 800-88 to dispose of the data from the hard disk drive is simply to perform at least single pass overwrite with a fixed data value, such as all null [7]. The controversial of the number of overwrite pass exists because some say that data can be recovered if the method of data deletion is done using single or two passes overwrite [6]. Therefore, the purpose of this study is to conduct the number of overwrite pass take in order to completely remove the data from the hard disk drive and secure from data recovery activities.

There are two objectives highlight in this research: i) to identify the data disposal methods from the hard disk drives to be disposed, and ii) to propose the best practise or standard of procedure for data disposal logically from hard disk drive that secure from any data recovery activities, easy-to-understand, follow the data sanitization guideline and standards. We organize this paper into five sections: Introduction, Research background, Methodology, Experimental Results and Analysis, and Conclusion and Future Work.

II. Research Background

There are four methods for deletion of data from digital devices which are: i) delete, ii) format, iii) data sanitization, and iv) physically destroying the media [7], [11], [12]. A generic approach for enterprises to automatically sanitize sensitive data in images and documents is browser-based cloud storage. CloudDLP utilizes deep learning methods to detect sensitive information in both images and textual documents [13]. FlashGhost [14], which is a software that apply a novel integration of cryptography techniques with the frequent colliding hash table is used to erase data in both client and server site. Consequently, data will be unreadable and unrecoverable by overwriting multiple times after its validity period has expired. In a shared IoT environment, [15] a data sanitization approach is present via hierarchical-cluster method to hide confidential information while finding the useful and meaningful information in the sanitized dataset

using multi-objective particle swarm optimization framework. Formatting [7] is the most common approach adopted to remove data in a hard disk. This is based on an interview among users where the results show that users are more prone to use the format method. The study on the comparison has been done [6], [8] between the three methods of deleting, formatting and sanitizing the data. A study [8] shows that the delete and format method are not safe due to lower security [8] compared to the method of sanitizing data which possess better security. Based on [7], [8], [9], [10] safest way to wipe data is by using sanitizing method. The sanitization of data is safer as both file navigator and the file are completely deleted from the media storage [8] by using the overwrite method by replacing the data with zero and one bit.

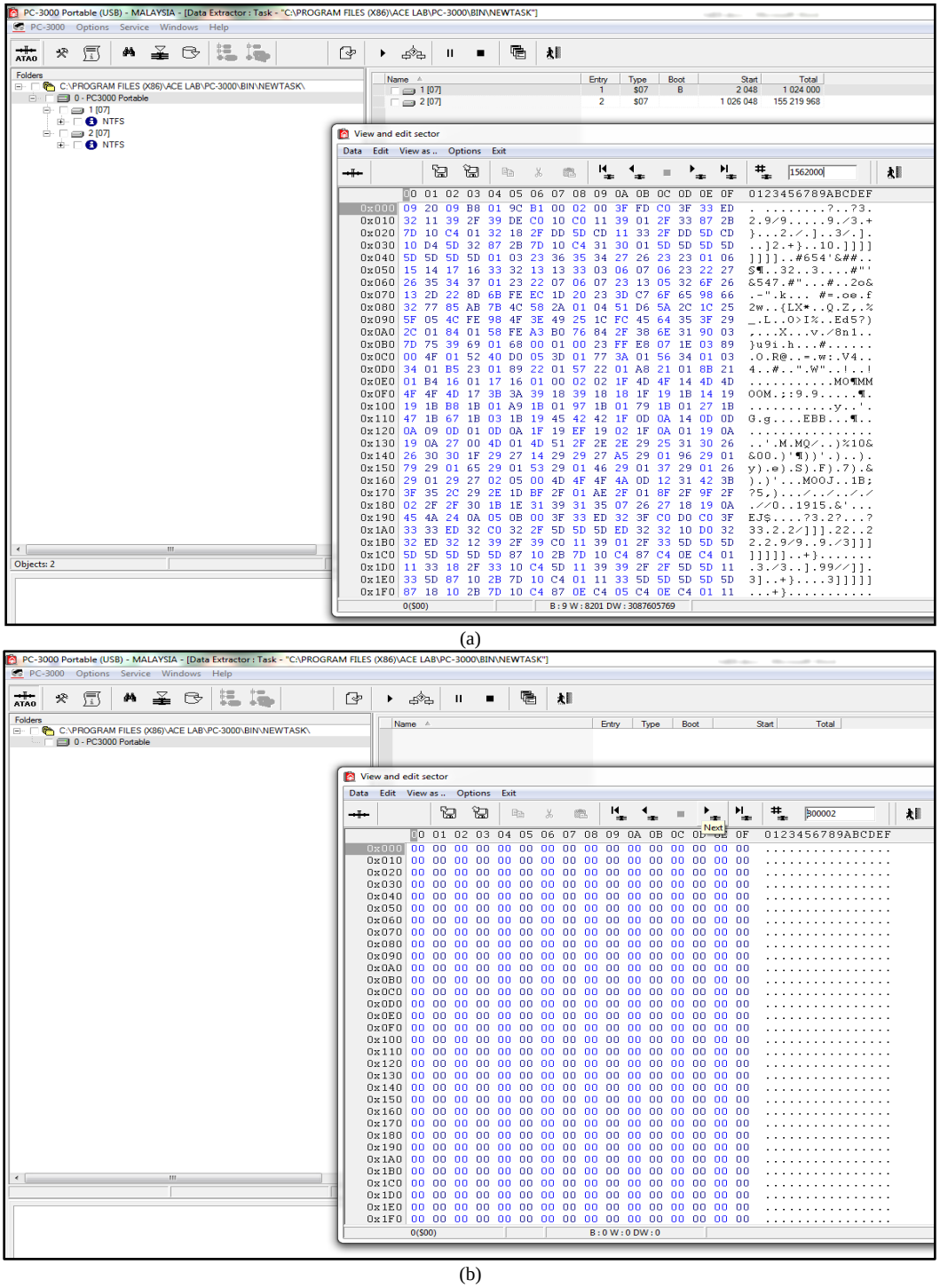


Fig. 1. The Pattern of Logical Data before Data Sanitization using Overwrite Technique. (b) The Pattern of Logical Data after Data Sanitization using Overwrite Technique.

There are a few standards that need to be followed in handling data sanitization [9] according to Department of Defense of United States of America. The first standard in handling data sanitization is DoD 5520 [9] from Department of Defense of United States of America. DoD 5520 states that clearing and sanitization matrix needs two ways to overwrite the data followed by validation [8]. However, revolution of hard disk renders DoD 5220 irrelevant as the first two ways did not function well and need to be updated [8], [9]. Next, DoD 5520.22-M is introduced to replace DoD 5220 in 1995 [12], [16]. Nevertheless, DoD 5220.22-M is not recognized by the Department of Defense of United State of America as the standard to sanitize the hard disk [12] and did not approve it as a secure method for hard disk drive [12]. DoD 5520.22-M is then improved to DoD 5220.22-M ECE in 2001 [12] by increasing the repetition of overwrite operation to seven times. However, DoD 5220.22-M ECE is also not applied as much as a single overwrite pass is already sufficient instead of overwriting it multiple times [12] to sanitize hard disk. Department of Defense of United State of America then replaced DoD 5220.22-M ECE with NIST 800-88 in 2006 [8]. NIST 800-88 emphasizes the method to sanitize the hard disk and other storage media electronic with minimum requirement of implementation of sanitization of media. NIST 800-88 [17] is rechecked again and improved to NIST 800-88 [11] in 2014. It has been used until today. NIST 800-88 [11] states that verification for each data sanitization process is a necessity. The use of tools or software for data sanitization that has been verified and recommended by organization [11]. There are two methods that can be used for verification by either doing: i) verification each time the re-write is performed, or ii) verification on sampling technique.

Data sanitization in theory is referring to overwrite technique [11], [18] to a logical data. Data sanitization method through overwrite technique, will overwrite all the existing logical data bit pattern with a fixed non-sensitive logical data bit pattern on the entire hard disk drives [11]. The following Fig. 1 shows the logical checks on the sectors in the hard disk drive before (Fig. 1(a)) and after the disposal of logical data (data sanitization) is executed (Fig. 1(b)).

In 1996 [10], Gutmann states theoretically, that data which has been overwritten can be restored by using specialized technique such as the magnetic force microscopy (MFM) technique. To avoid data recovery, Gutmann suggests that the secure data deletion method by overwrite pass is executed 35 times in order to make it difficult for recovery process. The overwrite method uses random passes before and after the erase process and in between it the bit is replaced with pattern as shown in Table I.

The Gutmann method takes too much time [6] [19] because the data needs to be overwritten up to 35 pass to delete the data. In 2014, NIST 800-88 [11] outline the minimum number of overwrite pass stated in the Minimum Sanitization Recommendation as in Table II.

NIST 800-88 [11] emphasizes that single pass is sufficient to prevent the deleted data from any data recovery attempts.

TABLE. I. GUTMANN THEORY [10]

Number of Overwrite Pass	Logical Data Bit Pattern
1-4	random value
5	01010101 01010101 01010101
6	10101010 10101010 10101010
7-9	10010010 01001001 00100100 01001001 00100100 10010010 00100100 10010010 01001001
10	00000000 00000000 00000000
11	00010001 00010001 00010001
12	00100010 00100010 00100010
13	00110011 00110011 00110011
14	01000100 01000100 01000100
15	01010101 01010101 01010101
16	01100110 01100110 01100110
17	01110111 01110111 01110111
18	10001000 10001000 10001000
19	10011001 10011001 10011001
20	10101010 10101010 10101010
21	10111011 10111011 10111011
22	11001100 11001100 11001100
23	11011101 11011101 11011101
24	11101110 11101110 11101110
25	11111111 11111111 11111111
26-28	same as overwrite pass 7-9
29-31	01101101 10110110 11011011 10110110 11011011 01101101 11011011 01101101 10110110
32-35	random value

TABLE. II. NUMBER OF OVERWRITE PASS RECOMMENDED BY NIST 800-88 [11]

Number of Overwrite Pass	Logical Data Bit Pattern
1	00000000 00000000 00000000

III. METHODOLOGY

There are two objectives to be achieved in this methodology, namely to propose a data sanitization framework and to perform a detail analysis on data disposal methods. Before proposing the data sanitization framework, a questionnaire is distributed to few agencies to gather information on how the agencies handle hard disk drive disposal and the current awareness on the importance of data sanitization. The data gathered is presented on the experimental results and analysis section. This questionnaire helped in proposing a better and secure data sanitisation framework. After that data disposal methods and data sanitization process will be tested and analysed. After the data sanitization process has been run and tested, an expert assessment and user acceptance is conducted by the expert from Cyber Security Malaysia (CSM) to verify the process.

Section 3.1, 3.2 and 3.3 illustrate the detailed steps taken for the methodology.

A. Questionnaire

There are two sets of questionnaires: Set A and Set B. The questionnaires were adapted from eight subsection under Section 4 Information Sanitization and Disposition Decision Making, NIST Special Publication 800-88 [6] Guidelines for Media Sanitization. Two agencies from government administrative agency, and an educational institution were selected as a respondent to this research. The questionnaires were distributed to the implementers (Set A) and supervisors (Set B) that involve in the agency ICT assets disposal.

B. Data Sanitization Tools used in this Research

19 units of hard disk drives to be disposed were obtained from two agencies for this work. All hard disk drives received with different size, manufacturer, type (SATA and IDE) and model [7]. Then, the hard disk drive was connected directly to the sanitization tool for sanitization process and format activity. Table III listed five data sanitization tools provided by CSM being used in this research.

In addition, PC-3000 which is a portable hardware software is used to perform logical checks and to perform data recovery on hard disk drives during this research to make sure that all the data in the drive that have been sanitized is completely removed.

C. Data Sanitization Process from Computer Hard Disk Drive

The proposed logical data disposal process is based on the results obtained from the six tests that have been performed. All the activities and processes are then organized into a proper flowchart. The chart shows the process of wiping the hard disk content logically. This is the proposed process without including the data recovery activity. Fig. 2 shows the proposed process to wipe the computer hard disk drive content logically.

1) Receiving hard disk drive: When a hard disk drive was received, all information such as the hard disk drive number, size, model and connection type (SATA / PATA) of the hard disk drives were recorded. The hard disk drive will then be tested whether it was working fine and readable by the computer. If the hard disk worked perfectly fine and could be read by the computer, then proceeded to next step.

2) Logical review of the content of the hard disk drive before the data disposal is conducted: PC-3000 was connected to a computer and then followed by the hard disk drive connected to the PC-3000 to check content of the logical drive and to check the number of partitions of the hard disk drive before disposing all of the data inside the hard disk drive.

3) Data disposal using commercial data sanitization tools: During the data disposal activity, all the data were wiped logically using special data sanitization tools. The data sanitization tools used provides single pass overwrite and multi pass overwrite. The hard disk drive was wiped using both the single pass overwrite and multi pass overwrite methods to compare the result. Information of the executors, data sanitization tools used, and result are all recorded in detail.

4) Implementation of logical data validation: Each data sanitization tool used had an indicator whether the operation performed successful or not. Nevertheless, NIST 800-88 recommends that verification methods to be performed using other tools than the original equipment used to dispose the data logically [11]. Logical data validation is implemented to ensure that data had been wiped entirely from the hard disk drive, and the existing bit had been overwritten by the data sanitization bit pattern. The validation exercise was based on logical data bit patterns after data disposal was carried out. The logical data validation that were performed in this research used the PC-3000 tools.

5) Formatting the hard disk drive: After the data in the hard disk drive had been disposed logically using the data sanitization tools, the hard disk drive could no longer be detected on the computer. This was because when the data sanitization tools wiping the data inside the hard disk drive and removing all the data, the file system of the hard disk drive was also wiped out. So, the hard disk drive needs to be formatted to re-install the file system. This scenario can be implemented in the agencies if the agencies desire to recycle their hard disk drives instead of disposing it physically.

6) File system review in the hard disk sector: The hard disk drive was then connected to the computer to check whether it was working fine, had a file system and could be read by the computer. If the computer could detect it, the hard disk drive was required to be format again.

7) Documentation: All the activities in Fig. 2 were recorded and then consolidated into a documentation for the revision of data disposal activities from the hard disk drive.

TABLE. III. DATA SANITIZATION TOOLS

No	Tools	Wiping Method	Purpose of Usage
i)	Ninja Y-1659	Erase All	i) Wiping the data logically from the hard disk drive.
ii)	Data Hapus	Sanitize	i) Wiping the data logically from the hard disk drive.
iii)	Tableau TD1	One Pass Write Multi Pass Write	i) Wiping the data logically from the hard disk drive. ii) Format the hard disk drive.
iv)	Tableau TD3	One Pass Write Multi Pass Write	i) Wiping the data logically from the hard disk drive. ii) Format the hard disk drive.
v)	Voom 3P Hardcopy	1 pass 4 pass	i) Wiping the data logically from the hard disk drive. ii) Format the hard disk drive.

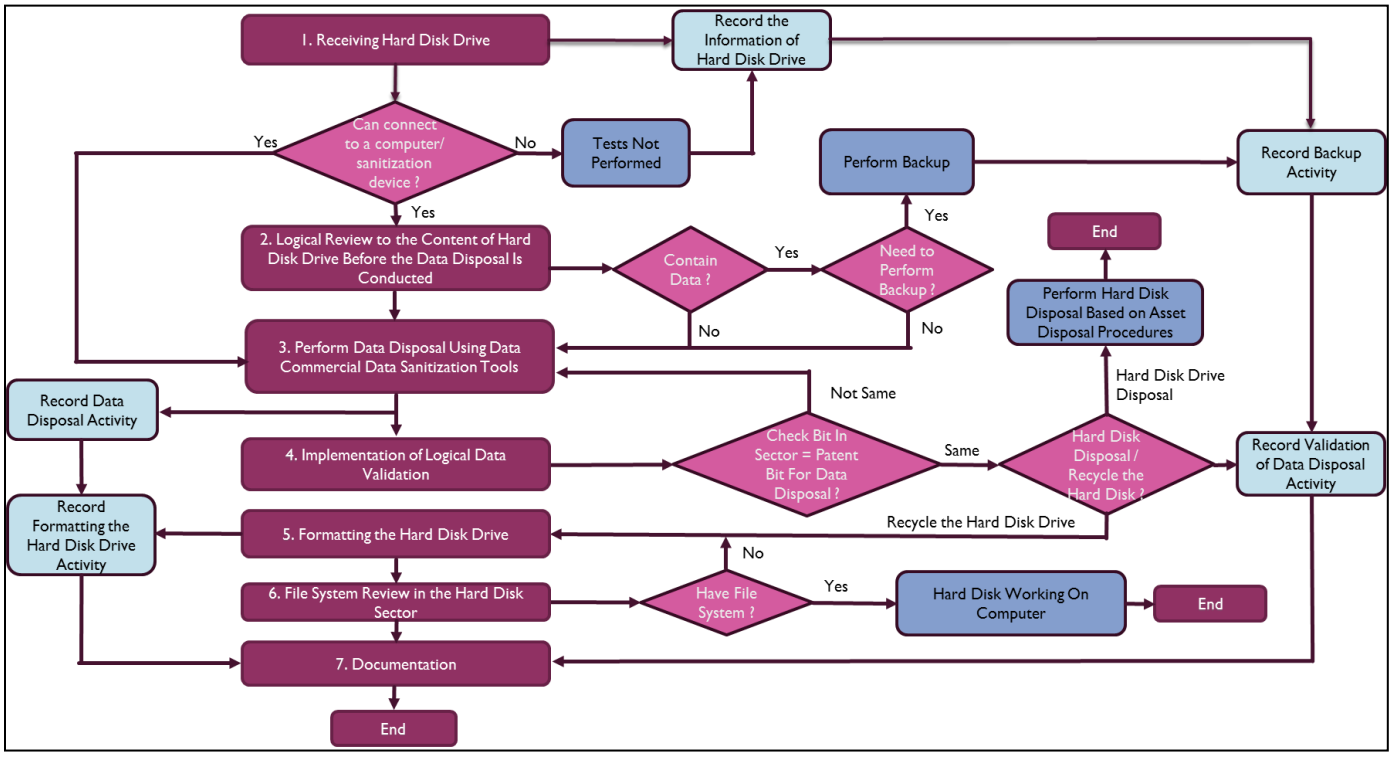


Fig. 2. Proposed Data Sanitization Process from Computer Hard Disk Drive.

IV. Experimental Results and Analysis

We have conducted a pre and post questionnaire for this experiment. The pre-questionnaire was to observe the method of data disposal from hard disk drive of the selected agencies, while the post questionnaire was to observe the relationship of the process of data disposal of hard disk drive between the implementer who conducted the data sanitization process and the supervisor who monitored the data sanitization process at two selected agencies. The result from the questionnaire shows the relationship between the data disposal implementation activities by the implementer and the monitoring activities by the supervisor at Agency B and Agency C and is presented in Table IV.

From the questionnaire, it could be seen that both agencies choose to format the hard disk drive to delete the data prior to disposal of the hard disk drive as opposed to performing a logical data sanitization. Both agencies also did not perform any validation after formatting the hard disk drive.

The analysis also shows that there seems to be discrepancies on the questionnaire between the implementation and monitoring activities in both agencies. Both agencies show a similar percentage of 37.50% and tend to agree to the point that data disposal is a must by the implementers while the percentage on monitoring data disposal exceeds 50.00% in both Agencies B and C. Section A to Section E questionnaire shows agencies implementing data disposal/ protection activities of hard disks drive by formatting it and/ or removing hard disk drives from ICT assets before it is disposed by the agency asset disposition procedure. Disposal of data in the agency does not take into account the level of security of hard disk content even though risk assessment is conducted periodically to determine the level of security of agency assets. Section F to Section H indicates the disposal of data in the agency does not cover the disposal of logical data/ data sanitization. This is because both agencies recognize that they do not have skilled and trained staff to carry out logical data disposal activities from hard disk drives.

The percentage of relationships and gaps between implementers and supervisors for the process of data disposal in both agencies are presented in Fig 3. Fig. 3 shows clear links between the two categories of respondents in data disposal of ICT assets in both agencies. Monitoring activity by the supervisors is at a high percentage for both agencies, which is 50% and 62.50%. Monitoring of data disposal of ICT assets in the agency refers to the method of data disposal using the format method for Agency B and the method of disassembling hard disk pieces for Agency C. All respondents from each agency acknowledged that no logical data disposal was conducted prior to the disposal of ICT assets including hard disk drives in accordance with the asset disposition procedure. The main factor is the absence of skilled and trained staff to carry out logical data disposal activities for hard disk drives. The lack of experts to conduct the logical data disposal makes the percentage of data disposal at both agencies low and equal to 37.50%. This is because the questionnaire developed take into account the good practices of data disposal that need to be taken when disposing data from hard disk based on NIST 800-800 [11].

TABLE. IV. A RELATIONSHIP BETWEEN THE DATA DISPOSAL IMPLEMENTATION ACTIVITIES BETWEEN AGENCIES

Section in Questionnaire	Agency B		Agency C		Average
	Implementer SET A	***Supervisor SET B***	***Implementer SET A***	***Supervisor SET B***	
Section A: Identifying the Needs for Data Disposal	Yes	Yes	Yes	Yes	Yes
Section B: Determination of Information Security Level	**No**	**No**	**No**	Yes	**No**
Section C: Reuse of Hard Disk Drive	Yes	Yes	No	Yes	Yes
Section D: Access Control to the Hard Disk drive	Yes	Yes	Yes	Yes	Yes
Section E: Hard Disk Drive Protection Based on DKICT	**No**	Yes	Yes	Yes	Yes
Section F: Logical Data Disposal of Hard Disk Drive (Data Sanitization)	**No**	**No**	**No**	**No**	**No**
Section G: Validation of Data Disposal	**No**	**No**	**No**	**No**	**No**
Section H: Documentation	**No**	**No**	**No**	**No**	**No**
Percentage Yes (%)	**37.50%**	**50.00%**	**37.50%**	**62.50%**	**50.00%**
Percentage No (%)	**62.50%**	**50.00%**	**62.50%**	**37.50%**	**50.00%**

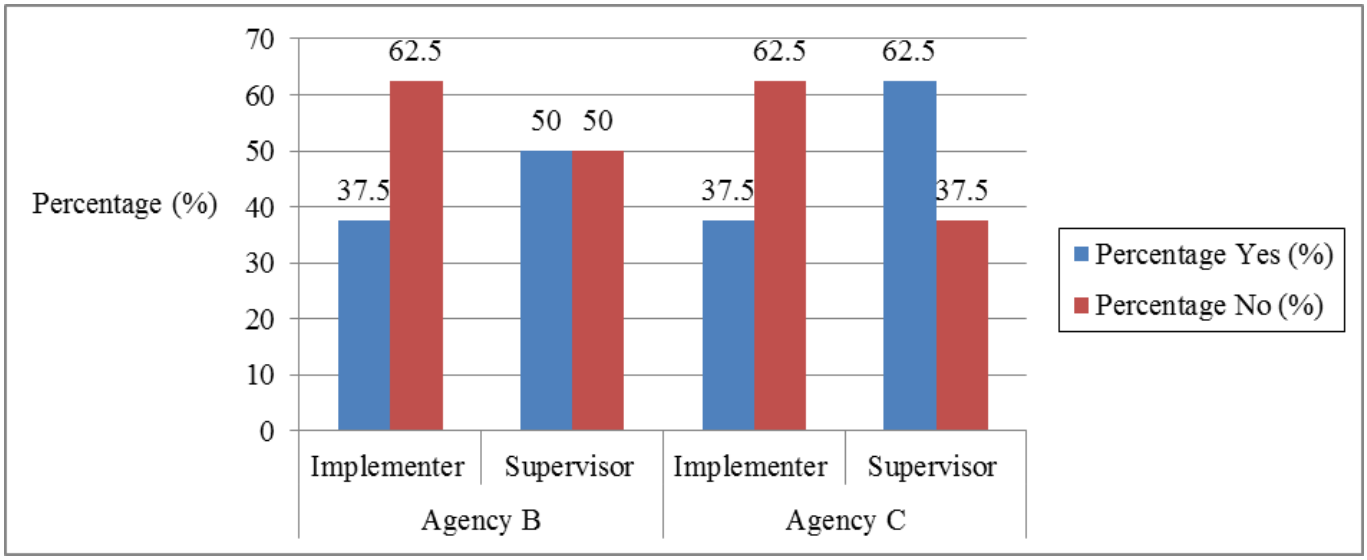


Fig. 3. Analysis Gap between Implementation and Monitoring for the Disposal of ICT Assets in the Agency.

A. Types of Testing

There are five types of tests performed on all hard disk drives received to achieve the second objective, including: i) checking the contents of the logical data before data sanitization, ii) data sanitization using commercial tools, iii) verify whether the logical data has been completely overwritten by bit pattern on the entire hard disk drives, iv) re-format the hard disk drives, and v) data recovery test. All these tests have been conducted at an accredited digital forensic laboratory of CSM using commercial and special hardware and software provided by the Digital Forensics Department. The result on the logical data checks shows that all hard disk drives received, still contains confidential information belong to the agencies which provided the hard disk drives. This study suggested that secure data deletion needs to be done properly by sanitizing the data using overwriting technique. This method wiped the data logically. Data sanitization method through overwrite technique will overwrite all the existing logical data bit pattern to a non-sensitive logical data bit pattern on the entire hard disk drives. Therefore, based on the results and observations from the five tests that have been conducted, a data sanitization process of hard disk drive has been developed. The data recovery test proved that this data sanitization method and process are able to completely wipe all the contents of the hard disk drive. Nineteen hard disk drives provided by selected agencies from

government sector and educational sector are used in the tests. Table V shows that from 19 unit hard disks, two of them still contain data and another two hard disk drives cannot be checked due to hard disk damage and connection failure. Total numbers of hard disk drives used for data sanitization experiment are 17 units.

After the data sanitization process was implemented on all the functioning hard disk drives, the process was then assessed by experts and a user acceptance test was conducted.

B. *Expert Assessment and user Acceptance on the Developed Data Sanitization Process*

This section shows the result of expert assessment and user acceptance analysis.

1) Expert assessment: The proposed data sanitization process had been presented to four experts from the field of digital forensic and data recovery from CSM. Evaluation was based on the presentations and the results of the research presented. The experts then gave their evaluation and feedback on the Expert Assessment Form provided. The evaluations of the four experts were analyzed based on the justifications and feedback options provided on the Expert Assessment Form. This form was filled by each panel with representatives as shown in Table VI.

All four experts agreed 100% on the justification provided in the Expert Assessment Form. The expert agreed that this proposed logical data disposal process complied with technical specifications of the logical data disposal from hard disk drive. It is also successfully adapted to the existing asset disposition procedure in Malaysia, and is suitable for continued implementation at the agency to protect the confidentiality of the information in the hard disk drive to be disposed of.

2) User acceptance: Three (3) agencies were selected for this user acceptance: i) Agency A, ii) Agency B, and iii) Agency C. Agency A refers to government laboratory that provides data sanitization services to the government agencies. They test and run the proposed data sanitization process and fill their evaluation in User / Agency Acceptance Form. The test results show that this proposed process helps them to run data sanitization activities work more systematically. User acceptance for Agency B and Agency C were based on briefing and without actual training. Three (3) agencies were selected for this user acceptance: i) Agency A, ii) Agency B, and iii) Agency C. Agency A refers to government laboratory that provides data sanitization services to the government agencies. They test and run the proposed data sanitization process and fill their evaluation in User / Agency Acceptance Form. The test results show that this proposed process helps them to run data sanitization activities work more systematically. User acceptance for Agency B and Agency C were based on briefing and without actual training. This was because both agencies did not have special data sanitization equipment and tools but they performed ICT asset disposal. The result of analysis and assessment on user acceptance for these three agencies shows in Table VII.

Analysis shows that five out of seven respondents accept 100% with the provided seven justifications of acceptance for the proposed data sanitization process. Two more respondents from Agency C accept 85.71% and 28.57% based on justifications of acceptance options provided in the User/ Agency Acceptance Form. This disadvantage may contribute to a lesser percentage of respondents C2 and C3 from Agency C.

TABLE. V. HARD DISK DRIVE USED FOR THE TESTING PHASE

Item	Government Sector	Educational Institution	Total
Total number of hard disk drives received	15	4	19
Number of hard disks that still have undeleted files/ unformatted	1	1	2
Number of hard disks damaged	2	0	2
Total hard disk used for testing	13	4	17

TABLE. VI. THE DEVELOPED INSTRUMENT FOR EXPERT ASSESSMENT FORM

	Justification (Agree or Disagree)
i.	This proposed data sanitization process is suitable to implement at all agencies.
ii.	This proposed data sanitization process is easier to understand and to implement.
iii.	The proposed data sanitization process is based on the guideline for media sanitization NIST 800-88.
iv.	The testing has been conducted on the actual hard disk drive that agency wanted to disposed
v.	The proposed data sanitization process including the verification process.
vi.	The proposed data sanitization process has been tested with data recovery testing.
vii.	The proposed data sanitization process including documentation activities.
viii.	The proposed data sanitization process makes the data disposal activity more organized based on the checklist provided.
ix.	Data disposal from hard disk drive using special data sanitization tools is easier.
x.	The proposed data sanitization process and documentation can be applied to the existing government ICT asset disposal process
xi.	Employees who involved with logical data disposal do not require high skills to use the proposed data sanitization tools.

TABLE. VII. USER ACCEPTANCE ANALYSIS

Justification for User Acceptance	Agency A	Agency B			Agency C			
	A1	*B1*	*B2*	*B3*	*C1*	*C2*	*C3*	*C4*
i. This proposed data sanitization process is very useful at my agency.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ii. This proposed data sanitization process is easier to understand and to implement.	Yes	Yes	Yes	Yes	Yes	Yes		Yes
iii. The proposed data sanitization process including the verification process.	Yes	Yes	Yes	Yes	Yes	Yes		Yes
iv. The proposed data sanitization process including documentation activities.	Yes	Yes	Yes	Yes	Yes	Yes		Yes
v. The proposed data sanitization process makes the data disposal activity more organized based on the checklist provided.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
vi. Data disposal from hard disk drive using special data sanitization tools is easier.	Yes	Yes	Yes	Yes	Yes	Yes		Yes
vii. Employees who involved with logical data disposal do not require high skills to use the proposed data sanitization tools.	Yes	Yes	Yes	Yes	Yes			Yes
Percentage (%) of user acceptance based on the justifications provided	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**85.71%**	**28.57%**	**100.00%**

Based on the analysis on Table VII, Agency A fully accepted the proposed data sanitization process since they are given the chance to fully test and run the data sanitization process. Hence, Agency A can fully experience and understand the process of data sanitization which aids in the acceptance of the proposed data sanitization process. Next, Agency B also fully accepted the proposed data sanitization method only based on the briefing without experiencing the data sanitization process and this is because they are experienced in ICT but they do not have the proper data sanitization tools. Other than that, Agency C, two of the respondents which is C2 and C3 respectively do not fully satisfied the proposed data sanitization process due to the lack of experience along with the absence of the data sanitization tools in the agencies. Besides that, other factor that might contribute to not fully accepted the data sanitization proposed method is because the respondents are not technical groups.

V. CONCLUSION AND FUTURE WORK

This research has found that logical data disposal methods using data sanitization methods by overwriting existing logical data bits into data sanitization bit patterns able to remove data permanently and securely from hard disk drives including file system. From the previous research [3], data can be recovered from hard disk drive through files system [3]. Based on the activities and tests carried out in this paper, the proposed logical data disposal framework has been developed in order to guide agencies/ implementers who carry out data disposal from hard disk drives.

This study has achieved both objectives set. The developed data sanitization framework for computer hard disk drive has been reviewed by experts and proved that this framework is comprehensive and secure to protect the deleted contents. In conclusion, content of the hard disk drive can be completely wiped through logical data disposal methods. This technique is carried out through reading and writing techniques using the special data sanitization equipment used in this study. Through this study, it also proves that the proposed data sanitization process is able to completely remove logical data with single pass overwrite. Thus, it may no longer be recovered through data recovery activities.

Therefore, individual and agencies should acknowledge and use this proposed data sanitization framework before discarding the hard disk drive, or before recycling the hard disk drive, or before disposing of the hard disk physically. This is important to protect the confidentiality of sensitive content and information belonging to the individual or agencies from being accessed by the third party.

However, this research was focusing on the hard disk drive from the normal PCs or servers. For future work, the research will look into data sanitization from mobile devices such as tablet PC and handphones for asset disposal process.

ACKNOWLEDGMENT

Special thanks to Universiti Kebangsaan Malaysia for the grant support PP-FTSM-2019 and UKM-AP-2017-005/2. Furthermore, special gratitude to all friends and colleagues who give the comments and support for this work as well as CSM for the used tools and laboratory, and to Center of Information Technology, Universiti Kebangsaan Malaysia (UKM), and two agencies representing government sector and educational institution for providing hard disk drives samples.

REFERENCES

[1] Akbar Khanan, Salwani Abdullah, Abdul Hakim H. M. Mohamed, Amjad Mehmood, and Khairul Akram Zainol Ariffin, “Big Data Security and Privacy Concerns: A Review,” in Proceedings of the 1st American University in the Emirates International Research Conference, 2017, pp. 55–62.

[2] Samar Kamil, Masri Ayob, Siti Norul Huda Sheikh Abdullah, and Zulkifli Ahmad, “Challenges in Multi-Layer Data Security for Video Steganography Revisited,” Asia-Pacific J. Inf. Technol. Multimed., vol. 7, no. 2–2, pp. 53–62, 2018.

[3] Kenan Kalajdzic and Ahmed Patel, “A Fast Practical Method for Recovery of Lost Files in Digital Forensics,” J. Internet Technol., vol. 10, no. 5, 2009.

[4] K. Kalajdzic and A. Patel, “A Fast Scheme for Recovery of Deleted Files with Evidential Recording for Digital Forensics,” in Proceedings of the Fourth International Workshop on Digital Forensics & Incident Analysis, 2009, no. WDFIA, pp. 9–19.

[5] Razana Md Salleh, Masnizah Mohd, and Kamarul Baharin Khalid, “Validation of Digital Forensics Tools for Android Tablet,” J. Inf. Assur. Secur., vol. 9, pp. 19–26, 2014.

[6] A. Al Anhar, M. Gandeva Bayu Satrya ST., and F. A. Yulianto, “Analisis perbandingan keamanan teknik penghapusan data pada hardisk dengan metode DoD 5220.22 dan Gutmann,” in eProceedings of Engineering, 2014, vol. 1, no. 1, pp. 607–613.

[7] S. L.Garfinkel and A. Shelat, “Remembrance of data passed: A study of disk sanitization practices,” IEEE Secur. Priv., pp. 17–27, 2003.

[8] G. F. Hughes, D. M. Commins, and T. Coughlin, “Disposal of disk and tape data by secure sanitization,” IEEE Secur. Priv., pp. 29–34, 2009.

[9] R. Raman and D. Pramod, “A study on data privacy, protection & sanitization practices during disk disposal by Indian educational institutes,” Int. J. Comput. Sci. Issues, vol. 10, no. 2, pp. 1–6, 2013.

[10] K. Sansurooah, H. Hope, H. Almutairi, F. Alnazawi, and Y. Jiang, “An investigation into the efficiency of forensic data erasure tools for removable USB flash memory storage devices,” in Australian Digital Forensics Conference, 2013.

[11] R. Kissel, A. Regenscheid, M. Scholl, and K. Stine, “NIST Special Publication 800-88 (Revision 1) Guidelines for Media Sanitization,” 2014.

[12] R. Stiennon, “Everything You Need to Know About the DoD 5220.22-M Wiping Standard & Its Applications Today,” blancco, 2017. [Online]. Available: https://www.blancco.com/blog-dod-5220-22-m-wiping-standard-method/. [Accessed: 30-Jan-2018].

[13] C. Liu, P. Han, Y. Dong, H. Pan, S. Duan and B. Fang, "CloudDLP: Transparent and Automatic Data Sanitization for Browser-Based Cloud Storage," 2019 28th International Conference on Computer Communication and Networks (ICCCN), Valencia, Spain, 2019, pp. 1-8.

[14] Y. Zhu, S. Yang, W. C. Chu and R. Feng, "FlashGhost: Data Sanitization with Privacy Protection Based on Frequent Colliding Hash Table," 2019 IEEE International Conference on Services Computing (SCC), Milan, Italy, 2019, pp. 90-99.

[15] J. C. Lin, J. M. Wu, P. Fournier-Viger, Y. Djenouri, C. Chen and Y. Zhang, "A Sanitization Approach to Secure Shared Data in an IoT Environment," in IEEE Access, vol. 7, pp. 25359-25368, 2019.

[16] M. Geiger, “Evaluating Commercial Counter-Forensic Tools,” Digit. Forensic Res. Work., vol. 1, pp. 1–12, 2005.

[17] R. Kissel, M. Scholl, S. Skolochenko, and X. Li, “Guidelines for Media Sanitization Reports on Computer Systems Technology,” 2006.

[18] P. Gutmann, “Secure deletion of data from magnetic and solid-state memory,” in Sixth USENIX Security Symposium, 1996, pp. 77–90.

[19] C. Valli and P. Patak, “An Investigation into the efficiency of forensic erasure tools for hard disk mechanisms,” in Proceeding of 3rd Australian Computer, Network & Information Forensics Conference, 2005, pp. 79–83.

Render Farm for Highly Realistic Images in a Beowulf Cluster using Distributed Programming Techniques

Enrique Lee Huamaní[1], Patricia Condori[2], Brian Meneses-Claudio[3], Avid Roman-Gonzalez[4]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—Now-a-days, photorealistic images are demanded for the realization of scientific models, so we use rendering tools that convert three-dimensional models into highly realistic images. The problem of generating photorealistic images occurs when the three-dimensional model becomes larger and more complex, so the time to generate an image is much greater due to the limitations of hardware resources, about this problem is implemented the render farm, which consists in a set of computers interconnected by a high-speed network that provides a strip of the global image distributed in each participating computers with the intention of reducing the processing time of highly complex computational images. The research was implemented in a high-performance Beowulf group of the Universidad de Ciencias y Humanidades using a total of 18 computers. To demonstrate the efficiency of a rendering farm implementation, scalability tests were performed using a 360° equirrectangular model with a total of 67 million pixels, the work is carried out to achieve highly complex renderings in less time to benefit the direction of the research.**

Keywords*—Distributed programming; computational parallelism; Beowulf cluster; high-efficiency computing; render farm

I. Introduction

In the last decades, a great demand began to grow in the generation of high-realism images where three-dimensional models require greater hardware resources to solve renderings of high complexity due to this need the rendering farms are implemented [1] which are groups of computers that they are interconnected with each other through a network that distributes the work of the images to each participating computers to obtain results of photorealistic images in less time.

The problematic of this investigation is the rendering of high realism images that due to the complexity of the project, the rendering processes can take hours or days to return a result of a complex image. Because of this problem, the research direction of the Universidad de Ciencias y Humanidades puts at your disposal the use of the high performance Beowulf cluster implemented by [2], which is located in the embedded systems laboratory. This work uses this architecture for the integration of a rendering farm using the Python programming language that will determine the amount of participating computers to distribute each strip of the image in order to achieve complex renderings in less time.

There are research related to the benefit of using rendering farms in a Beowulf cluster, as in the case of [3] that renders complex images using a total of six computers where it isn't necessary to buy specialized equipment at high costs, other research that try to emphasize the reduction of costs that it has [4] that compare the performance of their existing systems that of the high-ending machines used in the modern animation industry, likewise in this paper, it uses the existing hardware resources. There are different ways to get rendering farms, as is the research of [5] that uses a queue management software to get distributed renderings, instead this project uses distributed programming techniques using Python programming language of which the reader can customize the size of the image by each participating computer. There is also research related to cloud computing, a clear example is [6], which performs a hybrid rendering farm that can be adapted to the server cloud to increase computational capacity, therefore the use of a rendering farm in a Beowulf cluster will achieve greater efficiency in generating complex images and lower costs.

II. Methodology

For the realization of the rendering farms the Beowulf cluster architecture of the embedded systems laboratory of the Universidad de Ciencias y Humanidades, it was used with 17 slave computers and a master computer, all of them with the same hardware characteristics and under the Linux platform with Ubuntu distribution. The main elements that the machines must have are Blender[1], SSH open, ImageMagick and the Python programming language that will divide the original image into small strips of images that will be distributed to each of the participating computers to perform the rendering. Fig. 1 shows the overall architecture of the project.

A. Beowulf Cluster Physical Architecture

Beowulf's high-performance clusters are a group or conglomerates of computers interconnected through a high-speed network that are typically used to apply computational parallelism techniques in order to obtain results in a shorter time [7]. This architecture works with low-cost computers that try to give the similarity of having a supercomputer that is possible thanks to the manipulation of the central processing unit (CPU) therefore, this architecture has a total of 204 cores

[1] Blender Foundation, "blender.org - Home of the Blender project - Free and Open 3D Creation Software" 2019. [Online]. Available: https://www.blender.org/

to perform rendering, Fig. 2 shows the high performance cluster type Beowulf performing a rendering of a prototype of a three-dimensional image.

Each computers used in the Beowulf cluster architecture has the same hardware characteristics, as shown in Table I.

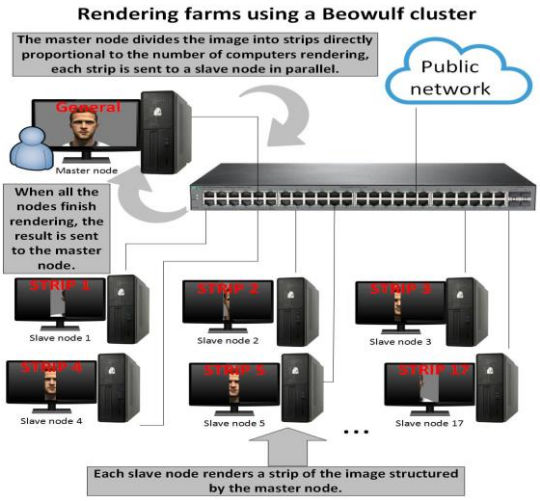


Fig. 1. Design of Rendering Farms with the Beowulf Cluster.

Fig. 2. Render Farm in a Beowulf Cluster.

TABLE. I. HARDWARE CHARACTERISTICS OF THE BEOWULF CLUSTER COMPUTERS

	Description
Modell	HP EliteDesk 800 G1 SFF
HDD	1 TB
RAM	8 GB
Processor	Intel(R) Core (TM) i7-8700
Total Cores	12
Type of Operating System	64-bit
Operative System	Ubuntu 18.04

B. Master Node

The computer in charge of distributing the strips of images to the participating computers is called the master node because it assigns the number of participating computers and the level of complexity to be solved.

C. Slave Node

The computer in charge of obtaining the strip of the image, rendering and returning it to its origin, it is called slave node which complies with the direct orders assigned to it from the master node, the slave and master nodes are interconnected through a high-speed network.

D. Dependencies

The following tools must be taken into account in each of the high-performance cluster machines.

1) Blender: As indicated by [8] blender is an open source 3D modeling and animation software, it is currently used by 2 million visual effects artists, animators and a growing number of astrophysicists. The advantage of using this software is that it is multiplatform and easy to install, it also includes methods for rendering complex surfaces and volumes, image composition, stereoscopic support for tracing graphics and the ability to export models and lighting in a variety of formats.

2) Python: It is an interpreted programming language where its syntax is legible code, this language is multiparadigmic [9] and easy to use, therefore is ideal for distributing the strips of images in each of the slave nodes, thanks to its simplicity the reader can configure and add new features to the algorithm to use it in different ways.

3) ImageMagick: It is a program built in open source C programming language, this software is used to create, edit, compose or convert images into bitmaps [10]. With respect to the project, the program executes the creation of the image in each one of the slave nodes to finally return them to the master node.

4) SSH protocol: Secure Shell (SSH) is a remote administration protocol that allows the reader to initiate commands and copy files from master node to slave nodes [11]. The SSH makes communication between computers secure due to a key generated from the master computer and copied to each of the slave nodes, this allows external devices can not access the Beowulf cluster because they do not have password identification.

E. Process Flow of a Rendering Farm

In the rendering process, distributed programming techniques are applied where a first, a validation is carried out to determine the number of slave nodes, if the condition of having more than one slave computer is met, the distributed rendering is carried out, but if only one computer is assigned, the traditional rendering will be used, in Fig. 3 the detailed rendering steps are explained in the form of a flow diagram.

Because this project is focused on rendering farms, the following distributed programming steps corresponding to the flowchart in Fig. 3 are explained.

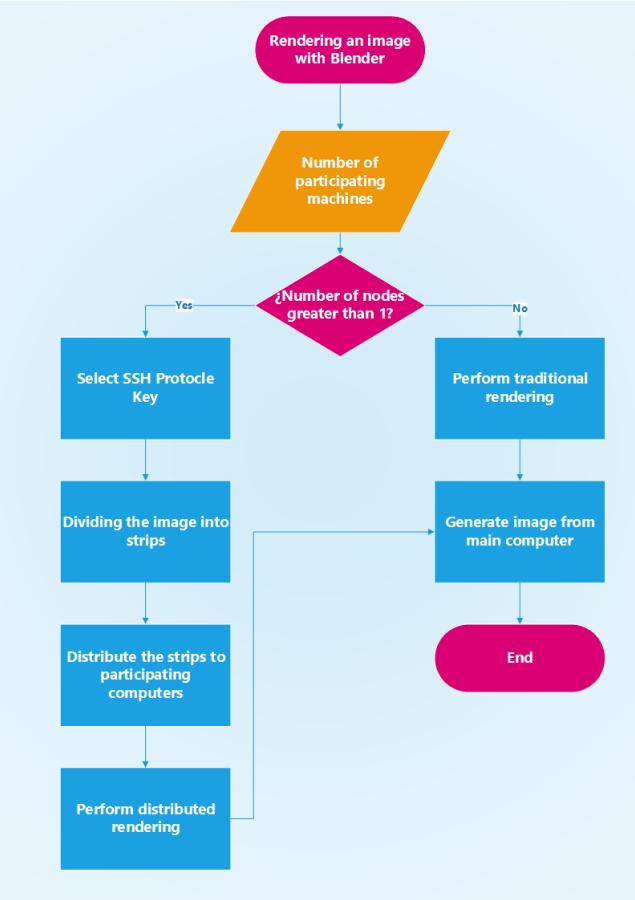


Fig. 3. Render Flowchart.

1) Selection of slave nodes: Before starting the rendering the reader must indicate the number of slave nodes that will participate. From the Linux terminal, it executes the Python code with the following sentence:

blender_network_render.py -m node1 node2 node3 node17

With this sentence, it indicates the number of nodes, in this case 17 slave nodes.

2) Selection of SSH protocol unique key: After classifying the participating nodes, they are assigned their identification key, this key allows unrestricted communication so it is assigned from the command line as following sentence:

blender_network_render.py -R

Which indicates the key that is shared in each of the slave nodes, this is done by security issues.

3) Strips of rendering on slave nodes: The image to be rendered is divided into vertical strips depending on the number of slave nodes assigned. Each strip of the image can take a certain size for rendering, this is useful for slave nodes with different hardware characteristics where the assignment of a strip of image must be inversely proportional to the rendering time of the image. With respect to the machines of the Beowulf cluster, all have the same hardware characteristic therefore the rendering images of high realism are not defined personalized strips for it will be distributed of equitable form as it is shown in the Fig. 4 in which the use of four computers is taken as example.

Depending on the complexity of each strip or on the characteristics in the hardware resources should be allocated an appropriate percentage for better efficiency. It will make the following assumption where 'Computer 0' and 'Computer 2' have fewer cores than 'Computer 1' and 'Computer 3' when the number of cores is not equal between the participating machines should be assigned a custom size of the strip image as shown in Fig. 5.

4) Obtaining the rendered image: Each machine renders a subset of the image as explained in the previous section, when all machines finish part of the render, the program collects all the strips in a final image, it is recommended that all participating machines have the same format with a color depth of 8 to 16 bits with the color specifier RGBA, with this format is possible to assemble images without interruptions, as shown in Fig. 6 to be configured in each slave node.

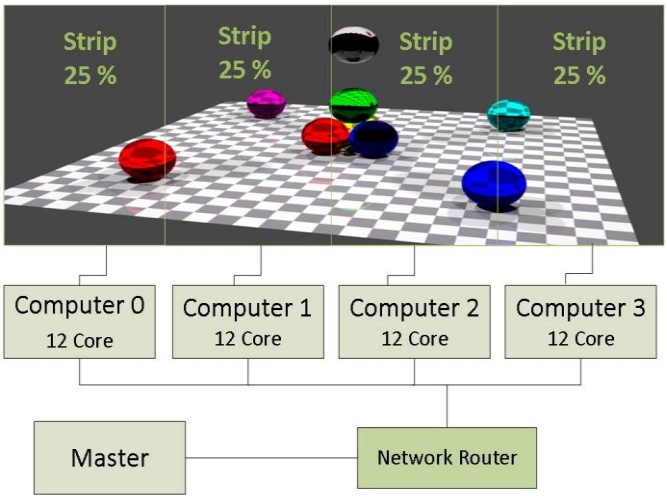


Fig. 4. Equitable Strips of Images.

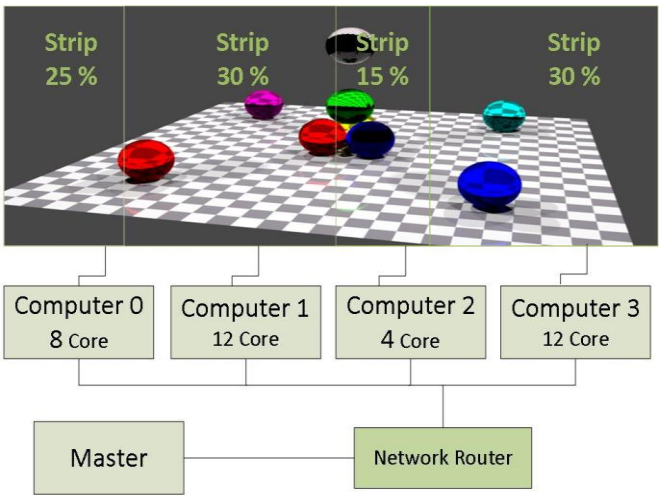


Fig. 5. Strips Images Proportional to the Number of Cores.

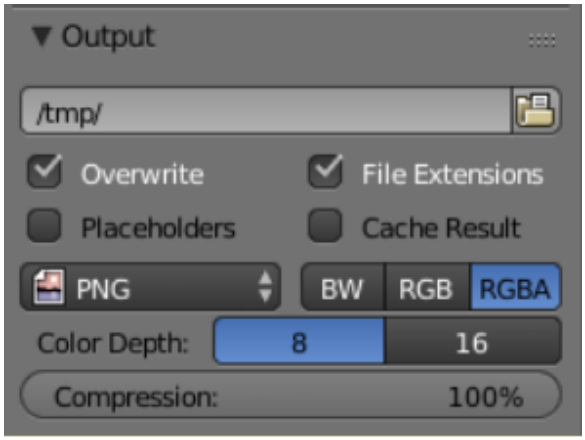


Fig. 6. Output Image Format.

F. Prototype of Rendering Tests

To measure the efficiency of the rendering farms in the Beowulf cluster of the Embedded System laboratory, scalability tests are performed using a large and complex image prototype that is a spatial 3D stereoscopic image with 360° equirectangular projection, which are images used in a dimension of 16384x4096 pixels equivalent to more than 67 million pixels, Fig. 7 shows the prototype designed with Blender software.

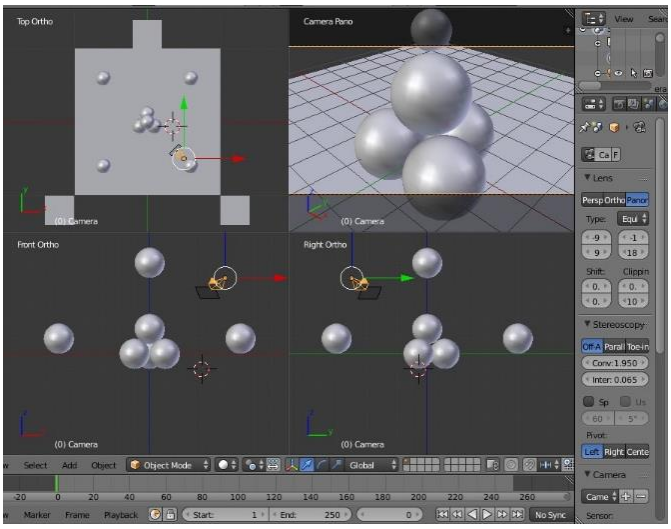


Fig. 7. Prototype of Equirectangular Image of 360°.

III. Result

In this section, the performance tests are carried out using as a project the 360° equirectangular image corresponding to Fig. 6. As shown in Table II, scalability tests are performed where the first column shows the number of participating nodes, the second column shows the time required for rendering and the third column shows the number of cores involved.

Fig. 8 shows graphically the render time with respect to the number of slave nodes involved.

As a result, it provides the image with respect to Fig. 9, which is a realistic 360° image with a dimension of 16384x4096 pixels, which demonstrates the reduction of time for high realism images using distributed programming techniques.

TABLE. II. High Realism Image Rendering Time

Number of Slave Nodes	Time of result	Core
1	1443	12
2	664.8	24
3	673.2	36
4	436.2	48
5	423	60
6	322.2	72
7	321	84
8	257.4	96
9	261	108
10	208.8	120
11	214.8	132
12	190.2	144
13	199.8	156
14	180	168
15	187.2	180
16	144	192
17	153.6	204

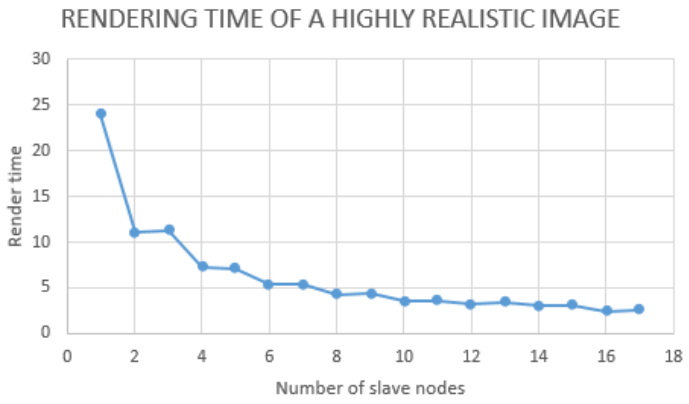


Fig. 8. Rendering Time of a Highly Realistic Image

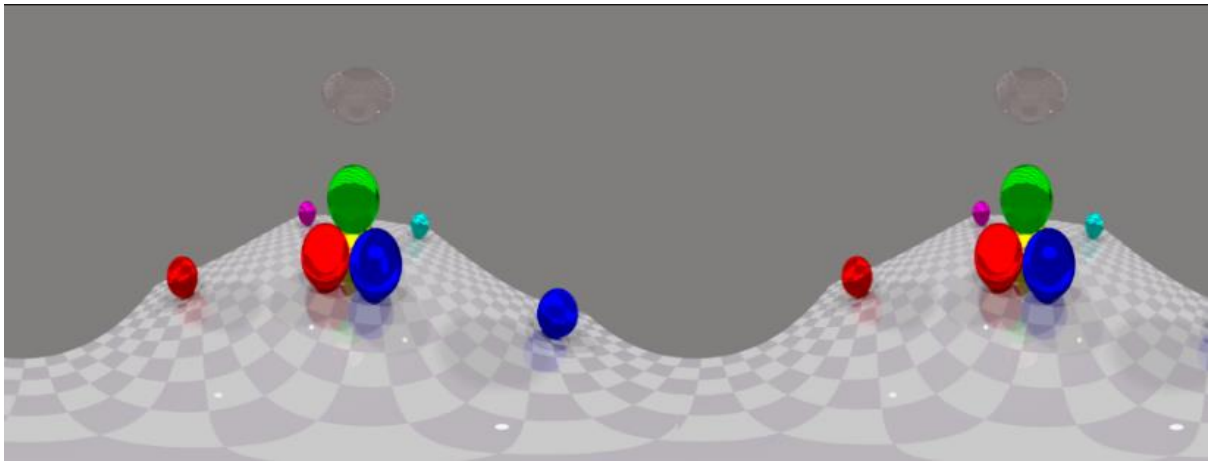


Fig. 9. Rendering Time of a Highly Realistic Image.

IV. Discussion and Conclusions

The farms of distributed rendering for images of high realism and complexity will be very useful for the research so the need to have more cores is beneficial to improve the rendering time so this project can be improved in increasing the number of cores without the need to use more computers, as it has the case of [12] that uses GPU-based engines which has a greater number of cores in order to reduce time. In the rendering of the prototype, it can have imbalances in a certain number of nodes in Table II, they are appreciated five moments of imbalance in the render these are due to several factors of which it has the issue of communication between nodes, more pixels in a strip of rendering and processes outside the rendering therefore it is desirable that the participating nodes are not doing other work that weaken to have an effective result. With regard to the future work of this research, high-realism animations will be carried out with respect to aerospace engineering prototypes, image processing, Big Data techniques and other investigations related to the existing lines in the direction of the investigation. These works will be carried out on the basis of distributed programming using the blender software as a rendering engine for the unification of rendered images and can be seen as a highly realistic animation in a shorter time. In this paper, it demonstrates that rendering farms under a Beowulf cluster reduces the time for obtaining large and complex images. Therefore, it is a benefit to use them in projects that require a realistic project model proposed for the research direction of the Universidad de Ciencias y Humanidades.

References

[1] M. Z. Patoli, M. Gkion, A. Al-Barakati, W. Zhang, P. Newbury, and M. White, “An open source grid based render farm for Blender 3D,” 2009 IEEE/PES Power Syst. Conf. Expo. PSCE 2009, no. April 2014, 2009.

[2] E. L. Huamaní, P. Condori, and A. Roman-Gonzalez, “Implementation of a Beowulf Cluster and Analysis of its Performance in Applications with Parallel Programming,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 522–527, 2019.

[3] N. Seticaya, “Implementasi Rendering Farm dengan Teknologi Cluster Computing Menggunakan Back Burner di Laboratorium Multimedia,” vol. 7, no. 2, pp. 41–56, 2018.

[4] G. V. Patil and S. L. Deshpande, “Distributed rendering system for 3D animations with Blender,” 2016 IEEE Int. Conf. Adv. Electron. Commun. Comput. Technol. ICAECCT 2016, no. March, pp. 91–98, 2017.

[5] A. Sheharyar and O. Bouhali, “A Framework for Creating a Distributed Rendering Environment on the Compute Clusters,” Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 6, pp. 117–123, 2013.

[6] K. Cho et al., “Render Verse: Hybrid Render Farm for Cluster and Cloud Environments,” Proc. - 7th Int. Conf. Control Autom. CA 2014, pp. 6–11, 2014.

[7] L. Chuquiguanca, E. Malla, F. Ajila, and R. Guamán-quinché, “Arquitectura clúster de alto rendimiento utilizando herramientas de software libre,” vol. 2, no. 1, pp. 1–8, 2015.

[8] J. P. Naiman, “AstroBlend: An astrophysical visualization package for Blender,” Astron. Comput., vol. 15, pp. 50–60, 2016.

[9] A. Asadulina, M. Conzelmann, E. A. Williams, A. Panzera, and G. Jékely, “Object-based representation and analysis of light and electron microscopic volume data using Blender,” BMC Bioinformatics, vol. 16, no. 1, pp. 1–9, 2015.

[10] J. Petke, M. Harman, W. B. Langdon, and W. Weimer, “Specialising Software for Different Downstream Applications Using Genetic Improvement and Code Transplantation,” IEEE Trans. Softw. Eng., vol. 44, no. 6, pp. 574–594, 2018.

[11] E. L. Huamaní, P. Condori, and A. Roman-gonzalez, “Virtualizing a Cluster to Optimize the Problems of High Scientific Complexity within an Organization,” vol. 10, no. 6, pp. 618–622, 2019.

[12] A. Martos and B. Ruiz, “Realistic virtual reproductions. Image-based modelling of geometry and appearance,” Proc. Digit. 2013 - Fed. 19th Int’l VSMM, 10th Eurographics GCH, 2nd UNESCO Mem. World Conf. Plus Spec. Sess. fromCAA, Arqueol. 2.0 al., vol. 1, pp. 127–134, 2013.

Performance Evaluation of IoT Messaging Protocol Implementation for E-Health Systems

M. Zorkany[1]
National Telecommunication Institute (NTI)
Cairo, Egypt

K.Fahmy[2]
Department of Electrical Engineering
Al-Azhar University, Nasr City
Cairo, Egypt

Ahmed Yahya[3]
Department of Electrical Engineering
Al-Azhar University
Cairo, Egypt

***Abstract*—Now-a-days, e-health and healthcare applications in the internet of things are growing rapidly. These applications are starting from remote monitoring of patient's parameters in home to monitoring patients during his life activities at work, transportation, etc. So we can monitor patients at any place outside of hospitals and clinical settings. By using this technology, we can save lives and reduce the number of emergency visits to hospitals. In contemporary time, there are great progress and opportunities for the internet of things (IoT) related E-health systems. Most IoT e-health platforms consist of three main parts; client nodes (patient or doctor), IoT server and IoT communication messaging protocol. One of E-health systems design over IoT challenge is choosing the most suitable IoT messaging protocol for E-health applications. In this paper, IoT remote patient and e-Health monitoring system was designed for monitoring physiological medical signals of patients based on most two famous IoT messaging protocols, MQTT and CoAP. These medical signals can be include parameters like heart rate signals, electro-cardio graph (ECG), patient temperature, blood pressure, etc. This practical comparison between CoAP and MQTT is to choose most suitable for e-health systems. The proposed approach was evaluated based on most significant protocol parameters like capability, efficiency, communication method and message delay. Practical and simulation results show the performance of the proposed E-health systems over IoT for different network infrastructure with different losses percentages.**

Keywords*—E-health; IoT; IoT protocol; CoAP; MQTT and remote patient monitoring

I. INTRODUCTION

IoT (Internet of Things) is an extension of the current internet by connecting millions of devices (things) together nevertheless the location of each thing. So IoT is a network between devices or things through any internet connection method (wire, wireless, WiFi, GPRS, 3G, etc). IoT applications grow in many applications like transportations, medical, shopping, smart home, smart cities etc. To build internet of thing platform for any application we need internet connections between clients and communication protocol to manage these connections through IoT server.

Nowadays the market of IoT in medical applications espial for e-health is growing rapidly ranging from medical smart sensors and remote patient monitoring to remote traceability and diagnosis of patients. The main objective of e-health over IoT is improving the health of patients through better disease management and self-help facilities.

IoT e-health systems can be classified as an Internet of Medical Things (IoMT). IoMT is a medical sensors and applications connected through IoT network. Also IoMT can be use Web Services, to captured patient data and analyzed it. The general architecture of e-health system [1] shown in Fig. 1 which consists of three main phases: Phase 1 monitoring the patient to acquire data from medical sensors, then this data is gathered and transferred through the internet to the application server or clients for data analysis, monitoring and investigation in third phase. Also some of data processing can be accomplished at patient side and send the results to doctor's side directly.

IoT platform has three main parts: clients, IoT server and IoT message protocols. As shown in Fig. 2 E-health monitoring system architecture [2] divided into three layers: sensing layer. It is a collection of medical sensors. server layers It links between the sensing layer "patient" and receiving data layer "hospital". Receiving data layer is the layer through which patient data are received. In this article, rely on the server layer to choose protocol most suitable for the electronic health system.

The messaging protocol is considered the main element in the Internet of things. These protocols play a big role to enable IoT all over life. One of E-health systems design over IoT is to choose the type of IoT messaging protocol. In the Internet of things the main factor is the incorporation of many technologies, due to the fast growth of the Internet of things information sharing has been sophisticated across systems where the internet is converted into a comprehensive future [3]. Internet of things has a great ability to grow and the main engine of this growth is the protocol [4]. So using dedicated protocols to the internet of things will increase the efficiency and ease of data transfer certified on the environment used. In recent years, health care applications have received a lot of attention, making researchers discuss the whole structure of the healthcare system through IoT and applicability [5].

The electronic health system has maintenance medical devices, way of managing hospitals and Patients were allowed arrival to medical care all hospitals in the world. Many hospitals have implemented an electronic health system to solve the problems of shortage of doctors, reduce costs for patients and increase efficiency [6]. The electronic health system aims to progress the health aspect where it monitors the organs of the human body and its vital functions, which helps to decrease the mortality rate. In the future, health care

applications across IoT will have a great effectiveness the state economy [7].

At this time, there is a lot of messaging protocols for IoT, some of which are distinguished according to a need for them. IoT protocol is divided into: TCP/IP protocol Such as: MQTT, SMQTT, AMQP and UDP protocol Such as: CoAP, SNMP, SSI. [8] Internet of things has major importance in electronic health so it was thought to this part of the application for the Internet of things to develop a Special Protocol for E-health to increase efficiency and care for patient cases in this paper, each protocol will be expounded separately to choose the most suitable protocol for health care system. There shall be a practical Comparison amongst the most renowned protocols it "MQTT, CoAP". This comparison will be based on the Special process components of each protocol.

In this research, e-health IoT platform will be implemented depending on most two famous IoT protocols MQTT (ex. TCP) and CoAP (ex. UDP). This practical comparison between CoAP and MQTT is to choose the protocol most suitable for e-health systems. The proposed approach was evaluated based on most important internet parameters like capability, efficiency and message delay. Practical results show a performance of E-health systems over it in different network infrastructure with different losses percentages. This comparison was presented to illustrate the integrity and health information towards the protocol most suitable for electronic health.

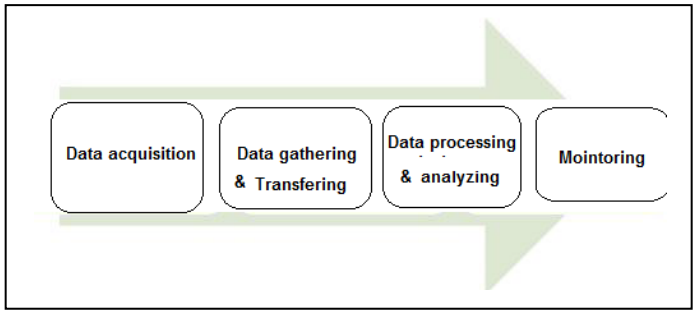


Fig. 1. General Architecture of e-Health System.

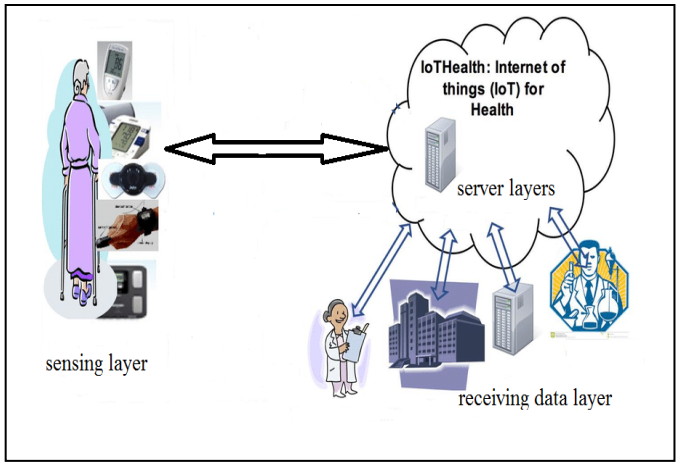


Fig. 2. e-Health Monitoring System Architecture.

This paper is regular as follows: Section 2 presents related of work IoT protocol. Section 3 discusses the famous IoT protocols. Then, Section 4 Proposed criteria for selecting the most protocol appropriate for the E-health application, in Section 5 set up scenario and practical results, the concludes of this research will be discussed in Section 6.

II. Related Work

Most e-health over IoT studies and researches in remote monitoring of patients although quite complete but are used for individual problems like as ECG or heart rate monitoring. Some studies developed for special features as develop prototypes for electrocardiography integrated with smart phone [9]. In recent years, many researches applied internet of things in many e-health applications. To develop these applications over IoT they used standard IoT protocols. For example, R.N. Kirtana, Y.V. Lokeswari in Paper [10] developed Heart Rate Variability (HRV) system to remote heart rate monitoring over IoT using MQTT protocol. But this system developed to monitoring one medical signal (heart rate) only. Antoine Jamin et al. in paper [11] presented an in-home aggregation IoT platform to monitoring more than one medical parameters (physiological and thermometer parameters). They used standard IoT protocol to implement their platform. So they send one message per parameter because MQTT is single topic protocol. In this case if patient has more than one emergency parameter in the same time, they must send one parameter in one MQTT message, and then transmit next parameter in other MQTT message. Which cause a delay in some message and occupy more internet bandwidth.

Kaleem Ullah et al. in paper [12] proposed framework for e-Health which used smart phone sensors with body sensors to monitor patient health. They used different protocol as HTTP and HTTPS. But they didn't use most famous IoT protocols like MQTT or CoAP. M. Hussein et al. developed IoT protocol based on proposed multi-topic IoT protocol which can send multi parameters in one message to overcome delay of multi messages and save the bandwidth [13]. Also, several researchers shared their efforts to develop the health care framework across IoT for example: P. Thota and Y. Kim compared the famous IoT Protocols MQTT and CoAP based on Raspberry-Pi and a temperature sensor [14]. The Pros and Cons of each of them in terms of the power consumption and data, flows from one node to multiple nodes, safely. CoAP more functional in terms of the energy, while the MQTT more suitable when data flows from a client to multiple nodes, but this consists of only two factors and neglect other factors.

K. Natarajan et al. used Raspberry Pi in E-health over IoT to collect patient data through the sensors, store it and send to the doctor [15]. Raspberry Pi is capable to collect diverse information from the patient through sensors. Mentioned how the data were collected but he did not talk about anything concerning the method of sending data over the protocol used in IoT. A. Al-Fuqaha et al. provide an overview about IoT, basic elements, tools and techniques needed from where communication technologies used in IoT like NFC, Z-WAVE, LTE-A, Bluetooth LE and RFIC [16]. In addition to the most renowned IoT messaging protocols as MQTT, CoAP XMPP,

AMQP. Description of different criteria for each protocol and explained the using of each of them in detail. Explained how to collect information by previous communication technologies and send off them through those protocols to achieve the goal of IoT. This paper didn't mention the challenges facing the protocols and the differences application between all these protocols. Also, they didn't run the protocols mentioned in real applications of the IoT, but mentioned in a theoretical way only. They didn't address the search of the best protocols from where energy and efficiency and the capacity to send data. It was necessary to mention the best protocol for each part of the different IoT.

T. Takpor and A. Atayero talked about the dangerous matter of integrating E-Health and educational institutions across IoT to effectively monitor students' health [17]. Because health issues affect students' academic performance, so exhibited a technique 'RFID' Radio Frequency Identification, and use it to know the medical data of students such as “blood”.

There a lot of work on different directions for the achievement of that project from sensors used and accuracy and how to compile data and keep it and how to transmit data Through IoT protocols. In this article Show the effect of IoT on improving electronic health, clarifying the communication models and protocols used in IoT and health care implementation. Today can be applied mobile devices in monitoring patients and the elderly and identifying their health status. The problem is how to join these devices to the internet world while preserving the safety of information. It was necessary to have a look at a collection of IoT messaging protocols used to solve the problems and to apply the application to transmit data [18].

From the above survey, we find that most researches didn't talk about the best protocols that ability to use in E-health applications. In our research, we will do a comparison between most IoT famous protocols to select the most appropriate protocol for the healthcare system. In the next section, clarification the most renowned IoT protocol will be discussed.

III. IoT Message Protocol

IoT message protocols can be divided into TCP/IP and UDP protocol. The most popular TCP protocol is "MQTT" and one of the most renowned UDP protocols is "CoAP".

A. MQTT "Message Queue Telemetry Transport"

It is a publish-subscribe messaging TCP/IP protocol. The messaging style "publish-subscribe" need to IoT server "broker". The publish send the message to the broker, the broker is accountable for classified messages to interested clients relying on the topic of a message as illustrated in the Fig. 3. It has many advantages as short massage, minimized data packets, faster response and throughput ‘Speed’, low power usage and lower bandwidth, Other advantages built into the MQTT protocol are retained messages and multiple subscriptions ‘multiplexed’ over one connection. many-to-many communication protocol for crossing messages between multiple clients through the broker.

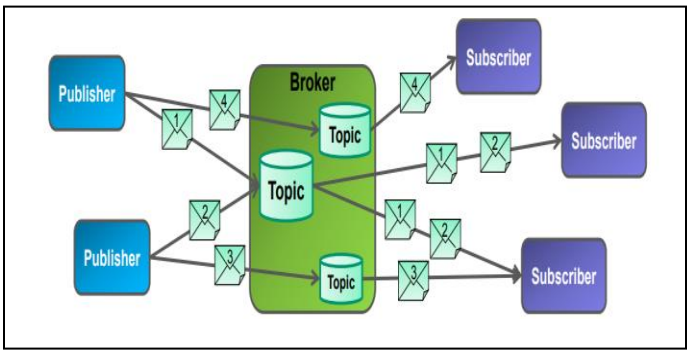


Fig. 3. General MQTT Architecture.

MQTT has featured it supports Quality of Service (QoS). The Quality of Service (QoS) is a method to secure the submission of the message between publisher and subscriber. There are three standards of Qualiy of Service (QoS), the standard of Quality of Service (QoS) is determined as in paper [19] within the scope of the research, it was found that there were some letters that showed official models of MQTT as well as how to pass messages and analyze them a constant analysis based on indications in paper [20]. MQTT protocol has the ability to save the order between the messages by adding the mark of the request and SEO which determines whether you want to keep the order between the messages or not.

B. CoAP "Constrained Application Protocol"

It is a "Request/ Response" protocol, it is used as an alternative to http in hardware-constrained devices to be simple to use with these devices so it is used in IoT. The CoAP provides four different types of messages:

- The message type defines four different method:

CON Message: refers to the “Confirmable” request. When a source node transmits a CON request, the recipient has to respond with ACK message.

NON Message: refers to “Non-Confirmable” request that when a source node transmits a non-request, the recipient is not required to respond back.

ACK Message: refer to “Acknowledgement” messages which are sent back as a response to a CON message. If processing successful, the recipient of the CON message should reply back with an ACK message. The ACK message can also contain the result of the processing time.

RST Message: refer to “Reset” message this type of messages is sent back when the future of a message encounters an error, does not understand the message or is no longer interested in the message sender.

- The message codes define one method: empty message.
- Request codes define four different methods:

(GET) This will retrieve the information from a specific resource specified by requested URI (POST) display information to be processed for a specific resource. The output result depends on the targeting resource, usually, results in the target resource that is created or updated. (PUT) requests that

resource particular by the URI be created with the carried information representation; (DELET) demand that the identified resource can be deleted.

- Response codes define two different methods :

Success message Contains four messages: CONTENT message response for GET message, CONTIUE message response for POST message, CREATED message response for PUT message, DELETED message response DELETED message.

Client error message Contains different message: such as bad request, not found, method not allow.

General CoAP architecture is exhibit in Fig. 4.

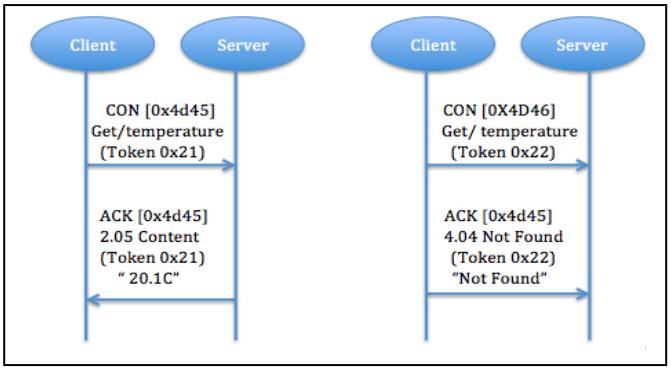


Fig. 4. General CoAP Architecture.

IV. Proposed Performance Evaluation Criteria for E_health over IoT

The proposed e-health system is An automatic wireless health monitoring system which can be used to measure patient's parameters (like body temperature and heartbeat, etc.) by using embedded system technology with internet of things. As shown in Fig. 5, The architecture of proposed system divides into five parts: Patient (sensor) node, IoT communication protocol, IoT server, application server and application node.

The patient node unit is medical sensors, internet access, controller and power and surrounded circuits. Medical sensors used to acquire medical parameters from patients. Selection of medical sensors depend on two issues decisions: which required sensing parameters need monitored, and which sampling frequency needed to send theses parameters to doctors/Hospital or sending these parameters only when required or at abnormal conditions. In order to the patient node achieve the objectives of the system; the scope of the node is summarized as follow:

Microcontroller system controls the operation of gathering the data from sensors and preparing it to be sent. It is integrated with some extra component to increase the functionality and execute at high frequency to improve the performance of the system so that the system is more reliable and efficient. Internet access like Wi-Fi or Bluetooth or GSM/GPRS/3G module used to transmit and receive data wirelessly in a long distance so the system is portable and easy to be operated.

A. IoT Protocol Architecture

In order that achieves better healthcare in the Internet of things, need to transfer data of patients accurately and effectively to remote servers. This paper proposed method to select the most convenient messaging protocol for the healthcare system. The architecture of the protocol system appropriate for electronic health depends on three layers, Client receiving data, Gateway / Translation Device "Protocols", Sensor data collection "sensing layer" as described in Fig. 6.

The sensing layer consists of many medical sensors such as pressure measurement, temperature, heartbeat, ECG, blood glucose level and breathing rate. Each sensor measures its own patient data and sends it to the doctor through the server layer, in this class, the data enable be collected and rerouted.

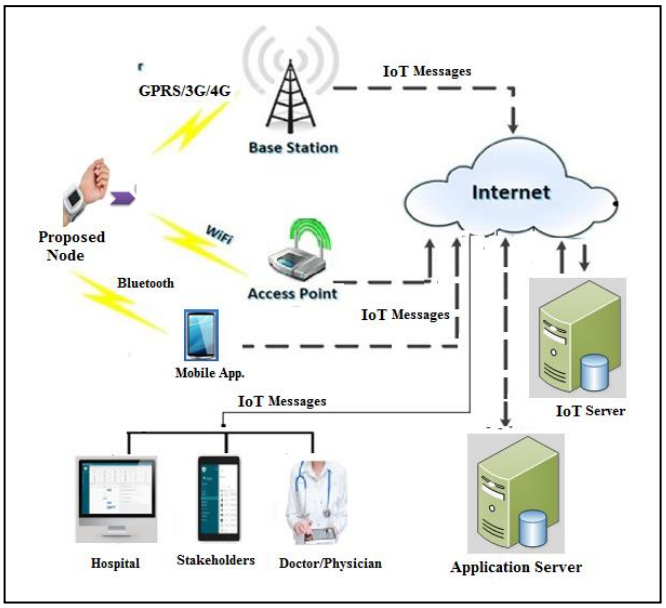


Fig. 5. Proposed IoT Patient Monitoring System.

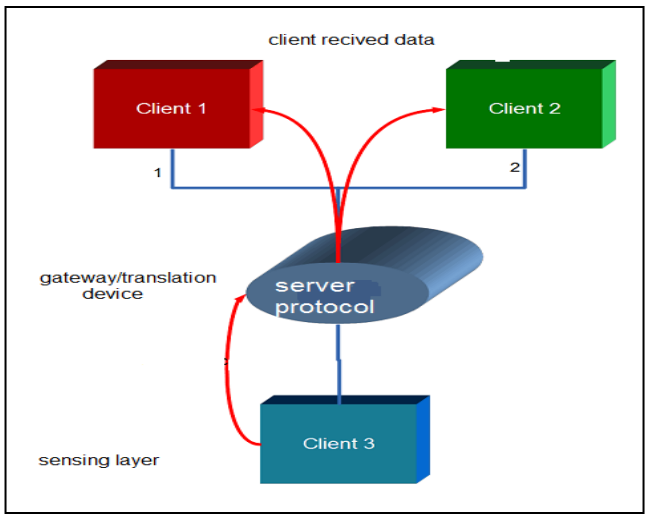


Fig. 6. The Proposed Methodology.

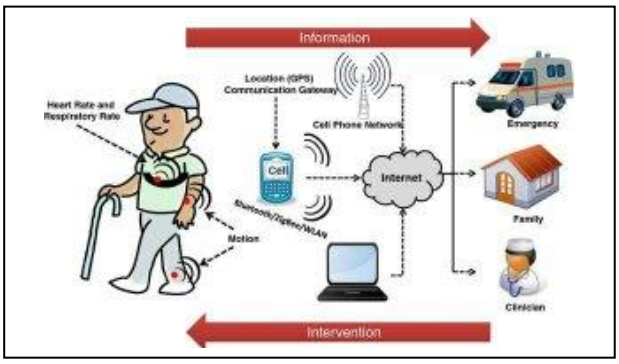


Fig. 7. Remote Health Monitoring.

In the server layer the focus is on communication protocols that authorize the reciprocation of information between devices, where the Internet of things is entered in many fields, whether industrial or health, where information exchange occurs between devices via cloud systems, the link in that is many communication protocols Which allow devices converse to each other and with the connection of millions of devices on the Internet, need for IoT protocols was very important such as MQTT, CoAP looking for whichever is more compatible for electronic health.

In the receiving data layer the information is received by the doctor that was sent in the sensing layer where the information is processed and stored continuously, occur a quick reaction to the emergency and other treatment by the doctor. Fig. 7 shows these three stages.

The first phase "The sensing layer" shows the patient's specific sensors and how to install it and measure the data. the second phase "server layer " describes how to transfer that patient-specific data that has been measured through the medical sensors via the IoT protocols. The third phase "receiving data layer" is the special phase in the hospital where the doctor receives the data of the patient was sent in the previous phase through special applications. Application "client" for each protocol to be the window to transmit data to patients Like MQTT FX data is then transferred to the public network to select most protocol convenient for transfer data. The researchers have made protocols IoT 'internet of things' to suit data transfer, enhance efficiency, safety, and energy consumption, as the devices used in electronic health often operate on batteries, the protocol should be suitable for energy saving and easy to use for the customer.

B. Metrics of the Proposed Method

The proposed comparison in this article based on three aspects metrics as follows:

1) IoT Success Message "ISM": With this metric, the speed of sending messages is evaluated, this is because the messages that arrive directly are quicker than the messages that need to retransmitting due to loss in the network. This measure is the rate of the total number of messages sent from client to server without the occurrence of retransmitting it, and the gross number of messages sent whether it has to resend or not. pronounced in the equation. all message sent M(s) consists of:

The successful message M(succ) from the first time and other message happen to be retransmitted M(re).

$$M(s) = M(succ) + M(re) \quad (1)$$

$$ISM = \frac{\text{No.of successful messages}}{\text{No.of retransmit and successful message}} \quad (2)$$

2) IoT Average Byte "IAB": With this metric, an medium number of bytes used in the message evaluated, because all protocol has its own format regardless of the tenor of the messages. This measurement is the rate of an average number of bytes used in messages sent from the client to the server without retransmit, and the medium number of bytes used in sent messages, whether resend it or not.

$$IAB = \frac{\text{average byte of successful messages}}{\text{average byte of retransmit and successful message}} \quad (3)$$

Pronounced in the equation. Packet contains a collection of byte it is a byte of payload B(p) and byte of 'fixed, variable' header B(h) specific protocol.

$$B = B(p) + B(h) \quad (4)$$

3) IoT Delay message "IDM": Message delay is evaluated in this metric, this is ratio of delay of messages sent from the client to the server without retransmission, and delay messages sent, whether resend it or not. This is because most important in E-health is the speed and size of the sending the message and the energy consumption add to the messages lost. Pronounced in the equation.

Delay consists of transmission delay D(t), signal processing delay D(s) and queueing delay D(q).

$$D = D(t) + D(s) + D(q) \quad (5)$$

$$IBM = \frac{\text{delay of successful messages}}{\text{delay of retransmit and successful message}} \quad (6)$$

As the speed of access to patient information to the doctor to follow the importance of medical, as well as the number of lost messages are a major risk in the diagnosis, the smaller the volume of messages sent the faster the access. Fig. 8 shows a comparison between most prominent IoT protocols MQTT, CoAP. The Patient-specific data "client1" will send to the specialist doctor "client2" through the server of the protocol used.

It is distinguished by the MQTT it supports (QoS). The Quality of Service (QoS) is to secure the transmission of the message between publisher and the future exist three-phases of Quality of Service (QoS), the publisher determine the level of the service.

QoS level 0: message sent from publisher to intermediary and no acknowledgment is sent to a publisher. This part has not been reviewed because it was working on a local network.

QoS level 1: a message sent by the publisher to the intermediary and then the mediator sends an acknowledgment to the publisher again until the publisher ensures that do not miss the message, this means that transmission is guaranteed.

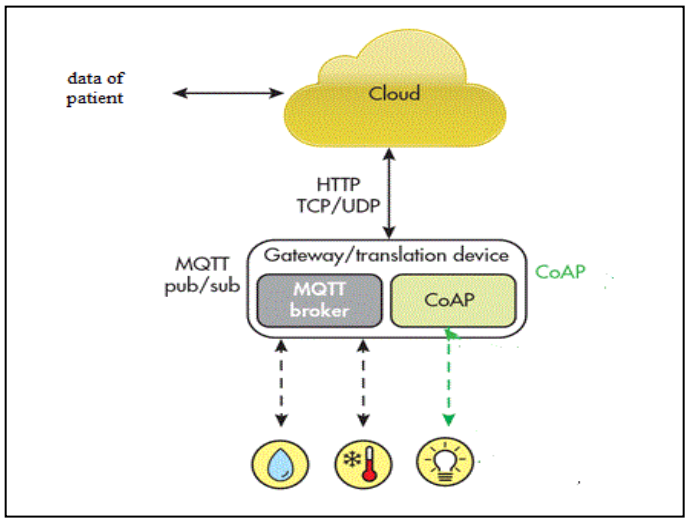


Fig. 8. The Actual Implementation.

QoS level 2: The highest level of service, where there is a chain of four messages between publisher and mediator to confirm sending and confirm receipt.

V. Implementation of IoT Protocol and Set up Scenario

A typical comparison between CoAP and MQTT and evaluated for appropriate selection, each client has a protocol and a private server.

MQTT implementation: have "MQTT FX" as a client and "MOSQUITO" server. MQTT FX" CLIENT through which the message sent from publisher "patient" and received it by subscriber "Doctor"; subscriber and publisher related to the server with special IP. MOSQUITO" SERVER mosquito is an MQTT broker used for handling data between publisher and subscriber. Receives data from publisher and sends it to the subscriber certified the subject of the message.

CoAP implementation: have "COPPER" client and "LIBCoAP "server. "COPPER" CLIENT through which the message sent from the patient to his doctor connects to the server using the IP of the server. CoAP client use 4 orders (GET) that retrieves information from a specific resource that is particular by the desired URI. (POST) provides the information to be processed to a private resource. The output result depends on target resource, usually, results in the base resource being created or updated. (PUT) requests to create the resource that is particular by the URI or updated with the carried information representation. (DELETE) requests that identified resource be deleted. "LIBCoAP" SERVER it is a special server for CoAP protocol, where data is sent from the patient to server using a command "put" and the doctor takes the patient's data from a server using a command "get".

To simulate the loss of network to show the performance of the proposed e-health systems over IoT at different network infrastructure with different losses percentages, some assisting programs were used as WANEM and WIRE SHARK.

WANEM is an extensive network emulator, designed to provide a real network experience. Which can be used to simulate WAN properties such as network delay, packet damage, packet loss, disconnecting, jitter, reorder of the package, etc. WANEM through which we can change the loss so as not to result in ideal case the WANEM must mediate Client and broker by routing table.

WIRE SHARK supports large number of the protocol. It analyzes the data sent and received through the protocol. Through it know: Public packets and subscribe, Timespan, Average packet per second, Average packet per size, average byte per second. From here we can calculate: IoT Success Message "ISM", IoT Average Byte "IAB" and IoT Delay message "IDM".

The scenario through which we have chosen the appropriate protocol for electronic health. It appears in Fig. 9.

The first scenario was executed using an MQTT protocol, where we have sent the patient's data from the client "MQTT FX" to the attending physician "doctor" through MOSQUITO server. Using WIRESHARK We can tracing the MQTT messages to know which of these messages received from the first time and which of them need re-sent. Through Wireshark, we were Managed these MQTT message to calculate the average byte used in sent messages and the time need to transmitting. All previous measurements in the ideal case, We have varied the loss ratio by WANEM server and re-do the practical simulations and re-measure an calculate the average byte and time to transmitting messages.

The second scenario was implemented using the CoAP protocol. We sent the patient's data from the CoAP client "CoPPER" to the doctor by the LIBCoAP server passing with WIRESHARK. By tracing the sent data to show which messages received at destenation "doctor". Also via WIRESHARK analysis we can calculate the average byte used to send messages and the user time to transmitte. all the previous measurements in the ideal case, then we have change the loss ratio using WANEM simulator and re-measure all the previous parameters to study the performance of the protocol in different network conditions.

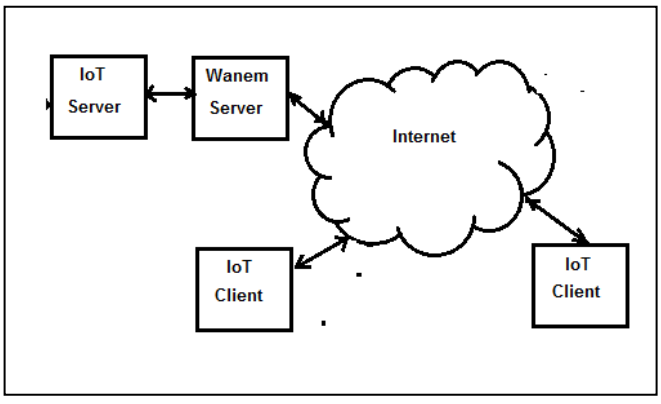


Fig. 9. Set up Scenario.

VI. Practical and Simulation Results

The main objective of the practical results and simulation is to clarify which protocols "MQTT, CoAP" are more suitable for electronic health where the results were evaluated by three factors. "ISM, IAB, IBM". From the practical and simulation results:

1) As shown in Fig. 10, the percentage of successful messages in case of the MQTT protocol greater than CoAP protocol, MQTT has a quality of service so the result is better in case of increasing the loss.

The CoAP protocol is special case of UDP protocol. And cannot return lost messages and does not contain QoS, if there is a loss of patient data, the data cannot be retrieved and therefore cannot be used CoAP protocol in medical cases because the doctor need all the patients data without missing anything so that he can Correct diagnosis and ensure improved health care.

2) As shown in Fig. 11, the rate of average byte in case CoAP greater than MQTT in the case of increasing the loss precentage. Despite what appeared in the first curve of the defect of the Coap protocol but it appears the second drawback is the large increase in the number of bytes used to send the message. Increasing the size leads to a large load on the network causing some messages to be lost. In spite of QoS 2 in case MQTT needs to send a series of messages to confirm receipt of message it needs a number of bytes less than the CoAP, does not mean increasing the size of messages in QoS 2 for their size in QoS 1 use QoS 1, However, we need to use QoS 2 in medical cases to ensure patient data is correct. There are also features in MQTT protocol it moves between Quality of Services automatically. You choose QoS 2 and the network does not allow to use because of a problem in it Selects QoS 1 so the network service is improved and then the choice is back QoS 2 again.

3) As shown in Fig. 12, The ratio delay of successful message in case CoAP greater than MQTT in the case of increasing the loss percentage.

Here is talk about the delay in time, it turns out that the delay in QoS2 is greater than QoS1 and CoAP because it sends a series of messages to ensure the arrival of the message carrying the patient's data. From here we can say that the applications that need to be accuracy in sending the message do not consider the time delay being used QoS 2, but the Applications that care more about time are used QoS 1. In both cases, MQTT is more appropriate than CoAP.

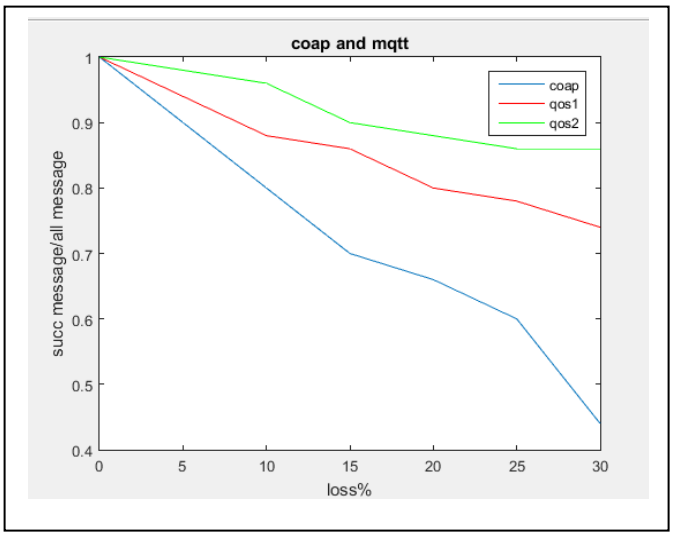


Fig. 10. IoT Success Message "ISM".

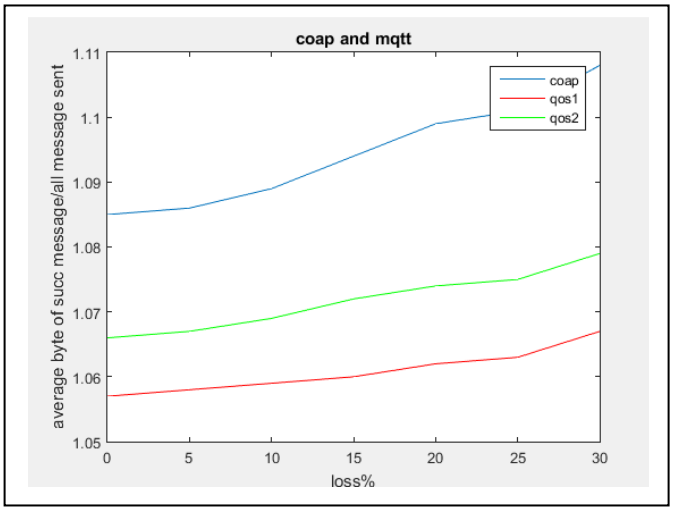


Fig. 11. IoT Average Byte "IAB".

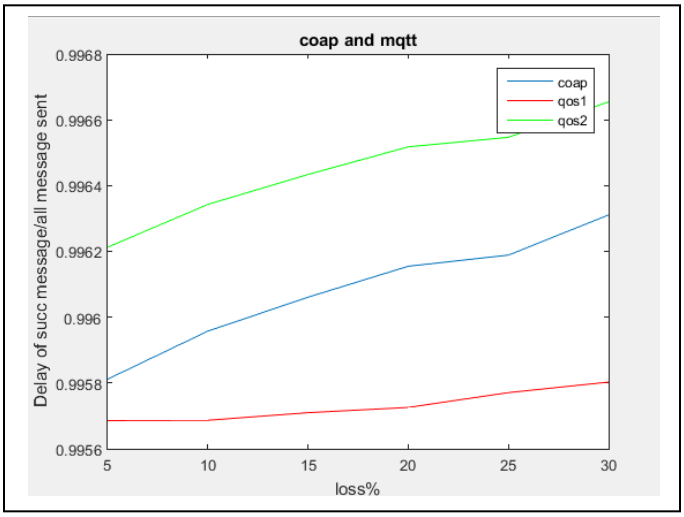


Fig. 12. IoT Delay Message "IDM".

VII. CONCLUSION

This paper presents design and implementation of e-health platform over IoT. A reliable e-health monitoring system over IoT has been designed and successfully implemented in this work based on MQTT and CoAP protocols. Whereas the proposed system is based on any type of internet access like WIFI, GPRS, 3G, so it is easily reconfigurable and can be extended to include more medical sensors and more parameters. The proposed e-health system is flexible enough to include such kind of modifications. Also, Comparing "MQTT & CoAP" protocol to select the best IoT protocol for E-health applications is investigated in this paper. In particular, it defines real results depending on practical simulations to select the best IoT protocol for e-health applications. From the practical implementation of these protocols, MQTT protocol provided good results than CoAP protocol in the delay, the number of messages lost and the number of bytes used in messages. Thus decrease the energy used, and reduce the needed bandwidth. So from these practical and simulation results, MQTT message protocol more suitable to design and implementation of e-health platform over IoT.

REFERENCES

[1] B. Lee, "Healthcare Framework on the IoT open Platform," Service Model, Architecture, International Journal of Applied Engineering Research, vol. 9, pp. 29783-29792, 2014.

[2] M.Bhay, M. Pat el and Chin tan Bhatt," Internet of Things (IoT): In a Way of Smart World" © Springer Science, Singapore 2016.

[3] D. Uckelmann, M. Harrison, F. Michahelles "An Architectural Approach Towards the Future Internet of Things", Architecting the Internet of Things, Springer-Verlag Berlin Heidelberg 2011.

[4] U. Tandale, Dr. B. Momin and D. P. Seetharam " An Empirical Study of Application Layer Protocols for IoT" International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017) , IEEE, 2017.

[5] B. Farahani , F. Firouzi , V. Chang , M. Badaroglu ,N. Constant , K. Mankodiya " Towards fog-driven IoT E-Health: Promises and challenges of IoT in medicine and healthcare", Future Generation Computer Systems , 2018.

[6] J. Shah, S. Soni, F. Darji, S. Chandak, A. Shetty "Smart Hospital Using Iot "International Conference on Innovative and Advanced Technologies in Engineering , IOSR Journal of Engineering (IOSRJEN), Volume 9, PP 26-32. 2018.

[7] C. I. Saidu, A. S. Usman and P. Ogedebe, "Internet of Things: Impact on Economy", British Journal of Mathematics & Computer Science, 2015.

[8] D. B. Ansari, A. Rehman, R. Mughal" Internet of Things (IoT) Protocols: A Brief Exploration of MQTT and CoAP", International Journal of Computer Applications (0975 – 8887) Volume 179 – No.27, 2018.

[9] Jorge Gómez, Byron Oviedo , Emilio Zhum, " Patient Monitoring System Based on Internet of Things", 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016), Published by Elsevier, 2016.

[10] R.N. Kirtana, Y.V. Lokeswari, " An IoT Based Remote HRV Monitoring System for Hypertensive Patients", IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP), IEEE, 2017.

[11] Antoine Jamin et al."An aggregation plateform for IoT-based healthcare: illustration for bioimpedancemetry, temperature and fatigue level monitoring", 2017.

[12] Kaleem Ullah et al.," Effective Ways to Use Internet of Things in the Field of Medical and Smart Health Care", International Conference on Intillegent System Engineering, IEEE, 2016.

[13] M. Hussein, M. Zorkany, N. Abdel Kader, "Design and Implementation of IoT Platform for Real Time Systems", International Conference on Advanced Machine Learning Technologies and Applications (Springer), Egypt, 2018.

[14] P. Thota, Y. Kim "Implementation and Comparison of M2M Protocols for Internet of Things" Intl Conf on Applied Computing and Information Technology,© IEEE, 2016.

[15] K. Natarajan, B. Prasath, P. Kokila "Smart Health Care System Using Internet of Things" Journal of Network Communications and Emerging Technologies (JNCET) Volume 6, Issue 3, ISSN: 2395-5317 , 2016.

[16] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, Moussa Ayyash " Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications", IEEE COMMUNICATION SURVEYS & TUTORIALS, VOL. 17, NO. 4, 2015.

[17] T. Takpor and A. Atayero "Integrating Internet of Things and E-Health Solutions Students' Healthcare", Proceedings of the World Congress on Engineering 2015.

[18] S. Lee, H. Kim, D. Hong, Hongtaek Ju " Correlation Analysis of MQTT Loss and Delay According to QoS Level", IEEE ICOIN, 2013.

[19] B. Aziz, " A formal model and analysis of an IoT protocol" Ad Hoc Networks, Elsevier B.V. ,2016.

[20] H. C. Hwang ,J. Park, J. G. Shon," Design and Implementation of a Reliable Message Transmission System Based on MQTT Protocol in IoT" Wireless Pers Commun, Springer Science, New York ,2016.

Analysis of Password and Salt Combination Scheme To Improve Hash Algorithm Security

Sutriman[1], Bambang Sugiantoro[2]
Master of Informatics Department
Sunan Kalijaga Islamic State University
Yogyakarta, Indonesia

***Abstract*—In system security, hashes play important role in ensuring data. It remains the secure and the management of access rights by those entitled to. The increasing power of hash algorithms, various methods, are carried out one of them using salting techniques. Salt is usually attached as a prefix or postfix to the plaintext before hashing. But applying salt as a prefix or postfix is not enough. There are so many ways to find the plaintext from the resulting cipher text. This research discusses the combination scheme other than the prefix and postfix between password and salt increasing the security of hash algorithms. There is no truly secure system and no algorithm that has no loopholes. But this technique is to strengthen the security of the algorithm. So that, it gives more time if an attacker wants to break into the system. To measure the strength generated from each combination scheme, a tool called Hashcat is used. That is the way known as the best composition in applying salt to passwords.**

Keywords—Security; hash; hashing scheme; salting; password

I. INTRODUCTION

Hash is an algorithm that changes the string becomes a series of random characters. It is also called a one-way function, or one-way encryption because it is only able to do encryption and does not have a key to decrypt. It works by accepting input strings that are arbitrary in length then transform it in a string of fixed length which is called hash value [1][2][3][4].

Hash is often used to provide security to the authentication process. An authentication is a process of ensuring a property is genuine, verifiable and trustworthy; deep conviction the validity of the transmission, message, or sender of the message. It verifies that the user should input entered from the system coming from a trusted source [1].

Authentication is one of several concepts needed to ensure the security of a system. Authentication along the accountability is the additional concept needed to support the CIA Triad. CIA Triad is a concept very well-known as the security, named the Confidentiality, Integrity, and Availability [1][5]. CIA triad is the basic model of Information Security and there exist other models that have the attributes of the CIA triad in common [6]. Despite the use of the CIA to determine goals security is well established.A few in the security sector feels that the additional concept is needed to present the picture completely [1].

Authentication is a very important process because besides maintaining information from unauthorized users. It also maintains the integrity data [7][8]. The use of algorithms and hashing techniques is needed to help the authentication process so that they can minimize the occurrence of broken data by the attacker. The authentication process utilizes the use of algorithms hash including the authentication of login (password), authentication file authenticity, password storage, key generation, pseudorandom number generation, authentication of tokens on services in a distributed system, digital signature, etc. [9].

In information systems, a hash is used for the authentication login process. Passwords are changed using certain hashing methods thus producing unique characters later stored in the database. Some common hash functions used include MD5 and SHA1 [3]. A message digest (MD) is the code which is created algorithmically from the file and represents that file uniquely. If the file changed, the message digest will change [10]. Message Digest describes the mathematical function that can take place on a variable-length string. The number five (5) simply depicts that MD5 was the successor of MD4. MD5 is essentially a checksum that is used to validate the authenticity of a file or a string. It is one of the most common uses [11]. The MD5 algorithm exhibits a lot of weaknesses such as its vulnerabilities to different attacks such as rainbow table, dictionary, birthday, etc. [12]. SHA is a series of cryptographic hash functions designed by the National Security Agency (NSA). The weakness in SHA family originated from this fact that possibility of two different input value will produce the same output value in the middle of algorithm and it is important to have a good diffusion. So, the output in each round will be spreaded out and not to be equal with the same output in the next coming stages [13]. MD5 and SHA1 are hash algorithm that do not recommend. MD5 and SHA1 have many vulnerabilities which allow attackers to easily get the system user password by knowing the hash value.

A new hash algorithm appears as time progresses with better security than the previous algorithm, among them are SHA2, SHA3, BCrypt, and others. Along with the development of the era, no doubt the new algorithms even the vulnerability of attackers will be found.

The use of hashes in the authentication process actually can reinforce by adding salt to the plaintext password before the hashing process is carried out. Salt in the cryptography is a

random bit that is used as a joint input password before the hashing process is done. Salt can be added to the hash to prevent a collision by uniquely identifying a user's password, even if another user in the system has selected the same password [14]. Salt is used as prefixed or post fixed against passwords before entering the process hashing. Each password has different salt.

The use of salt for passwords can increase password security in an application, but does not close possibilities that the attacker can crack against the generated values. Various kinds of password cracking tools who are currently circulating in cyberspace started by using password and salt combination for plaintext password guessing. Giving salt a prefix or postfix still has a vulnerability.

In [15] is previous research that has been conducted that discusses the method of exchanging passwords and salt. In that study, each index in the password and salt was exchanged and then stored in an array. Unfortunately, the research does not show the results of strength measurements that could assess the results of the study clearly. So, raising the question is a really measured thing. Then it can be added from the other parameters as a comparison.

II. Related Work

Password is one of the most important components of any classical security scheme designed to protect sensitive and confidential information from falling into the wrong person. Creating a strong password is the first step towards ensuring the protection of confidential user information [16]. The purpose is to improve password security and to contribute the research on password security, including salting techniques. Salting technique is a hedge against pre-computed dictionary attacks, the bedrock of which involves concatenating a random string of letters and numbers, a salt, to the beginning or end of a password before hashing it [17].

Abdelrahman Karrar *et al.* have published research on the security of hash algorithms by swapping every index for passwords and salts. Swapping Elements in an array algorithm consists of two main modules, the Hash input structuring module and the Salt rearrangement module[18].

The previous research on the use of salt strengthen hashes on passwords has also been carried out by combining with differential masking techniques [15]. The differential masking is basically the insertion of fake passwords associated with each user's account. When an attacker gets the password list, he retrieves many password candidates for each account but cannot be sure about which password is real[19]. The use of various methods and combinations is needed to increase the power of hash algorithms. However, only using salt is not enough because it can still be attacked using the Bruteforce attack technique for only a little longer. Using two salts, one public and one private can also protect the password against offline attacks [20].

III. Research Method

Research methods are a way to collect various data processed into information. The information is used as materials to solve the problems studied. The steps taken in this study include data collection, data grouping, analysis process and ending with conclusions obtained. The research method used in this study is shown in Fig. 1.

A. Research Tools

Research tools are used to carry out the research process. In the testing process, the specifications of the research tool affect the test results. The research tool used in this study is a computer laptop with specifications as following:

- Brand/Type : HP / G 240 G7
- Processor : Intel Core i7-8565U
- RAM : 8 GB DDR 4
- Storage : 256 GB SSD
- Operating System : Linux, Xubuntu 18.04

B. Generating Salt Process

In this process, the salt which will be combined with the password is made. The salt is made using a random function that is owned by the PHP programming language. It plays an important role in making hashed passwords stronger. By using the salt the more unique and the longer, a simple password will become stronger. In this study, the salt generated is limited to only 3 character numbers to simplify the testing process.

C. Rearrangement Process

This process is a process that will be carried out to strong the hash algorithm. The use of salt is generally only combined as a prefix or postfix for passwords. In other cases, the use of salt composition is very dependent on the programmer who built a system. By doing randomization between passwords and salt, it is expected that a plaintext will be carried out by the hashing process. It will be more difficult to describe by the attacker, but without adding too many characters, it does not increase the hashing time to be longer.

D. Testing

The tests that will be carried out in this study use a process that can find the plaintext of the password. The treatment for each scheme will be different based on each combination. Each combination will be described, and the time needed to be able to decipher the ciphertext of the saved password recorded.

The decomposition process, at this stage of testing, will use a simple algorithm made in Python language while for the hash parsing process a password recovery tool will be used, named OclHashcat. It is the fastest password recovery program based on GPU, built by atom and can run on GNU or Linux and MS Windows, 32 and 64 bit [21].

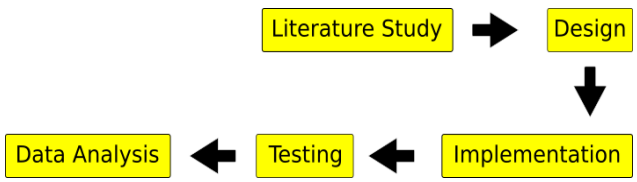


Fig. 1. Research Methods.

IV. Results and Discussions

A. Designing a Password and Salt Combination Scheme

Some draft password and salt combination schemes are needed to be able to compare the strengths of each design. The scheme created is a combination of several in the hash process between passwords and salt. The scheme used in this study can be seen in Table I:

TABLE. I. Password and Salt Combination Scheme

No	Scheme	Combination
1	Scheme 1	hash(rearrangement(password, salt))
2	Scheme 2	hash(rearrangement(password, salt)+salt)
3	Scheme 3	hash(hash(rearrangement(password, salt)))

Scheme 1 is the same scheme used by previous researchers. Schemes 2 and 3 are schemes created as a comparison by providing additional parameters. In scheme 2, the salt parameter is added to the exchange result before the hash process is performed. In scheme 3, a double hash is performed on the result of changing the password and salt position.

B. Implementation

Implementation is the process of implementing a scheme that has been designed into an application. The application made in this section is a simple application that can display the cipher text value of each designed combination scheme.

The implementation is done using the PHP and MySQL database. The use of databases in system implementation aims to get records of each result that is generated during the process of collecting data with the application that has been made.

1) Generating Salt: In this process, the salt which will be combined with the password is made. Salt is made using a random function that is owned by the PHP programming language. Salt plays an important role in making hashed passwords stronger. By using salt the more unique and the longer, a simple password will become stronger.

In this study, the salt was generated using the mt_rand() function that is already available in the PHP programming language and limited to only 3 character numbers to simplify the testing process. The mr_rand() function is used to generate a random number value between the given range. The function used to generate salt in its entirety can be seen in the program code below:

```
function randomSalt($length) {
        $result = '';
        for($i = 0; $i < $length; $i++) {
                $result .= mt_rand(0, 9);
        }
        return $result;
}
```

Script. 1. Generating Salt Function

2) Rearrangement Process: This process will be carried out to strengthen the hash algorithm. The use of salt is generally combined as a prefix or postfix for passwords. In other case, the use of salt composition is very dependent on the programmer built a system. By doing randomization between passwords and salt, it will be carried out by the hashing process will be more difficult to describe by the attacker without adding too many characters so that it does not increase the hashing time to be longer.

```
function rearrange ($arr1, $arr2, $n1, $n2){
        $i = 0;
        $j = 0;
        $k = 0;
        $arr3 = array();
        while ($i < $n1 && $j < $n2){
                $arr3[$k++] = $arr1[$i++];
                $arr3[$k++] = $arr2[$j++];
        }
        while ($i < $n1)
        $arr3[$k++] = $arr1[$i++];
        while($j < $n2)
        $arr3[$k++] = $arr2[$j++];
        $result = array();
        for ($i = 0; $i < ($n1 + $n2); $i++)
        array_push($result, $arr3[$i]);
        return implode("", $result);}
```

Script. 2. Rearrangement Function

The rearrangement() function above is called by including 4 parameters. The first and second parameters, namely $arr1 and $arr2 are passwords and salts. The third and fourth parameters namely $n1 and $n2 are the length of the password and salt. The above function works by rearranging the password and salt. For example, the password that is owned is abcdef and salt 123. It will produce an a1b2c3def string.

3) Data Collection: Data retrieval is the process of recording each cipher text value of each combination scheme for each predetermined password. The password used is a weak password taken from Splash Data's Top 100 Worst Passwords. From the collection of passwords found. The ones used in this study are those that match the criteria which are 6 characters long. The passwords used can be seen in Table II.

TABLE. II. Sample Password

No	Password
1	qwerty
2	monkey
3	abc123
4	123123
5	dragon
6	qazwsx
7	654321
8	harley

C. Testing

In this test, it is assumed that the examiner who acts as an attacker knows the combination scheme model and length and the type of data used for passwords and salts used as tested material. This aims are to limit research because the attacking process is very dependent on information and analysis obtained by the attacker. Thus, what was tested in this study was pure to measure the effect of the combination scheme on the strength of the hash algorithm.

1) Testing using Hashcat: Hash cat is a penetration testing tool that can be used to decrypt cipher text results from the hashing process. Hash cat has variety of attack methods with various hash decrypted models. In the testing phase of this study, what will be used is the Mask Attack model.

Mask Attack is an attack mode that uses a combination of characters to guess the plaintext sought. The Mask Attack on the Hashcat is similar to Brute force Attack. In traditional brute force mode, a character set is needed that contains all uppercase letters, all lowercase letters and all digits (mix alpha-numeric). So, it takes a long time.

In a Mask Attack mode, character sets can be arranged based on information about the target that has been obtained. In this test, not all mix alpha-numeric characters are used to carry out attacks but can be adjusted to the desired attack pattern based on the information that has been obtained. One example of the command used on Attack Mask Attack using Hashcat is as follows:

hashcat -a 3 -m 0 scheme3_monkey.txt ?l?d?l?d?l?d?l?l?l --force

For,

hashcat : Unique key used to call applications

-a : Attack Mode (3 to define the Mask Attack mode)

-m : Hash Type (0 to define that the target hash is MD5)

scheme3_monkey.txt : Chipertext as the target

?l?d?l?d?l?d?l?l?l : Characters used to determine the attack pattern,? l for letters a-z,? d for numbers 0-9.

2) Reverse Rearrangement Process: This stage is used because schemes that uses the rearrangement function before the plaintext is converted to the Ciphertext in the hashing process need to be rearrange. At this stage, the special scripts are created with python language which can be seen in the following script:

```
def main():
  input_string = raw_input("Input your string: ")

  start_time = time.time()
  start_time_print = time.strftime('%d/%m/%Y %H:%M:%S')
  plaintext_length = 6

  salt_length = 3
  string_to_list = list(input_string)

  list_of_result = []
  for x in xrange(0,len(input_string)):
    new_index = x
    if x != 1 and x != 3 and x != 5:
      list_of_result.insert(new_index,string_to_list[x])
  result = ''.join(list_of_result)
  print result
  print("\n\n\n--- start time: %s" % start_time_print)
  print("--- stop time: %s" % time.strftime('%d/%m/%Y %H:%M:%S'))
  print("--- time estimated %s seconds" % (time.time() - start_time))
```

Script. 3. Script for Reverse Rearrangement.

In the above function, the input string is a password and salt combination string. This function works by looping with the password length and salt parameters that were previously known. In the looping process, characters identified as part of the password stored in the new list and displayed as a result.

3) Test Results: The main purpose of the testing process is the time needed to get the plaintext for each scheme. The results of this test will then be analyzed in the next step. The test results are shown in the following Tables III, IV and V.

TABLE. III. TEST RESULT FOR SCHEME 1

No	Password	Salt	Time		
			Cracking Hash	***Reverse Rearrangement***	***Total***
1	qwerty	857	1h7m57s	0.000262975692749s	1h7m57.0003s
2	monkey	047	1h4m57s	0.00019907951355s	1h4m57.0003s
3	abc123	565	7s	0.000290155410767s	7.0003s
4	123123	345	5s	0.000843048095703s	5.0008s
5	dragon	372	1m16s	0.000288963317871s	1m16.0003s
6	qazwsx	512	3h35m17s	0.000334024429321s	3h35m17.0003s
7	654321	782	8s	0.000181913375854s	8.0002s
8	harley	456	1h6m14s	0.000261068344116s	1h6m14.0003s

TABLE. IV. TEST RESULT FOR SCHEME 2

No	Password	Salt	Time		
			Cracking Hash	*Reverse Rearrangement*	*Total*
1	qwerty	857	2h12m36s	0.000263929367065s	2h12m36.0003s
2	monkey	047	2h9m1s	0.000308990478516s	2h9m1.0003s
3	abc123	565	17s	0.000260829925537s	17.0003S
4	123123	345	16s	0.000319004058838s	16.0003s
5	dragon	372	12m33s	0.000279903411865s	12m33.0003s
6	qazwsx	512	4h18m40s	0.000226020812988s	4h18m40.0002s
7	654321	782	35s	0.000211000442505s	35.0002s
8	harley	456	1h50m22s	0.0001380443573s	1h50m22.0001s

TABLE. V. TEST RESULT FOR SCHEME 3

No	Password	Salt	Time		
			Cracking Hash	*Reverse Rearrangement*	*Total*
1	qwerty	857	1h48m39s	0.000224113464355s	1h48m39.0002s
2	monkey	047	1h57m11s	0.000231027603149s	1h57m11.0002s
3	abc123	565	18s	0.000243902206421s	18.0002s
4	123123	345	16s	0.000253200531006s	16.0002s
5	dragon	372	12m18s	0.000326156616211s	12m18.0003s
6	qazwsx	512	4h8m59s	0.000308990478516s	4h8m59.0003s
7	654321	782	29s	0.000277996063232s	29.0003s
8	harley	456	1h49m40s	0.000302791595459s	1h49m40.0003s

D. Data Analysis

At this step, the test results will be processed to be more informative. Data samples are eight (8) in total for each combination scheme. It was looked for an average value so that they can be easily compared with other combination schemes.

1) Look for the average value of each combination scheme: Before the process is done to find the average value, the total value of time which is still in the form of hours, minutes and seconds must be equalized in the form of seconds. The average is found by dividing the number of values from the entire data in one combination scheme, the amount of data, or the number of samples used. The formula used is as follows:

$$\bar{x} = \frac{\sum x}{N}$$

For,

$\bar{x}$ = Average value

$\sum x$ = Total value of x

N = the amount of data

The results of the average calculation for each scheme can be seen in Table VI.

2) Average Graph: In Table VI we can see the comparison of the average time needed to get the plaintext from each scheme. The average time indicates how strong the scheme has been designed. The higher required time can be interpreted the stronger the scheme.

The average time graph from Table 6 can be seen in Fig. 2 below:

TABLE. VI. THE AVERAGE ATTACK TIME FOR EACH SCHEME

No.	Scheme	Average Time (s)
1	1	3120.12
2	2	4832.50
3	3	4483.75

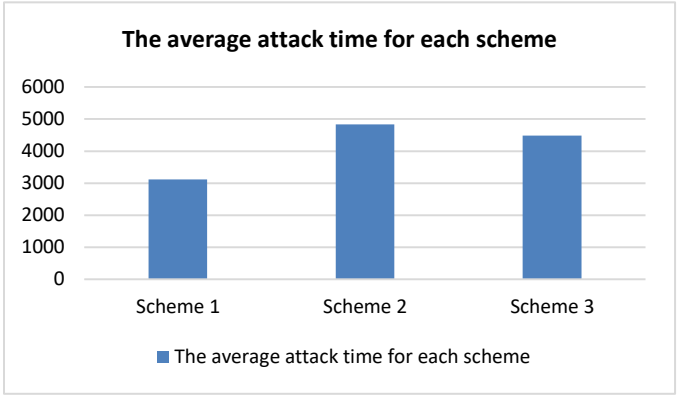


Fig. 2. The Average Attack Time for Each Scheme.

V. CONCLUSIONS

In this paper, we evaluate the strength of several combination schemes between passwords and salts. The scheme is tested based on the arrangement of the position exchange of each index between the password and the salt. With this research, it is hoped that knowledge can be obtained. The preparation and the addition of a few parameters can significantly increase the strength of the hash algorithm.

The hope for the future, there is further research on this field. For example, by adding another scheme, other parameters or with a truly random arrangement. Because this research is important in data security.

REFERENCES

[1] W. Stallings and L. Brown, Computer Security: Principles and Practice, Global Edition. 2015.

[2] N. Mouha, M. S. Raunak, D. Richard Kuhn, and R. Kacker, "Finding Bugs in Cryptographic Hash Function Implementations," IEEE Trans. Reliab., vol. 67, no. 3, pp. 870–884, 2018.

[3] P. P. Pittalia, "A Comparative Study of Hash Algorithms in Cryptography," vol. 8, no. 6, pp. 147–152, 2019.

[4] M. Cindy, A. Kioon, Z. Wang, and S. D. Das, "Security Analysis of MD5 Algorithm in Password Storage Security Analysis of MD5 algorithm in Password Storage," no. February 2013, 2015.

[5] S. Qadir and S. M. K. Quadri, "Information Availability: An Insight into the Most Important Attribute of Information Security," J. Inf. Secur., vol. 07, no. 03, pp. 185–194, 2016.

[6] J. Andress and S. Winterfeld, "The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice: Second Edition," pp. 1–217, 2014.

[7] P. Ramos Brandão, "The Importance of Authentication and Encryption in Cloud Computing Framework Security," Int. J. Data Sci. Technol., vol. 4, no. 1, p. 1, 2018.

[8] M. Trnka, T. Cerny, and N. Stickney, "Survey of Authentication and Authorization for the Internet of Things," Secur. Commun. Networks, 2018.

[9] A. Sahu, "Review Paper on Secure Hash Algorithm With Its Variants International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) Review Paper on Secure Hash Algorithm With Its Variants," no. May 2017, pp. 0–7, 2018.

[10] R. Mohanty, N. Sarangi, and S. K. Bishi, "A secured Cryptographic Hashing Algorithm," Analysis, p. 4, 2010.

[11] A. K. Kasgar, M. K. Dhariwal, N. Tantubay, and H. Malviya, "A Review Paper of Message Digest 5 (MD5)," Int. J. Mod. Eng. Manag. Res., vol. 1, no. 4, 2013.

[12] A. Bhandari, M. Bhuiyan, and P. W. C. Prasad, "Enhancement of MD5 Algorithm for Secured Web Development," J. Softw., vol. 12, no. 4, pp. 240–252, 2017.

[13] H. Mirvaziri, K. Jumari, M. Ismail, and Z. M. Hanapi, "A new hash function based on combination of existing digest algorithms," 2007 5th Student Conf. Res. Dev. SCORED, no. December, pp. 1–6, 2007.

[14] P. N. Patel, J. K. Patel, and P. V Virparia, "A Cryptography Application using Salt Hash Technique," Int. J. Appl. or Innov. Eng. Manag., vol. 2, no. 6, pp. 236–239, 2013.

[15] S. Kharod, N. Sharma, and A. Sharma, "An improved hashing based password security scheme using salting and differential masking," 2015 4th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015, pp. 1–5, 2015.

[16] E. M. W. R. Chowdhury, M. S. Rahman, A. B. M. A. Al Islam, and M. S. Rahman, "Salty Secret: Let us secretly salt the secret," Proc. 2017 Int. Conf. Networking, Syst. Secur. NSysS 2017, pp. 115–123, 2017.

[17] J. Zhang and S. Boonkrong, "Dynamic salt generating scheme using seeds warehouse table coordinates," 2015 IEEE 2nd Int. Conf. InformationScience Secur. ICISS 2015, 2016.

[18] A. Karrar, T. Almutiri, S. Algrafi, N. Alalwi, and A. Alharbi, "Enhancing Salted Password Hashing Technique Using Swapping Elements in an Array Algorithm," vol. 8491, pp. 21–25, 2018.

[19] D. Mirante and J. Cappos, "Understanding Password Database Compromises Technical Report," Tech. Rep. TR-CSE-2013-02, Polytech. Inst. NYU, 2013.

[20] K. Chanda, "Password Security: An Analysis of Password Strengths and Vulnerabilities," Int. J. Comput. Netw. Inf. Secur., vol. 8, no. 7, pp. 23–30, 2016.

[21] Radix, "Hashcat User Manual," no. August, 2011.

Capturing Software Security Practices using CBR: Three Case Studies

Ikram Elrhaffari[1], Ounsa Roudies[2]
Ecole Mohammadia d'Ingénieurs, SIWEB Team
University Mohammed V
Rabat, Morocco

Abstract—Generally, software security can be regarded as one of the most important issues in software engineering field since it may affect the software product effectiveness due to the various technological vulnerabilities and menaces. Most traditional software security approaches provide security activities through the software development lifecycle (SDLC) from requirements to design, implementation, testing and deployment. This paper focuses on embedding security concerns in the software development lifecycle (SDLC) using a bottom-up approach that is based on case based reasoning (CBR) paradigm. Thus, we study three high security-focusing cases for software projects, namely "e-shop", "Mobiling" and "intranet" using a structured case study method. Then, we populate these three cases in the proposed framework that is an excerpt of the case project base. Furthermore, this paper identifies the specificity of each case, discusses completeness of the proposed framework and proposes suggestions for improvement. Finally, usages scenarios are defined sustaining the use of the proposed framework.

Keywords—CBR; project features; case base; e-shop; mobiling; intranet; mutualize; security practices; security requirements

I. Introduction

Software engineering security has been discussed in many works with different perspectives. There are numbers of top-down security engineering approaches that cover the entire secure development life-cycle [1][2][3][4]. However, to the authors' knowledge, non bottom-up approaches have been performed on secure software development processes. In that sense, we put more focus on empirical approaches to consider security concerns in software development field.

The software engineering team has to learn more about software security in order to adopt, express, conduct, apply, review, and judge it properly [5]. According to [6] each technical member of a project (developer, tester, etc) should have a basic software security knowledge including concepts like security design, threat modeling, secure coding, security testing, etc.

The original intention was to act in the real world and change the software engineering team attitudes toward security aspects. So, the overall objectives were as follows:

- Allows project managers to maintain security practices repository for software engineering projects;
- Assist software engineering team in order to select and apply security practices in building software products.

The proposed bottom-up approach was inspired from the case based reasoning paradigm (CBR) [7][8][9]. The idea is to learn through experience and use it to make a new one successful. In this work, we focus on the first and second steps in the CBR cycle that correspond to retrieve and reuse security requirements and practices suggested by similar previous project cases.

In this way, we propose a framework that is an excerpt of the case project base. It helps the software engineering team to populate a project case base with the pertinent elements. It serves as a support to describe all project case characteristics that correspond to case descriptors. The proposed framework is structured as a class diagram, which models the project case features, in particular, security requirements and practices.

According to [10], structured method is adopted to select and design the case studies and to collect the related data. We emphasize on case quality that is indeed critical. Therefor, qualitative aspect is more important for us than quantitative. We experiment the proposed framework using three real case studies: "e-shop", "mobling" and "intranet". We notice that these cases have been selected due to their high level of reusability. In addition e-shop, mobling and intranet cases are security focused. Next, the framework is reviewed in order to ensure completeness. At the end, some usage scenarios are provided to ensure the use of the framework (see Fig. 1).

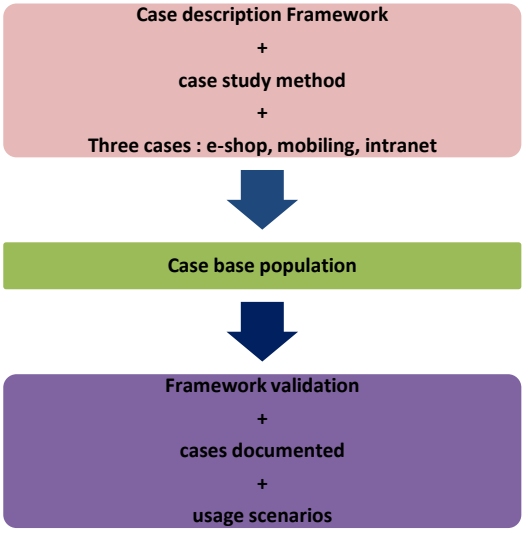


Fig. 1. The Proposed Approach.

The remainder of this paper is organized as the following: Section II presents the methodological approach that was adopted for the cases selection. Section III describes the "e-shop", "Mobiling" and "intranet" case studies through the proposed framework. In Section IV we discuss the completeness and validity of the framework. Section V presents some usage scenarios of the framework. Section VI provides a review of related studies. Finally, Section VII summarizes and set plans for future research.

II. Case Study Method

Yin [1] defines case study as "an empirical enquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident". In software engineering field, this can be particularly verified. In [1], the authors provide guidance for case study research in software engineering. this paper, we have followed the main suitable steps of this method.

A. Rationale

Case-based reasoning (CBR) can be seen as "solving a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation" [11]. In order to benefit from CBR in security engineering field, we propose a structured framework for software projects that describes the set of project characteristics, in particular, security requirements and related security practices (see Fig. 2). The idea is to retrieve security knowledge in similar previous situation (project cases) and reuse this knowledge in new project cases.

B. Objectives

The original objective is to investigate how to change the software engineering team attitude toward security and how to assist them to perform security practices through the entire software development life cycle. The objectives of this research are as the following:

- Objective 1: Validate the completeness of the framework and reveal any lack, ambiguity or superfluous elements;
- Objective 2: Populate the case base by the three cases e-shop, Mobiling and intranet;
- Objective 3: Illustrate use of the framework by two scenarios.

C. The Cases and there Context

In this paper, three projects are studied in a Moroccan company in three different application domains, both security focusing and using security practices. The initial definition of the context was only tentative to define all security project characteristics.

The company operates in the postal, banking and e-gov services. It takes careful consideration of information security issue and adopted ISO 27000 standard [12] since 2009. It is involved in global Moroccan information security concerns. The company is involved in the governmental initiatives for enforcing Digital Economy. We notice that first author of this paper is project manager and software practitioner in the company in which these projects took place.

The cases e-shop, Mobiling and intranet have been selected due to their high level of reusability. In addition, these projects are both security focusing and absolutely vital for the selected company. Finally, these three projects relates to heterogeneous domains of the company's management.

D. Theoretical Frame of Reference

Recent literature review reveals several research studies tackling software engineering security issues. There are several works about embedding security in software development lifecycle, but they often propose an ordered series of security activities that start from the SDLC input and are closely linked to SDLC phases [3][5].

Although extra attention is given by the academic literature to security engineering issues, there is a lack of theories devoted to bottom-up approaches. These approaches starts with real engineering projects and aims to capture and mutualyse organisation's know how. So, the idea is to retrieve security practices in order to capitalize them for new similar project cases. We build a case description framework according to Case Based Reasoning (CBR) theory.

E. Research Questions

We focus on the following research questions:

- RQ1: What are security requirements/practices captured from each case?
- RQ2: Is the framework accurate i.e. properly describing these security practices and the main case characteristics?
  - Is the framework appropriate and comprehensive?
  - If not, what are its limitations?
  - Are there any lacks or ambiguous aspects?
  - Are there any superfluous elements?
- RQ3: Are the selected cases pertinent?

F. Data Collection and Case base Populating

Data were collected through interviews with some project stakeholders, in particular, Project managers and software engineers and security engineers. we also received some documentations and deliverables from the project stakeholders.

Overall, the time needed to carry out interviews and to collected all possible data was 5 days hours.

III. Case Base Description and Population

In this section, we will use the proposed framework for mutualizing the selected cases and for illustration of applicability. Fig. 2 presents the proposed framework that is an excerpt from a global base case. This framework includes all classes and attributes needed to describe and document the case projects. The abstract class "feature" represent all case characteristics. We can clearly observe three main classes : "scope", "requirements" and "progress". The class scope is

used to define the case. it provide specific features like “man days”, “title”, “software engineering method”, etc. The class "requirements" is a special class that describes all project requirements. The class "progress" provides an overview of all characteristics defined in the project progress including steps and delivrables. The association "perform" links a security requirement to its security practices. This association allows to access performed security requiremements through the SDLC lifecycle.

We notice that the propoed framework is a template in which generic features can be enriched by additional information, so to provide more specific features. For example, we separated the generic feature “requirements” as “functional requirements” and “non functional requirements” (see Fig. 2).

According to the framework, we populate the Tables I, II, and III from metadata sources in order to characterize these different types of projects with the typical values for each feature. Table I illustrates an excerpt from case base that descripbe the “project scope” feature and related sub-features. We observe that, for each case, wide scope of functionalities are gathred. The “man day estimation” feature gives an indication about the size of the project.

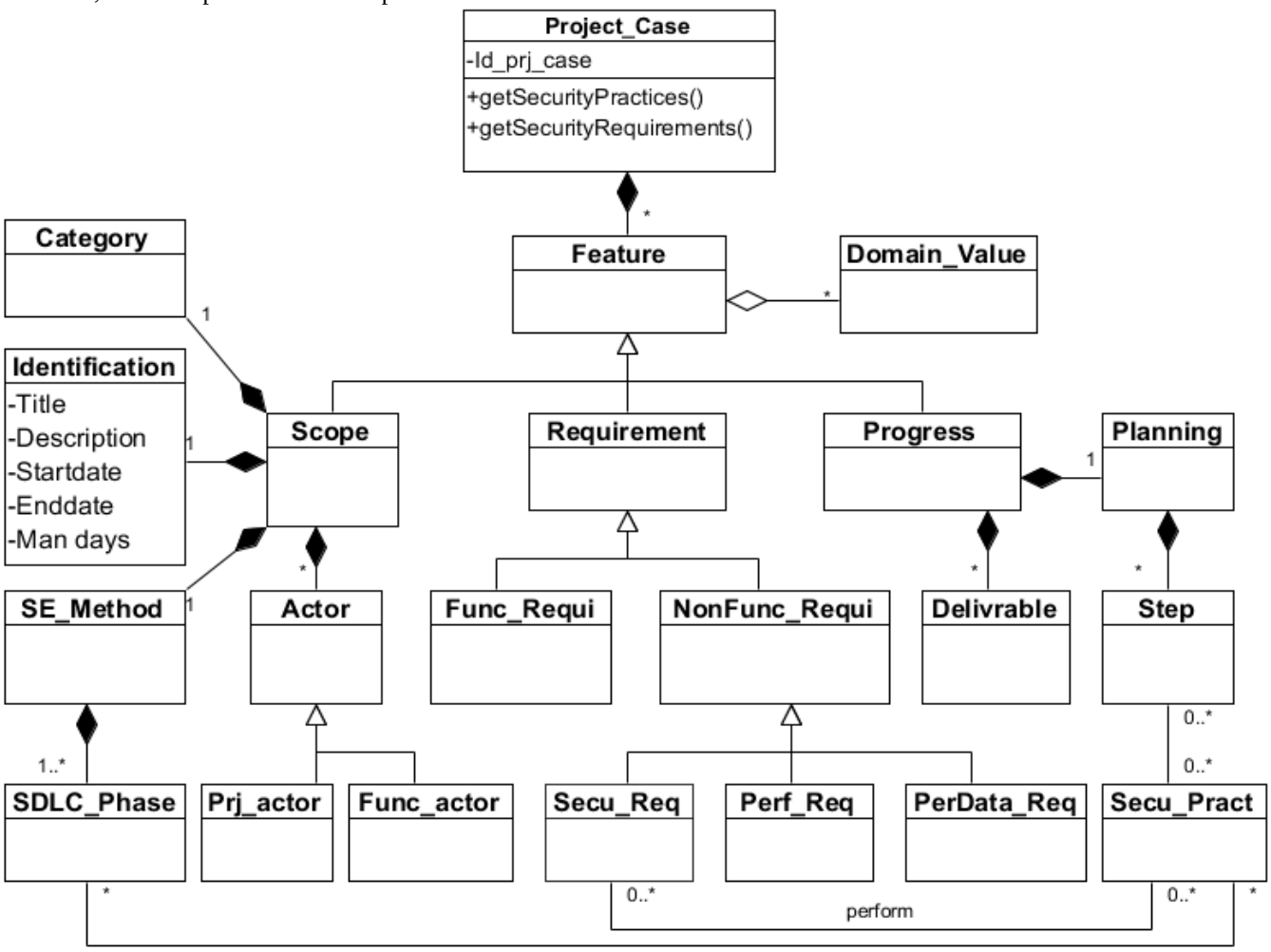


Fig. 2. Case Description Framework.

TABLE. I. EXTRACT FROM CASE BASE : PROJECT SCOPE

Project scope				
Section	***e-Shop project***	***Mobiling project***	***Intranet project***	***Values domain***
1. Project title	e-shop	Mobiling of postal distribution activities	Intranet	
2. Description	e-commerce website for third	Provide a solution that allows postmans to carry out their operations in the field using a mobile terminal	Provide a collaborative space within an organization, Various stakeholders are involved.	
3. Man days estimation	270	360	750	
4. Functionalities	e-marketing (promotion, best sales); e-Shop customer registration; basket management; Online payment; Order tracking; Merchants registration in Interbank center; Newsletters management (customer and others)	Mobiling of postal distribution activities; Monitoring postman activity; Frontoffice and backoffice solutions	Suggestion box, forum for the discussion,Newsletter, human resources data, exchange of experience through shared platform, intranet document repository	
5. Customer (target)	Enterprises (cooperatives…)	Postman, supervisors	Employees	Employees, enterprises
6. Actors				
6.1. Functional actors	merchant, customer, functional administrator, technical administrator, commercial	Postman, supervisor, manager, functional administrator	Employees, functional approver, functional administrator, technical administrator	Functional administrator, technical administrator, manager
6.2. Project actors	Project manager(provider side), developers, system administrator, Project manager (owner side), tester	Project manager (provider side), developers, system administrator, Project manager (owner side), security advisor, tester,	Project manager (provider side), developers, system administrator, Project manager (owner side), tester	Developer, project manager, system administrator, tester
7. Project type				
7.1. Category 1	Outsourcing	Outsourcing, backsourcing, internal development	Outsourcing , internal development,	Internal development, outsourcing, backsourcing
7.2. Category 2	Web-based	Domestic, mobile	Domestic	Web-based, domestic, mobile, app for smartphone, app for tablets
7.3. Category 3	Third	Business	Collaborative	Third, e-gov, collaborative, business
8. Software engineering method	Agile	Agile	UP	Agile, UP
9. Software lifecycle coverage	Total	Total	Total	Total, partial
10. Tools/technologies	Websphere e-commerce, DB2 (IBM), Jee, web services	Visual studio .NET, windows mobile, SqlServer, GPS, 3G, 4G, Wifi	SharePoint, SqlServer	

Table II contains a section for each requirement category. Across interviews, we identified three non functional requirements categories : security requirements, performance requirements and personal data requirements. We can conclude that several security requirements are commun to both the three cases, other security requirements are expressed for a specific use. For example the security requirement “the connection to the payment website should be encrypted by a certificate delivered by a certificate authority” is dedicated for “online payment” funcltionality. We can clearly distinguish between organisationla and operational security requirements. An example of organizational security requirement is “demilitarized zone (DMZ) Front Office must be isolated form demilitarized zone (DMZ) Back office”. This requirement must be performed when we have two platforms : one exposed to internet and the other reserved to backoffice users and services.

Performance requirements are crucial when we cannot predict simultaneous user connections. The typical example is: e-shop case project.

In Table III, we describe some elements dealing with project progress. For instance, one of these features is “deliverables” which provide a guidance for each step in the project planning. In addition, deliverables serve as a gateway between all project stakeholders.

TABLE. II. EXTRACT FROM CASE BASE : REQUIREMENTS DESCRIPTION

Requirements				
Section	***e-Shop project***	***Mobiling project***	***Intranet project***	***Values domain***
11. Functional requirements	- Display category and related products - Customer registration - User registration for newsletter - e-marketing (promotion, best sales) - Reporting (orders, customers…)	- Automatic Transmission of the postman rounds on their mobile devices - delivery operations using mobile terminal and transmission of real-time information - Monitoring postman activity	- All stakeholder contributions must be validated by the functional approver -	
12. Non functional requirements				
12.1. Security requirements	- Financial transactions between the customer's bank and the merchant bank on a closed network and not open on the Internet - Merchant does not have access to banking information of its customers either in consultation neither in treatment - The connection to the payment website should be encrypted by a certificate delivered by a certificate authority - All transactions must be traced (log files, logging) - Licensing and code ownership	- Automated connectivity with online and offline modes - User access must be restricted - All data exchange must be encrypted - demilitarized zone (DMZ) zone Front Office must be isolated form DMZ Back office - The password rules must be in line with the organism password policy - The reset or change of password must be available for users - All security rules must follow the organism security policy - contractual requirements for quality and code security level	- Single sign on (SSO) - User authentication system includes Lightweight Directory Access Protocol (LDAP) directory - User access must be completely restricted - Data must be transmitted to the user over a secure connection - External access must be denied - Code Test must be performed to detect potential malware or Trojan horse code	
12.2. Privacy [13]	- The compliance process must be initiated	- The compliance process must be initiated		
12.3. performance requirements	- 1000 simultaneous user connections	- 500 mobile users, 100 BackOffice users - Fast response time		

TABLE. III. EXTRACT FROM CASE BASE : PROJECT PROGRESS

Project progress				
Section	***e-Shop project***	***Mobiling project***	***Intranet project***	***Values domain***
13. Planning (macro steps)	Project scope, requirements analysis, design, development, test, production	Same	Same	Project scope, requirements analysis, design, development, test, production
14. Deliverables	Installation procedure, backup procedure, user guide, acceptance document	User guide, operation procedures, acceptance document, functional specification document, technical specification document	Acceptance document, functional specification document, technical specification document	User guide, acceptance document,
15. Tests				
15.1. Functional /Non-functional tests	See Table I			
16. Pre-production				
16.1. incidents/bugs	some product images are not displayed, adding individual product is not operational	Mobile configuration are not adapted, some data are not updated on the mobile	intranet document repository link is not available,	
17. Post-production				
17.1. incidents/bugs	Low performance of web service calculating the charge of distribution; hard disk space exceed the limit	unavailability or data transmission, unavailability of the network	Unauthorized user can access to restricted data,	

IV. Discussion

After studying, documenting the cases and clarifying the need of a project case base, the proposed framework must be verified. The validation will be done by answering the research questions defined in Section II.

- **RQ1:** What are security requirements/practices captured from each case?

The interviews which we conducted give information on security requirements and practices related to the selected cases. Security requirements are also described in the "functional specification document", but no document was found for security practices or how they perform the security requirements (see Fig. 3, "perform" association). We present in Table IV, the main security practices through interviews and related SDLC phases.

- **RQ2:** Is the framework accurate i.e. properly describing these security practices and the main case characteristics?

On the positive side, the framework has been defined to guive some guidance on how to document the cases. Thus, fundamental knowledge has been gathered during the case studies and the execution of framework experiment (see Tables I, II and III). This knowledge includes generic features (scope, requirements and progress) and specific features (software engineering method, functional requirements, security requirements, project actors, functional actors, etc.).

On the negative side, it is interesting to observe that, the framework lacks several elements that could play a leading role in documenting and selecting cases. Thus, the framework shows how the security requirements can be integrated through the SDLC phases (see "perform" association). However, the framework does not explicit who is performing these requirements (involved roles).

Another two important features of a case base of security projects are security bugs and incidents. So, to ensure a high security level of the software product, software team has to learn about security bugs and incidents of previous cases, in particular critical ones. The framework does not seem to cover this. In Table V, we illustrate some examples of security bugs and incidents that are collected through the case studies.

The class "category" in the framework describes the case category. This feature is not fully clear since many interpretations were provided by the interviewers. Fig. 3 shows example values for this feature. Thus, we suggest a three-level categorization for this feature.

There is a correspondence between security and technologies used in almost all software project and. For example, when we use web services, several security activities must be performed to deal with this technology. Therefor we can capitalize them for future similar situations.

In general, additional classes must deal with all security aspects involved in a project case including actors, technologies, project categories, bugs and incidents, etc. these features will be used as case descriptors (characteristics) for eventual new cases in order to perform similarity rules and to select the suitable similar cases.

Depending on the kind of the organization and its main strategies, there is a set of security aspects related to software security. For example, standards, policies and guides, requirements, principles, practices and activities.

The framework has very limited support in this way. The only aspect that the framework covers is that the security practices should perform the secure requirements. In the authors' opinion, the framework must deal with all software security issues.

TABLE. IV. The main Security Practices Gathered

Project case	*security practices*	*SDLC phases*
Mobiling	Use secure coding guidelines during implementation	Implementation
	Reduce priviliges	Design
	The reset or change of password must be available for users	Design
E-shop	Specify operational environment	Design
	Reduce priviliges	Design
	Perform manuel code inspection	Implementation
	Perform penetration testing	Test
	Perform database log files	Implementation, Maintenance
Intranet	Reduce priviliges	Design
	Perform database log files	Implementation, Maintenance
	User authentication system includes LDAP directory	Design, implementation

TABLE. V. Examples of Security Bugs and Incidents Gatherded through the Case Studies

Project case	**Security bug/incident (pre and post-production)**
e-shop	• Low performance of web service calculating the charge of distribution • Hard disk space exceed the limit
Mobiling	• Unavailability or data transmission • Unavailability of the network
Intranet	• Unauthorized user can access to restricted data

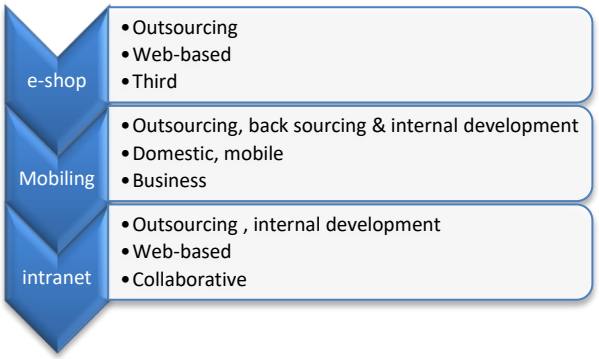


Fig. 3. Values Gathered for the "Category" Feature.

- **RQ3:** Are the selected cases pertinent?

The case pertinence can be analysed from different points of view. On the one hand, the selected cases are pertinent as they are leader in their respective fields. For instance, if you start any new project using mobile technology, you can find guidance about deadlines, project actors, requirements, security practices, technologies, etc. To assess the optimal reuse of both the "Mobiling" and "e-shop" cases, we introduce in Section V, a usage scenario that provide baselines about security requirements to be performed.

TABLE. VI. QUALITY EVALUATION OF CASE STUDIES

Quality criteria	*e-shop*	*Mobiling*	*Intranet*
A theoretical basis including research questions is described	3	3	3
Triangulation is ensured by using multiple sources of evidence (data collection and interpretation)	1	2	1
A chain of evidence is designed with traceable reasons and arguments	2	2	2
The case study research is fully documented	3	3	3
The case study report is compiled through an iterative review and rewriting process	0	0	0

On the other hand, software engineering team is not oriented to achieve security activities. We observed this in the context of the case studies. Unfortunately, the selected cases are not appropriate to provide guidance on software security. Thus, one of the problems that we have encountered is the weakness of a strucuctured approach for embedding security in software engineering field.

Author in [1] summarizes Yin's quality criteria for a good case study. In Table VI, we evaluate the case studies regarding these criteria. We consider 4 levels to evaluate the case studies as follow:

- Level 0: Lack; Level 1: low; Level 2: medium; Level 3: strong.

V. EXPERIMENT : REUSE OF CASE KNOWLEDGE

We present here briefly some usage scenarios of the cases knowledge.

A. *Reuse scenario 1 : Identify Security Requirements for Mobiling e-Shop Project*

In this scenario, we consider a new e-shop project on mobile. A typical question that can be asked by the project manager is: what security issues should we focus on? So, for a project team, it is of crucial importance to have access to security requirements and best practices identified in both e-shop and Mobiling experiences.

In Fig. 4, we illustrate this scenario and we represent some useful security requirements for the Mobiling e-shop project.

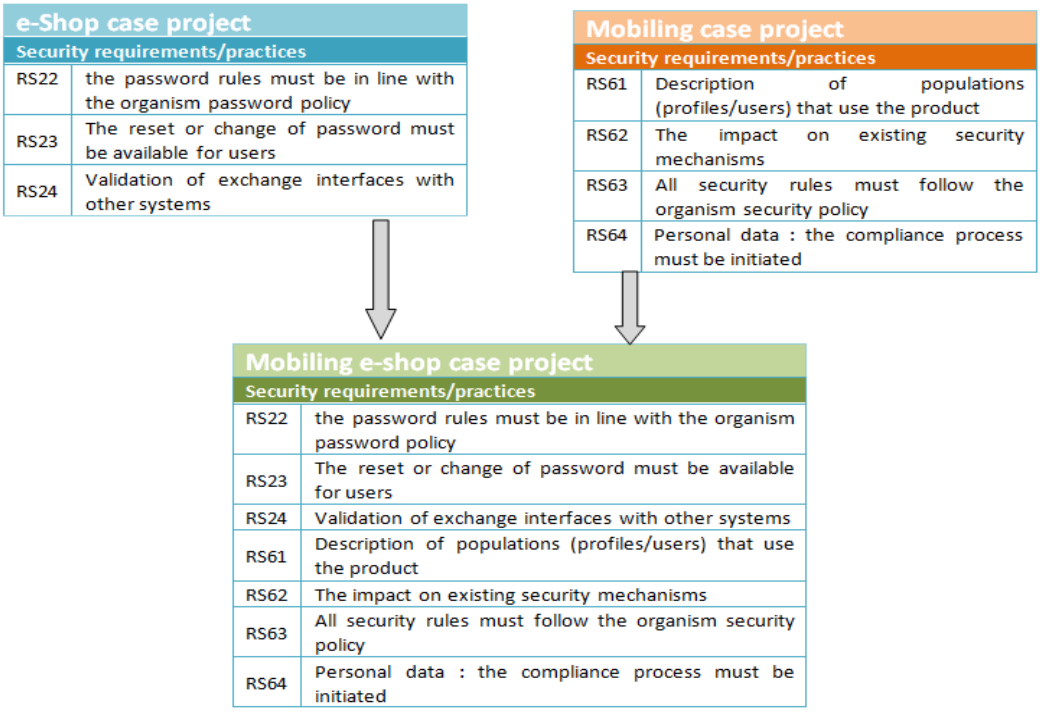


Fig. 4. Capturing Security Requirements for Mobiling e-Shop Project.

B. Reuse scenario 2 : Security Requirements for Outsourcing Projects

According to Table I, the “e-shop” case is only outsourcing, while the "Mobiling" and "Intranet" cases are both outsourcing and internal projects. Thus, by analyzing the Tables I and II, we can conclude that several security requirements and derived security best practices are extremely important when we decide to develop an outsourcing project, for example:

- An outsourcing project is managed by several contract documentations such as “contractual requirements for quality, code security level”, “licensing”, “code ownership” and “intellectual property rights”.
- All security rules must follow the organism security policy. In particular, The password rules must be in line with the organism password policy.
- Code test to detect potential malware or Trojan horse code.

VI. Related Works

Our research builds on multiple streams of related research, including software security, reuse, case based reasoning, and case study research.

In the context of software security, a good number of approaches and tools have been proposed to support security into the software development lifecycle. These approaches tend to be centered on building security activities and best practices gathered from standards, security processes and methodologies [5][6][13][14]. In this way, each security best practice is mapped with the corresponding phase of development. However, although security in the SDLC can be fine-built and managed by experienced teams, we have found that the integration of security best practices still difficult, even for skilled software engineering teams [6][15][16]. Reusing previews experienced best practices can reduce the complexity of security adoption in the SDLC. In that respect we propose to reuse security requirements and practices suggested by similar previous project cases.

One of the aspects that we focused on is the adoption of the CBR paradigm. In this way, there are currently several approaches focusing on the CBR approach to ensure similarities and to improve quality of search in the software engineering field [7][8][17].

As mentioned in section II, the case study methods provide guidelines that are helpfull to researchers when deciding how to select and evaluate case studies. Author in [10] outlined a case studiy method in software engineering and provided usefull examples.

VII. Conclusion

Security aspects are critical to secure software products and to meet end-user needs. The global aim is implementing a detailed case base for security focused software project that helps both software engineering and security teams to maintain a knowledge repository.

In this work, we have tested the effeteness of the proposed framework that is an excerpt of the case base through three cases e-shop, mobiling and intranet. These cases were selected using a structured case study method. We recall that we conducted this empirical study within a Moroccan company.

As a validation of the framework, we discuss a number of fundamental elements including cases pertinence, framework lacks and completeness. We also suggest improvements, which the framework could benefit from. At the end, we provide two usage scenarios to illustrate of the effectiveness of the framework.

This study indicates that several security requirements and practices can be adapted for specific projects. However, It is clear that both software engineering and security teams have to work synchronized with an incremental approach in order to provide a comprehensive project cas base.

This work can serve as a good reference for many organisations to implement their own case base for security focused projects. In this way, the authors plan to experiment this work based on validation by panels. Furthermore, the development of a software tool to support the managenemnt of the case base will be initiated.

References

[1] Mohammed, N. M., Niazi, M., Alshayeb, M., & Mahmood, S. (2017). Exploring software security approaches in software development lifecycle: A systematic mapping study. Computer Standards & Interfaces, 50, 107-115.

[2] Baca, D., Boldt, M., Carlsson B., Jacobsson, A.: A novel security-enhanced agile softwaredevelopment process applied in an industrial setting. In: ARES 2015, pp. 11–19 (2015).

[3] Moradian, E., 2012. Integrating Security in Software Engineering Process: The CSEP Methodology. Ph.D. dissertation, KTH Royal Institute of Technology, Sweden.

[4] D. Noopur, Secure Software Development Life Cycle Processes, Published: July 05, 2006 | Last revised: July 31, 2013, Official website of the Department of Homeland Security, Available on https://buildsecurityin.us-cert.gov/articles/knowledge/sdlc-process/secure-software-development-life-cycle-processes.

[5] Mohammed, N. M., Niazi, M., Alshayeb, M., & Mahmood, S. (2017). Exploring software security approaches in software development lifecycle: A systematic mapping study. Computer Standards & Interfaces, 50, 107-115.

[6] Julia H. Allen, Sean Barnum, Robert J. Ellison, Gary McGraw, Nancy R. Mead (2008). Software Security Engineering: A Key Discipline for Project Managers. part of the IEEE Reliability Society 2008 Annual Technology Report.

[7] Bitar, Ibrahim EL. CBR4WSD: Une approche de découverte de services Web par Raisonnement à Partir de Cas. Diss. Ecole Mohammadia d'Ingénieurs-Université Mohammed V de Rabat-Maroc, 2014.

[8] Martin Andreas, Sandro Emmenegger, and Gwendolin Wilke. "Integrating an enterprise architecture ontology in a case-based reasoning approach for project knowledge." Enterprise Systems Conference (ES), 2013. IEEE, 2013.

[9] de Mantaras, Ramon Lopez. "Case-based reasoning." Machine Learning and Its Applications. Springer Berlin Heidelberg, 2001. 127-145.

[10] RUNESON, Per, HOST, Martin, RAINER, Austen, et al. Case study research in software engineering: Guidelines and examples. John Wiley & Sons, 2012.

[11] A. Aamodt and E. Plaza, “Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches,” Artificial Intelligence Communications, vol. 7, no. 1, pp. 39–59, 1994.

[12] Alexandre Fernandez Toro, Management de la sécurité de l'information, implémentation ISO 27001 : Editions Eyrolles, 2018.

[13] Michael Howard and Steve Lipner. The Security Development Lifecycle, Microsoft Press, 2006.

[14] Uzunov, Anton V., Katrina Falkner, and Eduardo B. Fernandez. "A comprehensive pattern-oriented approach to engineering security methodologies." Information and Software Technology 57 (2015): 217-247.

[15] Busch, Marianne, Nora Koch, and Martin Wirsing. "Evaluation of engineering approaches in the secure software development life cycle." Engineering Secure Future Internet Services and Systems. Springer, Cham, 2014. 234-265.

[16] Mazni Mohamed Jakeri, Mohd Fadzil Hassan, "A Review of Factors Influencing the Implementation of Secure Framework for in-House Web Application Development in Malaysian Public Sector", Application Information and Network Security (AINS) 2018 IEEE Conference on, pp. 99-104, 2018.

[17] Benjamin Maraza-Quispe, Olga Melina Alejandro-Oviedo, Walter Choquehuanca-Quispe, Alejandra Hurtado-Mazeyra and Walter Fernandez-Gambarini, “e-Learning Proposal Supported by Reasoning based on Instances of Learning Objects” International Journal of Advanced Computer Science and Applications(IJACSA), 10(10), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101035.

An Efficient Model for Medical Data Classification using Gene Features

Ensemble Model for Text Classification (EMTC)

Kosaraju Chaitanya[1], Rachakonda Venkatesh[2]
Department of CSE
Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur-522009
Andhra Pradesh
India

Thulasi Bikku[3]
Department of CSE
Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur-522009
Andhra Pradesh
India

***Abstract*—In the medical field to solve the new issues, the novel approaches for managing relevant features by using genomes are considered; using the sub-sequence of genes the outcome of interest is analyzed. In this implementation part of the model, we have given the MEDLINE and PubMed archives as inputs to the proposed model. A large number of MESH terms with gene and protein are utilized to characterize the patterns of a large number of medical documents from a large set of records. Standard datasets with different characteristics are used for examination study. The characteristics and inadequacies of different techniques are noted. Feature selection techniques are given in perspective of data composes and region traits by applying proper rules. Feature context extraction through name element distinguishing proof is an essential errand of online therapeutic report grouping for learning disclosure databases. The parameters are identified to compare with other models implemented on these datasets and the results prove that the proposed method is very effective than existing models. The primary point of the proposed ensemble learning models is to characterize the high dimensional information for gene/protein-based disease expectation in light of substantial biomedical databases. The proposed model uses an efficient ranking algorithm to select the relevant attributes from a set of all attributes; the attributes are given to the classifier to improve the accuracy based on the users' interest.**

Keywords*—Classification; Hadoop framework; biomedical documents; feature selection; gene features; medical datasets

I. INTRODUCTION

In the bioinformatics field, the Dimensionality reduction procedures have turned into a need. The relevant features are used for effective reduction of training time and overfitting of the model. In the medical area, a colossal measure of information is produced and put away in the medicinal archives. The reports or archives contain the data about the indications of a patient and furthermore numerous medical test reports that might be produced. The dominant part of genuine classification issues require algorithm of supervised learning in which the hidden class probabilities and conditional probabilities are ambiguous, which are not clearly mentioned, and each instance in test data is related with a class name. A candidate feature is neither superfluous nor repetitive to the objective idea; an irrelevant component does not influence the objective idea at all, and irrelevant attributes or repeated attributes does not add anything new to the objective idea. In many applications, due to the large datasets, the learning algorithms never give accurate results because of irrelevant and spurious attributes, so it is essential to segregate the relevant features from large data repositories. Diminishing the quantity of superfluous/repetitive features definitely decreases the algorithm's training time and yields a negative impact on the objective function [1]. The feature selection methodology does not make novel attributes, it only chooses unique features. One of the difficulties comprise for the development of model is to choose the features when selection is required. Selections of relevant features are measured to check the properties of the data might be required to settle on a decision.

In the meantime, feature classification, selection, and feature construction can be used together to improve the execution of the objective function and lessen the dimensionality of the datasets. The model construction can be in three distinctive ways: 1) before constructing the model feature classification and selection are performed; 2) feature construction should be done; and 3) in the meantime performing feature classification, selection, and construction. These aides in showing signs of improvement knowledge into the basic idea of a true classification issue. These days, the development of the high-throughput innovations has brought about exponential development in the reaped information as for both dimensionality and test data size. Feature selection techniques endeavor to choose a subset of attributes that are applicable to the objective idea. With the quick growth of biomedical text documents, automatic text analyzing systems is increasing at an exponential rate to find relevant patterns from the distributed biomedical data [2]. Data mining and machine learning models have been successfully implemented in various bio-medical domains such as cancer detection from pathology reports, disease diagnosis and predicting complex diseases [3]. Irrelevant features not only results high true negative rate but also lead to inaccurate classification results. Text classification and feature prediction are the important issues of biomedical text documents in a large number of distributed biomedical applications [4]. Eliminating irrelevant features from the large bio-medical databases facilitates noise reduction and optimizes the classification accuracy. Distributed

data mining is a functioning exploration region for structure high computational and for proficient decision-making design patterns for learning based applications in the bioinformatics [5].

Because of the exponential development of models in the distributed environment, a large number of server farms have been shaped and collected huge measures of the information. It is important to find hidden patterns of knowledge to improve decision making from those databases in the bioinformatics [6]. The programmed arrangement of restorative reports into predefined classes is developing quickly on online biomedical repositories [7]. The current arrangements that require earlier information of classification accuracy for different sorts of relevant attributes, which is difficult to acquire day by day. The commitment of the proposed research work for biomedical databases is feature selection and classification and Hadoop-based decision tree model; which is used to construct the patterns based on the user requirements in large data sets. In this scenario, one of the most concerning issues persuaded by the exponential development of repositories about biomedical data is to help researcher in finding valuable data from the databases of MEDLINE utilizing PubMed web search tool in making decisions including functions about gene, patterns of diseases using genes, associations among the genes and MeSH knowledge discovery patterns [8].

The search results using the synonyms of drugs, genomes, disease names, and medical terms browsed by the users in the internet are shown in Fig. 1. The user searches the medical documents based on the MeSH terms or general medical terms in the web, based on the user interface the medical documents are extracted from the biomedical repositories. The document extraction is based on the ranking algorithms and synonyms of the medical terms based on the user requirement.

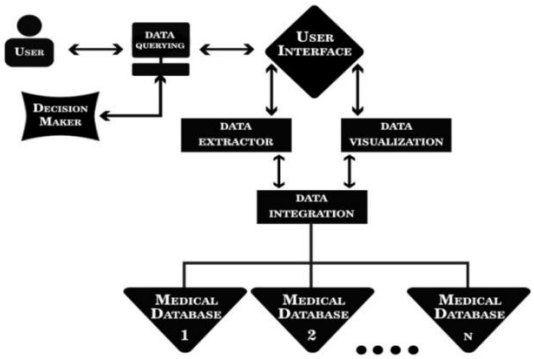


Fig. 1. Users Search for Medical Documents.

Traditional Hadoop based distributed ensemble models such as Random forest, Fuzzy c-means and Multi-objective ensemble classifier; Bayesian ensemble models were implemented to discover cluster based decision patterns with a few feature vectors. In this dynamic model, an ensemble medical document classification model was developed to discover the clustered feature vectors with patterns in the distributed medical datasets. As the traverse of the online biomedical databases is creating well ordered, analyzing interesting patterns in a fundamental structured or unstructured biomedical applications have ended up being more troublesome with traditional algorithms [9]. Additionally, because of the accessibility of various biomedical reports in the PubMed and MEDLINE documents, it is all the more testing to investigate information, foresee patterns and translate the documents data utilizing the customary information mining by using classification and clustering.

A typical framework of a distributed approach works as follows: it performs the analysis of the local dataset at each distributed peer node, and then the discovered information is migrated to a central site where the integration of the distributed local nodes is performed. The computed results are returned to the distributed data sets so that all nodes contain the updated information. Peer-to-peer (P2P) medical system involves a large number of peers in the network without any centralized control, in which each peer node exchanges and share data throughout the network. Each peer node is linked directly to a large number of peers within the overlay network. It is impractical to gather all the data distributed in the P2P network into a centralized node or site and then perform the conventional data mining techniques. P2P overlay networks emerged as a promising mechanism to manage and share the high dimensional data in the client-server architecture [10]. A high-performance Hadoop framework has become extremely important for most of the large scale data applications such as Medline and Pubmed [11].

A typical framework of a distributed approach works as follows: it performs the analysis of the local dataset at each distributed peer node, and then the discovered information is migrated to a central site where the integration of the distributed local nodes is performed. The main ideology behind implementing the distributed environment in bioinformatics is easy and effective to implement, reliable, fault tolerance, scalable in different dimensions. To process the large amounts of data, the single cannot run in stipulated time, it has many limitations like processing speed, storage etc., so we go for distributed environment to retrieve the knowledge patterns. The biomedical datasets contain heterogeneous data, the data from multiple sources must be integrated and the data is shared between the repositories. The primary sources of medical databases are GENEBANK, EMBL, NCBI, NDB and SWISS-PROT. The DNA information of the living being is encoded by genomes, to know about their heritage of gene, so that it will be easy to predict their diseases and health conditions of the patient.

II. Literature Survey

The literature survey on biomedical repositories contains the information about the preprocessing techniques of the medical documents. The preprocessed data converts the raw input into output as useful structured format by avoid missing data and to provide clean data [12]. Bioinformatics is the field of study about biological data involving gene ontologies and genomics. It contains the data of DNA and protein sequences, gene expression data, images which are incomplete in nature. The bioinformatics data can be stored, managed, analyzed using algorithms and statistical methods and tools of computer science. The biomedical data is organized and stored in the databases; the new entries of data are also stored in the large repository.

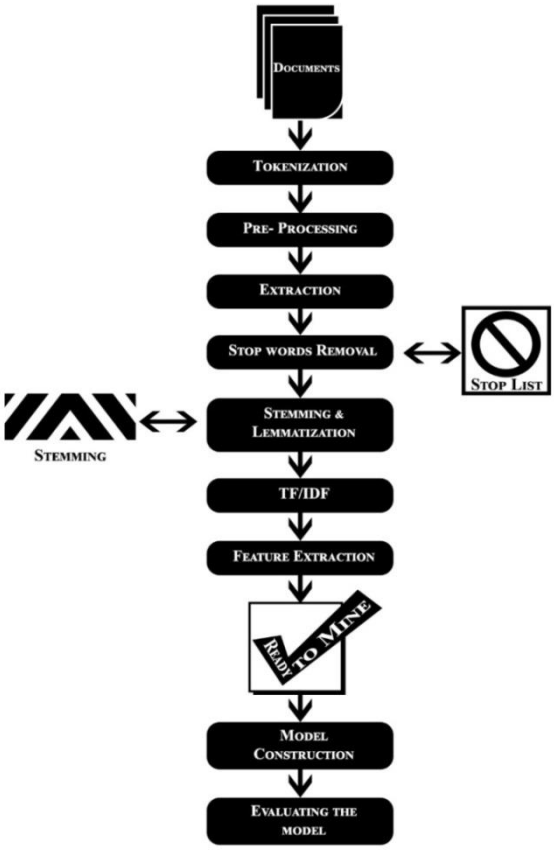


Fig. 2. The Preprocessing Steps.

The identification of gene can be done by using the name of the gene or its equivalent word in the title part of the record in the datasets [13]. The mining used to discover patterns for the discovery of genes can be performed in two stages: Initially, every one of the terms identified with genes out of a MEDLINE database is distinguished and extricated. Here, TF-IDF (Term Frequency-Inverse Document Frequency) and Z-scores methodologies are actualized for the identification of related genes. In the subsequent advance, the report preprocessing approach is executed to frame the gene features shown in Fig. 2.

The data repository stores the information of different formats like genomics, medical images, incomplete data, which should be cleaned in the preprocessing stage. Thus the data obtained having many dimensions having irrelevant data. The feature selection of data is not an easy procedure, which would not be possible manually. High quality data gives good results, removing the irrelevant data may not affect the accuracy of the classifier. The medical documents may contain the stop words i.e. they are commonly used words like articles and prepositions, sometimes contain symbols like ‘,’, ‘.’, ‘_’, ‘-’ are removed. The stemming is the process of removing the words having the extensions like ‘ed’, ‘ing’. The tf/ idfare used to identify the key terms of the medical document based on the frequency of the terms used in the document. Now the preprocessed data is ready to mine based on the user required queries.

The traditional Relational Database Management System (RDBMS) and Object-Oriented Database Management System (OODBMS) are not suitable to retrieve the genome data due to the complexity of the data. So the researches are done to discover method to store, manage the data based on the users’ requirement by extending the existing methods. The pre-analysis of information consequently chooses or unselects various alternatives and gives valuable data to the client dependent on his prerequisite. It begins by checking if all the gene expression patterns have a similar number of conditions and a valid identifier. It rectifies them generally by altering the number of components in each row and including a unique identifier when necessary.

Map-Reduce are a framework that enabled the processing of large-scale databases for the first time [14]. This framework processes datasets in a distributed platform with introducing two phases, Map and Reduce functions. Named entities in the medical text documents are structured and unstructured due to its boundary detection and feature classification. Pietari Pulkkinen and Zhihua Wei, 2008 implemented name entity identification approaches from biomedical databases using support vector machine classifier [15]. Junshan Zhang, 2013, proposed a probabilistic model of name entity identification with decision tree approach. This model proposes the unsupervised system to design an index based clinical documents classification from the large corpus [16]. Cheng Hua Li, 2014 implemented a topic modeling using principal component analysis to achieve correlation between the feature vectors in different databases [17]. A number of biomedical documents have been used in the development of parallel computing and data mining models on the MapReduce framework. Typical features of textual machine learning models include iterative data scanning and feature extraction, which require high performance and computation over the distributed data.

By using the Biomedical Name Entity Recognition (BNER) used to filters the various biomedical entities such as genes, proteins, cell-type, etc. by using. The taggers which are commonly used are GENIA, NLProt, ABNER, and BANNER [18]. GENIA predicts a biomedical name entity using Natural Language Processing (NLP) tasks such as tokenizing, chunking, part of speech (POS) tagging and machine learning measures for entity detection. Both ABNER and BANNER outlined and built up an algorithm for data learning model utilizing contingent irregular fields with relevant attributes [19]. Nonetheless, the parameter estimation for contingent arbitrary fields is costly, when the complexity of space and time are concerned. This issue winds up harder while training on numerous biomedical databases with constrained processing resources and assets. At long last, GENETAG is used to filter the channels of gene or protein labels from biomedical records by word reference or pattern matching coordinating models of GenBank substances.

Information extraction in the medical domain includes extraction of proteins, genes, and extraction of the relationship between these entities. Herman implemented a system that retrieves and visualizes domain knowledge from literature and database using gene names with high recall and low precision and it is difficult to detect the semantic relationships among the

entities [20]. Another approach is a rule-based technique, which aims to detect and extract bimolecular relations using inhibit relation achieved 91% of precision and 56% of recall.

The supersets of all the terms in the bioinformatics are stored in the medical database known as Unified Medical Language System (UMLS). The searching operation is performed on the UMLS to select the terms used to retrieve the medical text files. If a term is detected in the UMLS, that specific term is considered as a biomedical term and included in the document. Else, the terms are ignored and deleted from the document.

A different decision tree-based multi-objective estimation of conveyance algorithm (DT-MEDA) for improvement concerns through uninterrupted features is generated. Decision tree-based probabilistic models are utilized to convert contingent conditions stuck between factors in DT-MEDA [21]. As of late, a substantial number of statistical machine learning strategies are connected to the content classification framework. The utilization of the most punctual machine learning strategy is naive Bayes classification (NBC). Along these lines, all the essential machine learning calculations have been connected to the field of content classification, for instance, *K* nearest neighbour (KNN), neural network system (NNS), support vector machine (SVM), decision tree, kernel learning, and some others [22]. The expectation maximization (EM) algorithm is used to distinguish themes in unaligned biopolymer groupings. Late examination have given opposing outcomes when contrasting Bayesian estimators with Expectation Maximization (EM) for unsupervised Hidden Markov Model (HMM) Parts Of Speech (POS) tagging, and we demonstrate that the distinction in detailed outcomes is to a great extent because of contrasts in the training data size and the number of states in the Expectation Maximization - Hidden Markov Model (EMHMM). In the proposed model (EMTC), a large number of documents are processed with high true positive and precision. Experimental results prove that the proposed approach has a high recall, precision, and less standard error compared to conventional ensemble approaches. The proposed models used for biomedical classification using the Hadoop framework can be extended further for semantic entity mining and deep learning as follows.

III. Proposed Model

The proposed method mainly concentrates on Ensemble Model for Text Classification (EMTC) shown in Fig. 3. Let document D is segmented into a set of medical phrases or sentences in a peer document set. Let D= {p1, p2, p3,....,pn} and T={t1,t2,t3,...., tm} represents all the term occurs in a document set, where n is the number of phrases and m is the number of terms in a document D. The major challenge in the document classification is a physical representation of the document set in the distributed environment, which will not work properly in a high dimensional vector space model.

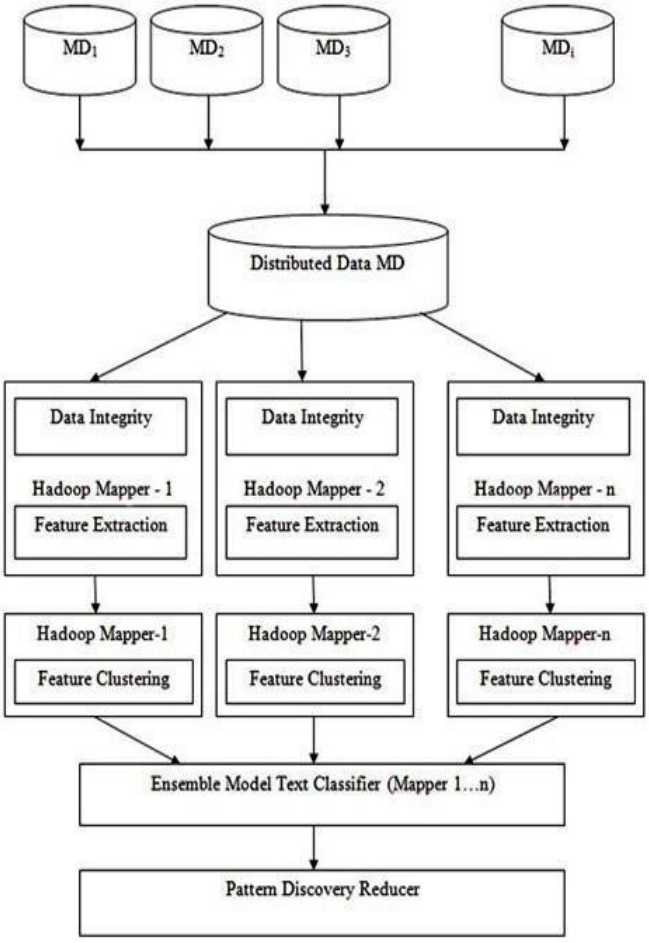


Fig. 3. Proposed Method (EMTC) for Feature Clustering.

Similarity measures are used in text related document repositories and applications such as information retrieval, language processing, text mining, and text clustering and prediction process. An enhanced MapReduce framework was used to discover the textual patterns from a large collection of structured or unstructured medical document sets. Also, multiple Hadoop mapper interfaces are assigned to each distributed data source for domain knowledge identification. Medical documents are integrated and preprocessed using the feature extraction algorithm of the Hadoop's Mapper interface; each Mapper phase is responsible to prune the structured, unstructured and inconsistent features.

After the feature extraction process, an ensemble cluster-based classification model was applied to discover the interesting patterns for medical documents. Ensemble integrates the results of multiple classifiers; thus assists to overcome the possible issues of classifiers in medical databases. The proposed model is evaluated on two datasets, namely GENETAG and Medline distributed datasets.

A. Combining Multiple Distributed Data Integration

Input: Heterogeneous Medical datasets
Output: Single repository with integrated data
Procedure: The medical repositories with names BD_1 and BD_2
for attribute-m in BD_1 do
for attribute-n in in BD_2 do
if(attribute-m≠NULL&&attribute-n≠NULL) then
if(type(attribute-m)==type(attribute-n)) then
Similarity (attribute-m, attribute-n) = (Parent(attribute-m/attribute-n)*Parent(attribute-n))*Correlation(attribute-m,attribute-n);
// the similarity between two attributes are measured
The correlated attributes are mapped using
Map((attribute-m, attribute-n), similarity(attribute-m, attribute-j));
Done;
To integrate the attributes, we merge the attributes having maximum similarity.
Select attributes having maximum similarity pair (attribute-m, attribute-n)
Repository contains attributes integrated with attribute-m and attribute-n mapped with map ((attribute-m, attribute-n), similarity (attribute-m, attribute-n));
Done;

B. Proposed Model (EMTC)

Input: Dynamic Document sets, number of data sources
Output: Preprocessed integrated dataset
Mapper1 (1...m, index)//Hadoop mappers indexing from 1 to m
Step1: for each document 'd'in the data source DS do
//for feature selection, where did is document id, tf is term frequency, itf is inverse term frequency and wt is weight measure
<did,index,tf,itf,wt>=featureselection(d);
Document representation should be done as <did, index,tf,itf,wt>
MD(1…n)=addDocument(<did,index,tf,itf,wt>);
done;
Step2: for each documented in MD(1…n)
do
Remove stop words
Remove special characters
List the medical features'' in the document using Medical Subject Headings (MeSH) terms.
Extract the synonyms' 'to the document features'' using GENETAG database.

Step3:find the feature clustering using the equation (1) as do
Conditional probabilityCP (tfi, θ j)= ∑ (prob(tfi ∩ θ j)/prob(tfi))*wtj;
Fclust(ρ i,fj)= ∑ j=1Cov(θ , ρ j)*CP(tfi, θ j);
//Covariance provides a measure of the strength of random variables between two or more sets
List<id,Fclust>
Mapper2<1...m,list<id,Fclust>>done

Step4: Estimate the decision parameters using co-decision matrix D from the instances set of all classifiers. For each data, source computes the patterns using the following models do
Apply Neural Network NN, Random Forest RF, Ensemble learning model, and Genetic algorithm based support vector machine GA_SVM and HBEC models.
Reduce<model, patternlist>done
Reducer<model, patterns>
Step5: For each model in the list do
Integrate and visualize the top k patterns in the pattern list;
done;
Given a set of medical document features
Feature, F={feature-1,feature-2…feature-n},the set contains n features.
For each document term t in d do
MI(t,d)=log(p(t/d)/p(t))=log(p(t,d)/p(t).p(d));
if(MI(t,d)>0)
wt=log(N/tf)*itf;
//tf-term frequency, itf-inverse term frequency and N-number of documents
Else continue;
done;

C. Distributed Databases (PubMed and Medline) used to Extract the Top Most Genes using Ranking Algorithm

Input: Database contains gene synonyms (GENETAG-DB), Gene Documents document sets from PubMed and Medline.
Output: Top ranked Distributed Gene-Synonym Documents
Procedure:
For each synonym used to describe genomes in GENETAG-DB do
For each Gene Documents document set do
GeneDocumentlist[i] =Gene Documents;
For each token t in GeneDocumentlist[i] do
getGeneSynonym[]=GeneSynonym(t);
PgetGeneSynonymDocuments[]=PubMed(Url(disease,getGeneSynonym[]);
MgetGeneSynonymDocuments[]=Medline(Url(disease,getGeneSynonym[]);
To check the mutual independence of two genes in the given contextual document use mutual information.
The common probability of the Document i, Pro(Di ∩ getGeneSynonym[j])
The sum of all probability of Document i, Pro(Di ∩ getGeneSynonym[j])
Probability of the total gene synonym documents Di Pro(Di)
Thres=0.5;// user-defined threshold for ranking
If(MFR>Thres)
Then
Add GESynonymDocumentList (MFR, PgetGeneSynonymDocuments, EgetGeneSynonymDocumentss);
End if
Done
Done;

IV. Results

After The Medline and GENETAG datasets are used for the experimental results. The data consist of a large collection of medical documents with different features. A total of one million medical documents were processed for training and testing in these experimental results. In this implementation part of the model, we have given the MEDLINE and PubMed archives to the proposed model. To implement the model, we utilize the present Apache Hadoop system with Amazon AWS server. The design of Amazon AWS server consists of cluster nodes ranging from 10-50, each with 10 CPU cores and 24 GB RAM is connected to the mapper node. A large number of MESH terms with gene and protein are utilized to characterize the patterns of a large number of medical documents from a great many records. These compared algorithms are done using Map/Reduce framework using Java Programming on Cluster Nodes.

From Table I, it is clear that the percentage of accuracy is achieved by the proposed model (~97%) are higher than the traditional ensemble model on different document set. As the size of the document increases the average runtime is decreased for the proposed model as compared to traditional models.

Document classification measures such as true positive rate, recall, precision are summarized in Table II. Also from the table, it is clearly observed that the proposed model has high computational efficiency (at least 5-10%) than the traditional models.

TABLE. I. Accuracy Terms of Document Processed Time

Algorithm	Accuracy		Documents Size in MB	Average Running Time (secs)
	40K	100K		
Multi-objective DT	72.75	76.35	5	235
Multi-text Bayesian Classifier	83.1	80.84	6	199
EMHMM	86.34	83.5	7	178
Neural Networks	90.28	88.33	8	204
Proposed Model (EMTC)	96.86	96.98	10	135

TABLE. II. True Positive, Precision and Recall of Proposed Method Compared with Traditional Methods

Algorithm	True Positive Accuracy	Precision	Recall
Multi-objectiveDT	73.65	71.65	75.25
Multi-text Bayesian Classifier	86.36	84.76	84.08
EMHMM	87.83	89.43	90.43
Neural Networks	89.64	90.53	91.87
Proposed Model (EMTC)	95.75	94.98	96.37

From Fig. 4, it is clear that the true positive rate scores achieved by the proposed model (~95.75%)are higher than the traditional ensemble model on different node configurations with different metrics.

From Table III, it is clear that the proposed model has high computational efficiency in terms of time and memory is concerned.

The proposed model (EMTC) completely eliminates the duplicate documents in the distributed databases (PubMed and Medline) using Top k-ranking algorithm.

From Fig. 5, it is clear that the proposed model performance in terms of time and memory is concerned. Proposed model completely eliminates the duplicate documents in the distributed databases (PubMed and Medline) using Top k-ranking algorithm. The proposed model is far better than traditional algorithms.

Table IV describes the feature extraction process of the gene to disease datasets using the proposed feature extraction models to the existing parameters. From the table, it is clearly observed that the proposed feature extraction model has high computational efficiency and less average runtime than the traditional gene-disease extraction measures.

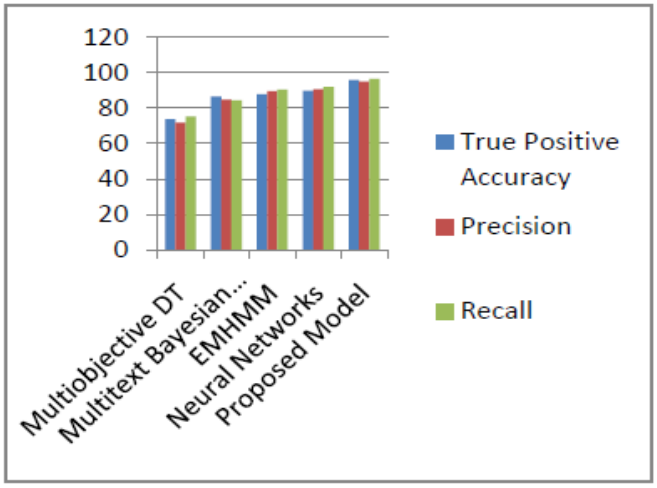


Fig. 4. Comparing the Proposed Model with Traditional Model.

TABLE. III. Performance Analysis of Top Ranking Gene-based Methods

Algorithm	Duplicate Filtered (Documents)	Time (ms)	Memory (Kb)
Multi-objective DT	6524	12763	25.76
Multi-text Bayesian Classifier	6935	11863	21.86
EMHMM	5987	12865	28.34
Neural Networks	5709	10273	23.86
Proposed Model (EMTC)	3876	8963	17.95

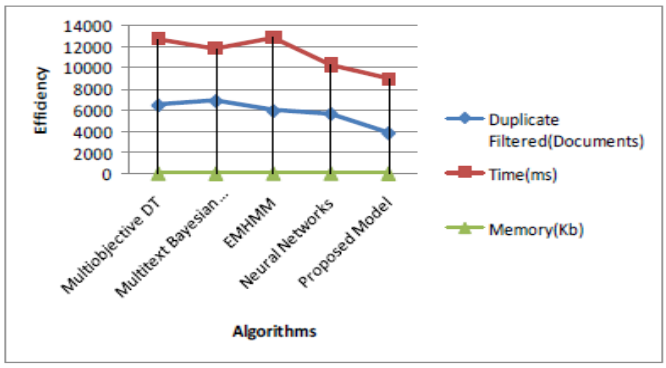


Fig. 5. Performance Analysis of Top Ranking Gene-based Documents.

TABLE. IV. GENE FEATURE EXTRACTIONS AND ITS ACCURACY

Parameters	Feature extraction Avg Accuracy	Avg Runtime(ms)
Information Gain	87.45	9743
Gain Ratio	89.29	8974
Chi-square	91.73	8689
Correlation	85.87	9217
Rough set	92.75	8248
Gene mutual information	97.53	7936
Gene Chi-square Measure	95.25	8013

In the proposed model (EMTC) is compared with different traditional algorithm models like Multi-objective DT, Multi-text Bayesian Classifier, EMHMM and Neural Networks, a large number of documents are processed with high true positive and precision. Experimental results prove that the proposed approach has a high recall, precision, and less standard error compared to conventional ensemble approaches. The proposed models used for biomedical classification using the Hadoop framework can be extended further for semantic entity mining and deep learning also. The proposed work can also be reached out to locate the perplexing qualities DNA-ailments designs utilizing parallel semantic mining model. Sem MedDB, a repository of semantic predictions used to find the relationship between the DNA sequence to disease, gene-disease and its clinical trials using the Hadoop framework. A novel Hadoop based semantic mining model on biomedical databases for multi-pattern evaluation using gene-disease patterns. An extension of parallel semantic mining model is used to find the relationship between two semantic entities (e.g. genedisease1, gene--disease2, etc.) and to check the existence of the disease patterns in more than three repositories. The proposed work can be extended to work on the disease to DNA or Protein sequences using distributed databases. The proposed work can be extended to deep learning models convolution neural networks for biomedical text classification. The research can be advanced to the disease-related patterns and its auto clinical decision system for medication. This work can be extended to optimize computing resources such as memory and time, when the data size is large (>30million records).

In future work, we will develop a multi-objective ensemble classification model to improve accuracy and outliers.

V. CONCLUSION

Here we propose a novel model, which uses a different classification and feature selection approaches makes it possible to select relevant genes with high confidence. A new feature selection and classification algorithm EMTC is implemented and evaluated by comparing with related feature selection algorithms. Our proposed model demonstrates its efficiency and effectiveness in dealing with high dimensional data for classification. As a result, text mining has evolved in the field of biomedical systems where text mining techniques and machine learning models are integrated using high computational resources. The main purpose of the work is to enhance the techniques used for feature selection and classification of high dimensional datasets of biomedical repositories and compared with different traditional algorithms. The results prove that the proposed model is efficient in selecting the relevant features and attains the outcomes as per users' interest with very low runtime. The accuracy of any model using relevant features is better than the complete features in the entire dataset.

ABBREVIATIONS

BOW: bag-of-words
MEDLINE: Medical Literature Analysis and Retrieval System Online
MeSH: Medical Subject Headings
XML: extensible markup language
BNER: Biomedical Named Entity Recognition
NLP: Natural Language Processing
POS: parts of speech
SVM: support vector machines
NMF: Nonnegative Matrix Factorization
sLDA: supervised Latent Dirichlet Allocation
UMLS: Unified Medical Language System
NER: Named Entity Recognition
DT-MEDA: decision tree-based multi-objective estimation of conveyance algorithm
NBC: naive Bayes classification
KNN: *K* nearest neighbour
NNS: neural network system
EM: expectation maximization
HMM: Hidden Markov Model
EMTC: Ensemble Model for Text Classification

ACKNOWLEDGMENT

We thank our college management, especially Dr. Lavu Rathaiah Garu, Chairman of Vignan's Group of Institutions and our beloved Principal Prof. Paturi Radhika Garu for supporting us, who put their faith in us. This work is mainly based on biomedical data extraction from the corpus, so we thank the referees in this wide area. We thank each and every one who supported us to accomplish this work successfully. This material has not been published in whole or in part elsewhere and not funded by any organization or committee. The author(s) received no specific funding for this work.

REFERENCES

[1] Kamath, U., De Jong, K., & Shehu, A. (2014). Effective automated feature construction and selection for classification of biological sequences. PloS one, 9(7), e99982. https://doi.org/10.1371/journal.pone.0099982.

[2] Bikku, T., Nandam, S. R., & Akepogu, A. R. (2018). A contemporary feature selection and classification framework for imbalanced biomedical datasets. Egyptian Informatics Journal, 19(3), 191-198. https://doi.org/10.1016/j.eij.2018.03.003.

[3] Wagholikar, K. B., MacLaughlin, K. L., Kastner, T. M., Casey, P. M., Henry, M., Greenes, R. A., & Chaudhry, R. (2013). Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. Journal of the American Medical Informatics Association, 20(4), 749-757. https://doi.org/10.1007/s10916-011-9780-4.

[4] Chen, C. M., Lee, H. M., & Chang, Y. J. (2009). Two novel feature selection approaches for web page classification. Expert systems with Applications, 36(1), 260-272. https://doi.org/10.1016/j.eswa.2007.09.008.

[5] Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 43(1), 15-23. https://doi.org/10.1016/j.jbi.2009.07.008.

[6] Bikku, T., & Paturi, R. (2019). A novel somatic cancer gene based biomedical document feature ranking and clustering model. Informatics in Medicine Unlocked, 100188. https://doi.org/10.1016/j.imu.2019.100188.

[7] Hu, C., Xu, Z., Liu, Y., Mei, L., Chen, L., & Luo, X. (2014). Semantic link network-based model for organizing multimedia big data. IEEE Transactions on Emerging Topics in Computing, 2(3), 376-387. https://doi.org/10.1109/TETC.2014.2316525.

[8] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., .& Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.

[9] Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5. https://doi.org/10.1145/2481244.2481246.

[10] Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. ACM Computing Surveys (CSUR), 38(1), 3. https://doi.org/10.1145/1132952.1132955.

[11] Bikku, T., Rao, N. S., & Akepogu, A. R. (2016). Hadoop based feature selection and decision making models on big data. Indian Journal of Science and Technology, 9(10). https://doi.org/10.17485/ijst/2016/v9i10/88905.

[12] Feldman, R., & Sanger, J. The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data. https://doi.org/10.1017/CBO9780511546914.

[13] Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. (2012, February). Modeling how students learn to program. In Proceedings of the 43rd ACM technical symposium on Computer Science Education (pp. 153-160). ACM. https://doi.org/10.1145/2157136.2157182.

[14] Bikku, T. (2017, August). A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets. In IOP Conference Series: Materials Science and Engineering (Vol. 225, No. 1, p. 012161). IOP Publishing. https://doi.org/10.1088/1757-899X/225/1/012161.

[15] Pulkkinen, P., & Koivisto, H. (2008). Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms. International Journal of Approximate Reasoning, 48(2), 526-543. https://doi.org/10.1016/j.ijar.2007.10.004.

[16] Qian, D., Zheng, D., Zhang, J., Shroff, N. B., & Joo, C. (2013). Distributed CSMA algorithms for link scheduling in multihop MIMO networks under SINR model. IEEE/ACM Transactions on Networking (TON), 21(3), 746-759.

[17] Zhang, L., Luo, Y., Tao, F., Li, B. H., Ren, L., Zhang, X., ... & Liu, Y. (2014). Cloud manufacturing: a new manufacturing paradigm. Enterprise Information Systems, 8(2), 167-187.

[18] Raja, K., Patrick, M., Gao, Y., Madu, D., Yang, Y., & Tsoi, L. C. (2017). A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries. International journal of genomics, 2017. https://doi.org/10.1155/2017/6213474.

[19] Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Biocomputing 2008 (pp. 652-663). https://doi.org/10.1142/9789812776136_0062.

[20] Leroy, G., McDonald, D. M., Su, H., Xu, J., Tseng, C. J., & Chen, H. (2005). Gene Pathway Text Mining and Visualization. https://doi.org/10.1007/0-387-25739-X_18.

[21] Zhong, X., & Li, W. (2007, December). A decision-tree-based multi-objective estimation of distribution algorithm. In 2007 International Conference on Computational Intelligence and Security (CIS 2007) (pp. 114-11). IEEE. https://doi.org/10.1109/CIS.2007.136.

[22] Shih, B. (2011). Target sequence clustering (Doctoral dissertation, Ph. D. Thesis, Machine Learning Department, Carnegie Mellon University).G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

Mutual Authentication Security Scheme in Fog Computing

Gohar Rahman[1], Chuah Chai Wen[2]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
86400, Johor Malaysia

***Abstract*—Fog paradigm is a new and emerging technology that extends the services of cloud computing near to edge network. This paradigm aims to provide rich resources near to edge devices and remove the deficiencies of cloud computing for example, latency. However, this paradigm is distributed in nature and does not guarantee the trustworthiness and good behavior of edge devices. Thus, authentication and key exchange are significant challenges in front of this new paradigm. The researchers have worked on different authentication and key exchange protocols. Recently Maged Hamada Ibrahim proposed an authentication scheme that permits fog user to authenticate mutually with fog server under the authority of a cloud service provider. Alongside, Amor et al proposed an anonymous mutual authentication scheme. In this scheme, the fog user and fog server authenticate each other without disclosing the user real identity, using public-key cryptosystem. But, we demonstrated that Maged Hamada Ibrahim does not preserve the user anonymity, hence, it was exposed to man in the middle attack. Amor et al. scheme is computationally complex as it is using public key cryptosystem that has low throughputs and requires large memory, which not suitable to employ for fog computing that connecting internet of things with small memory, and requires high throughputs. Therefore, to overcome the above-aforementioned security problems internet of things constraints, an improved mutual authentication security scheme based on advanced encryption standard and hashed message authentication code in fog computing has been proposed. Our scheme provides mutual authentication between internet of things devices and fog servers. We proved that the proposed improved scheme provides secure mutual authentication using the widely accepted Burrows Abdi Needham logic. In this study, the properties i.e. performance, security, and functionality are analyzed and compared with existing and related mutual authentication schemes. Our scheme provides better in security, functionalities, communication and computation cost as compared with the existing schemes.**

***Keywords*—Fog computing; mutual authentication; man in the middle attack; key exchange; ban logic**

I. INTRODUCTION

Fog computing is decentralized computing architecture where the data is processed between the internet of things (IoT) devices and cloud servers. This computing paradigm brings the services of cloud computing near to the edge devices. The edge devices i.e. switches, routers and gateways act as computing nodes along with the cloud data center [1]. Compared to cloud computing, the computations of fog computing provide better results (location awareness, geographical accessibility, low latency, mobility support, etc.). It is because the fog computing nodes are near to the IoT devices. A typical IoT cloud architecture works in three phases. IoT devices reside in the first phase where sensors collect the information and forward the collected information to the cloud servers. In the second phase, cloud servers analysed the received information. In the third phase, the cloud servers processed the information and send back to the IoT devices. In this case, cloud computing compromise on high latency, security, and privacy of data [2][3].

Like the cloud, fog computing gives processing and storage services to the IoT users. Fog computing depends on giving processed information and storage locally to the fog devices as opposed to sending them to the cloud [4]. The architecture of fog computing consists of three-layers as well. The first layer contains IoT devices such as sensors, actuators wearable devices, smartphone, smartwatches, etc. The second layer called the middle layer consist of fog nodes where the computation is performed in a real-time manner. The last layer includes the cloud server where the data is stored for future use [5]. Fog computing is believed to be an extension of cloud computing and the security problems in the cloud are inherited to fog computing. As fog computing is decentralized in nature, therefore, the same methods applying to cloud computing is not applicable in fog computing [6, 7]. In fog computing, when a user opens their resources, the attackers easily comes and attacked on the fog nodes [8]. One of the important malicious attack is man in the middle attack (MitM) [9, 10]. In this type of attack, the attacker is passed out through malicious inner user between two computers, secretly relays, and pretends himself to be the legitimate [11, 12].

MitM can be either categorized as eavesdropping and manipulation. Eavesdropping is a passive attack as the attacker is merely concerned in the information passed. While in manipulation attack, the attacker makes changes in data sent to it and pretend it as the original sender. Detection and preventions of MitM attacks are critical in fog computing [13, 14]. The architecture of fog is characteristically analogous to a MitM attack, as fog node is intermediate in the cloud and IoT devices, allowing the attacker to camouflage easily. Nodes of fog transform personal information intensely like the medical history of a patient, prescription and health status of a person. Whatever the case is, such kind of information might prove terrible in the erroneous hands [15]. Therefore, the security must need necessary while designing fog computing specially in the edge fog layers.

MiTM attack can be prevented with mutual authentication [16]. Where the literature reveals that none of the mutual authentication protocols cannot attain the complete security requirements in fog edge cloud environment. This work studies the protocol presented by Maged Hamada Ibrahim [17]. Maged Hamada Ibrahim scheme cannot resist MitM. Thus has motivated us to propose a modified mutual authentication protocol that achieves the weaknesses of Maged Hamada Ibrahim scheme. Our proposed mutual authentication scheme may resist known security attacks and offer better functionalities. For example (a) user anonymity, (b) mutual authentication and session key establishment, (c) less computation cost. The proposed scheme is using the hash message authentication code (Hmac) for integrity and authenticity between fog users and fog servers. For authentication between fog servers and registration authority, public key infrastructure (PKI) is used. Moreover, the advanced encryption standard (AES) is used for encryption and decryption. Lastly, key derivation function (KDF) is used for session key creation to establish secure communication between fog user and fog server.

The rest of the paper is divided into sections. Related work is discussed in Section 2. A brief overview of the Maged Hamada Ibrahim scheme is presented in Section 3. The proposed mutual authentication scheme is discussed in Section 4. Security services of the proposed scheme are discussed in Section 5. The mutual authentication security proof based on Ban logic is presented in Section 6. The security functionalities and computation comparison of the proposed mutual authentication are discussed in Section 7 and Section 8. In the end, Section 9 counts the conclusion.

II. Related Work

Amor et al. [18] proposed an anonymous mutual authentication scheme where the fog user and the fog server authenticate each other without disclosing the user real identity using public-key cryptosystem. According to Albakri et al. [19], the computation cost of a public-key cryptosystem is high. The internet of things has limited memory, processing, and battery power. Therefore, this scheme is impractical in fog computing.

A framework was proposed by Dsouza et al. [20] based on policy-based security in fog computing. The proposed framework was intended to support the secret exchange of information, collaboration and reuse of data in a fog computing. The framework uses attribute-based authentication where first the fog user will be detected and then accesses the resources or services based on the fog user attribute. The framework comprises of numerous modules. The fundamental goal of these modules is to define rules and store them for user services and to send information to various fog nodes. The authors stated that these modules are important for real-time computation, and one of the modules is referred to as a policy enforcement module that is mapped to either fog nodes, cloud data servers or IoT devices. However as stated by the authors, this is a preliminary framework which does not take into account all the nuances of the federal fog ecosystem.

Jimoh and Abdul Hamid [21] proposed a conceptual framework in fog computing, based on swarm intelligence optimization technique also known as dragonfly algorithm. The main aim of the proposed framework is to detect the man in the middle attack (MitM) in fog computing. The authors used ifog simulator to detect the presence of MitM in the fog-computing environment. The authors considered two scenarios to determine the threshold consumption of CPU in MitM. In the first scenario, the memory consumption is tested without the presence of MitM between fog and cloud. While the second scenario is tested when there is a MitM attack takes place. The author claimed that the memory consumptions is less in normal communication as compared to the scenario where the MitM attack take place. However, this study is only based on preliminary results and outcomes. There is no solution to prevent the MitM attack in fog cloud IoT environment.

Alwaris et al. [22] proposed an efficient key exchange protocol based on ciphertext policy attribute encryption scheme (CP-ABE) for securing groups of fog nodes and cloud. To achieve better results in term of confidentiality, authentication, verifiability and access control they integrate CP-ABE with the digital signature. But Nikkah et al. [23] claimed that the aforesaid algorithm might be deliberated for obtaining the security of fog nodes and cloud because they are usually resourced consuming and are not appropriate for deployment in the IoT layer of the fog computing architecture.

Lee et al. [24] explored various unique security and privacy problems such as man in the middle attack (MitM), malicious detection technique, intrusion detection, data protection, and data management problems. However, the authors ignored the proper solutions that tackle these problems.

Lu et al. [25] proposed a lightweight privacy- preserving data aggregation scheme (LPDA) in fog computing. This scheme use three-techniques, namely, one-way hash chain, homomorphic Paillier encryption, and Chinese Remainder Theorem. The main goal of the scheme is to prevent data injection attack in the network edge. Besides, the scheme provides better results in terms of communication overhead and computation cost as compared to the basic paillier encryption scheme. However, the limitation of this scheme is that the traceability is not considered.

Stomojo et al. [26] highlighted different features of fog computing in smart grid, smart homes, and traffic control system. They examined the stealthy feature of man in the middle attack (MitM) on CPU and other memory consumptions fog computing devices. The authors explained that MitM is easy to be detected but difficult to be addressed. This study not gives any proper solution to prevent the MitM.

III. Review on Maged Hamda Ibrahim Scheme [17]

This section briefly review “Octopus: An Edge-Fog Mutual Authentication Scheme” for fog computing environment proposed by Maged Hamada Ibrahim. The scheme has three main phases namely, system initialization phase, registration phase and authentication phase. We present

the brief overview of these phases to understand the security weaknesses of Maged Hamada Ibrahim scheme.

A. System Initialization Phase

In this phase of the scheme, the system is initialized. Registration authority RA contains their own public key and private key (〚PK〛_RA 〚,SK〛_RA). Similarly, fog server also contain private key and public key PK_{FS}, SK_{FS}. Registration authority knows the public key PK_{RA} of fog server. The registration Authority RA picks a unique identity ID_{FS} for fog server, sends the ID_{FS} signed with RA signature key SK_{RA}.

B. Registration Phase

In this phase of the scheme the fog user with identity, ID_{FU} approaching to registered with the *RA*. While every fog network having identity ID_F and containing a set of fog servers having identity ID_{FS}. The registration process (as shown in Fig. 1) is discussed below.

Step1: Fog user FU shows his identity ID_{FU} to the *RA*.

Step2: Upon receiving the request, *RA* pick random master secret key name K_{FU}, and send it to fog user FU.

Step 3: Fog user *FU* receive, the message from registration authority *RA* and store the master secret key K_{FU} and ID_{FU} on his smart device/card.

Step 4: RA calculates the secret key of FS and FU where K^{FU-FS} =$H\ (ID_{FU}, ID_{FS}, K_{FU})$.

Step 5: Upon receiving the message from *RA*. The *FS* verify the signature using *RA* public key PK_{RA}. After successful verification, decrypts the received tuple and stores ID_{FU} and K^{FU-FS}.

C. Mutual Authentication Phase

This phase executes several steps to achieve mutual authentication and session key establishment for all parties involved in the fog network. All the steps of this phase is discussing below which is shown in Fig. 2.

Step 1: FU picks a random number r_{FU} and broadcast the message $(Hello, ID_{FU}, r_{FU})$.

Step2: When fog server *FS* is within the range of the fog user FU, the fog server *FS* checks the identity of fog user *FU* which is ID_{FU} . If the identity ID_{FU} of fog user FU is registered, the communication is continuing otherwise abort the communication.

Step3: FS fetches the key K^{FU-FS} for fog user *FU*.

Step 4: *FS* picks a random number r_{FS}.

Step 5: The FS encrypts the r_{FS} and r_{FU} using symmetric key K^{FU-FS} , replies with message $(ID_{FU}, ID_{FS}, ID_F\ E(K^{FU-FS}, r_{FS}, r_{FS})$ to fog user FU.

Step 6: Upon receiving the message from FS , FU calculates the symmetric key K^{FU-FS} locally where $K^{FU-FS} = H\ (ID_{FU}, ID_{FS}, K_{FU})$. Secondly, *FU* decrypts the received tuples and checks for the validity of random number r_{FU}. If the check is failed the session is aborted. Otherwise, fog user picks the session key K_S. Encrypt the session keyK_S by using symmetric key K^{FU-FS} send the tuples ID_{FU}, ID_{FS}, K_{FU} , $E\ (K^{FU-FS}, K_S, r_{FS})$ to the *FU*.

Step 7: Upon receiving the message from *FU*, the *FS* decrypt the received tuples and check for validity of r_{FS} if the matched of r_{FS} successful the session key K_s will be accept otherwise session is rejected.

D. Weakness of Maged Hamda Ibrahim Scheme

The main drawback of the Maged Hamada Ibrahim is the identity of each *FU* and *FS* is publically transmitted on unsecure channel. Therefore, the attacker may perform a middle attack during communication between fog user and registration authority. When fog user sends their identity to the registration authority, the attacker compromised the identity of fog user and obtain the master secret key from the registration authority. While in the authentication phase, the identity of fog user is also transmitted on unsecure channel. The attacker in the middle get the identity of fog user and fog server and reached to obtain the session key among fog users and fog servers. Therefore, this scheme is unsecured due to man in the middle attack (MitM).

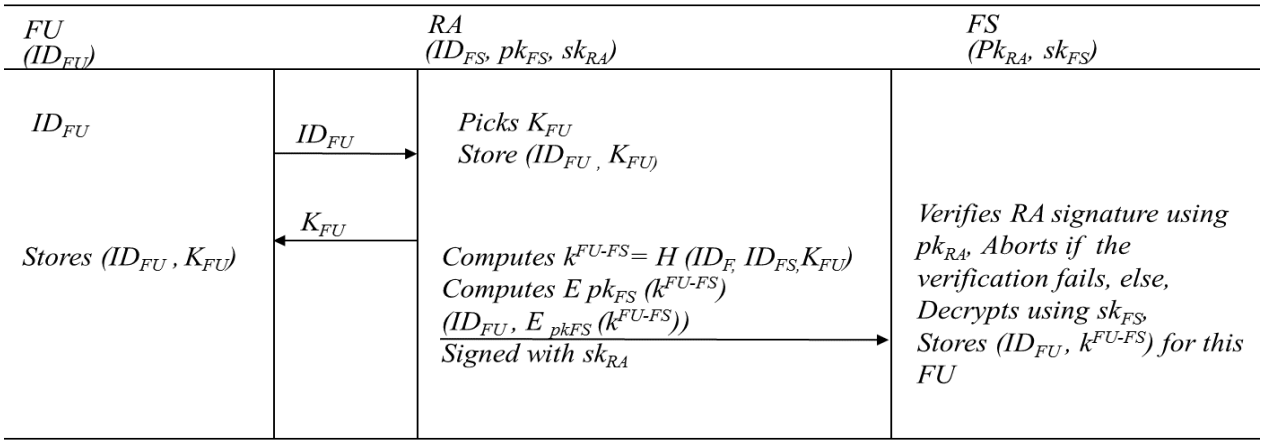


Fig. 1. Registration Phase of the Maged Hamda Ibrahim Scheme [17].

RMC Universiti Tun Hussein Onn Malaysia (UTHM)

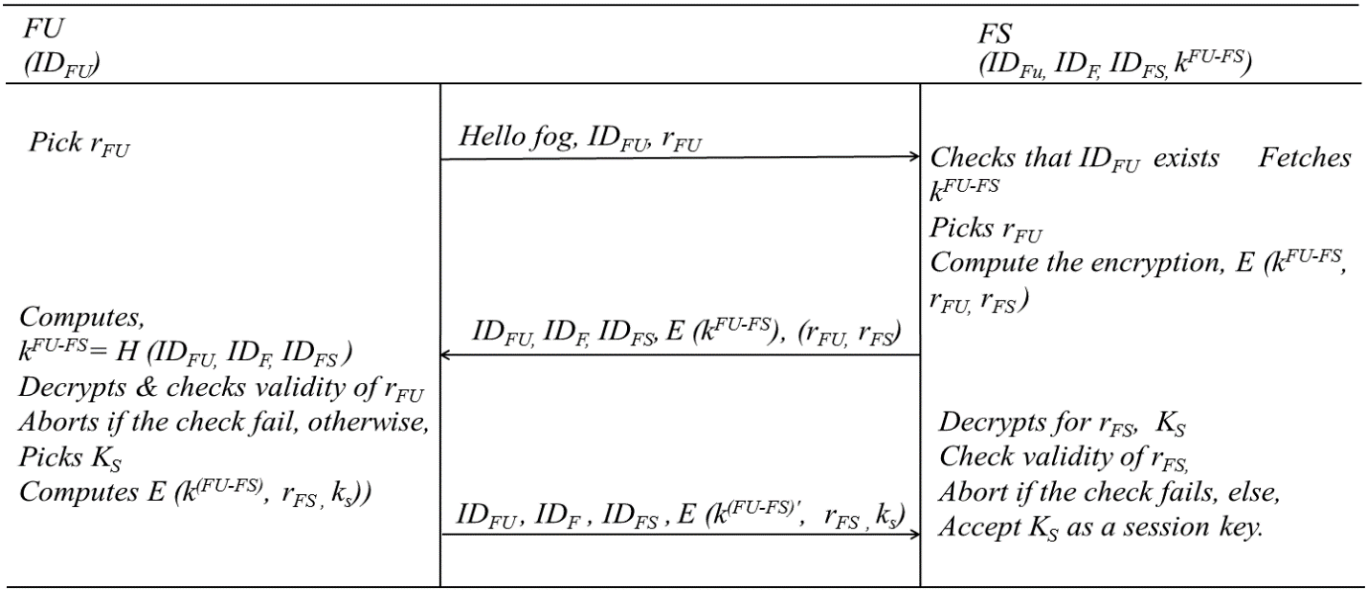


Fig. 2. Mutual Authentication Phase of the Maged Hamda Ibrahim Scheme [17].

IV. Proposed Scheme

This section briefly discussed the proposed mutual authentication scheme for fog computing to overcome the security weaknesses of Maged Hamada Ibrahim that is discussed in Section 3. The proposed scheme provides a lightweight strong mutual authentication among the fog users and fog servers. Our proposed scheme also divided into three phases as in the Maged Hamada Ibrahim, namely, 1) System initialization phase, 2) fog user registration phase, 3) Mutual authentication and session key exchange phase. Table I referred to the notations used the proposed scheme. Each phase of the proposed scheme is discussing below in details.

A. System Intialization

Before fog user registration, the registration authority RA has to perform system initialization phase. The *RA* has public key PK_{RA} and private keys SK_{RA} and K_{FU-RA}. The *RA* public key PK_{RA} is known to the *FS*. Fog server *FS* has also public key PK_{FS} and private key SK_{FS}. We assumed that, the *RA* private key K_{FU-RA} is shared with *FU* in advance.

TABLE. I. Notations used in the Proposed Scheme

Symbols	Description
K_{FU-FS}	Master secret key
K_{FU-RA}	Secret key shared between *RA*, and fog user
PK_{FS},SK_{FS}	Public and private key of the fog server
ID_F	Identity of fog network
ID_{FS}	Identity of fog server
ID_{FU}	Identity of fog user
n_1,n_2, n_3	Random numbers
Ks	Session key
Ts	Timestamp
$Hmac$	Hash message authentication code
kdf	Key derivation function
$\|\|$	Concatenation operator

B. Registration Phase

Registration phase is *FU* and *FS* registering themselves with *RA* as shown in the Fig. 3. The authentication process is carried out between *FU* and *FS* with the registration authority *RA* as follows:

Step1: $FU \rightarrow RA$: $E(K_{FU-RA}, ID_{FU}||n_1) = c$,
$Hmac(K_{FU-RA}, c) = T_1$.

Each fog user *FU* generate a random number n_1. Compute c and T_1 . Bothe c and T_1 sends to the *RA*.

Upon receiving message, c , T_1 . The *RA* computes the valid tag T_1'. Compared with the received tag T_1. If it is valid tag *RA* decrypts the message $c, D(K_{FU-RA}, c)$ and stored the identity of fog user ID_{FU}.

Step 2: $RA \rightarrow FU$: $E(K_{FU-RA}, K_{FU-FS}||n_1) = c_1$,

$Hmac(K_{FU-RA}, c_1) = T_2$

RA picks a long-term master secret key K_{FU-FS} and encrypts it with random number n_1 by using shared symmetric key K_{FU-RA}. *RA* Sends the parameter c_1 with the $Hmac$ tag T_2 where tag $T_2 = Hmac(K_{FU-RA}, c_1)$ to *FU*.

Upon receiving the message c_1 and T_2. *FU* verifies the T_2' and compares it with the received tag T_2. If the verification successful, the *FU* decrypts c_1 using K_{FU-RA} and stored the long-term secret key K_{FU-FS} and ID_{FU}.

Step: 3: $RA \rightarrow FS$: $E(PK_{FS}, ID_{FU}, ID_{FS}, ID_F, K_{FU-FS}) = c_2$,

$Sign(SK_{RA}, c_2) = sig_{C2}$

RA picks the ID_{FU}, ID_{FS} and long term secret key K_{FU-FS}. Encrypts it by using the public key of *FS*. *RA* signs the encrypted message by using his own secret key SK_{RA} and sends to the fog server *FS* . Upon receiving message c_2 and sig_{C2} . *FU* verifies the signature sig_{c2} by using public key

PK_{RA} of the *RA*. If the signature verified the *FS* decrypts the received message and stored the ID_F, ID_{FS} and long term secret key K_{FU-FS}.

C. Mutual Authentication

When a registered fog user *FU* want to access the services of fog server *FS*, *FU* needs to exchange a mutual authentication key request message to the *FS*. Fig. 4 shows the mutual authentication process and brief detail of the steps as follows.

Step1: $FU \rightarrow FS: Hmac(K_{FU-FS}, ID_{FU}||n_2 = T_3.$

FU generates the random number n_2 and sends the tag T_3 to *FS* where T_3=$(K_{FU-FS}, ID_{FU}||n_2$. Upon receiving the tag T_3. If tag T_3 is valid *FS* picks a random number n_3, identity ID_{FS}, and the identity of fog network ID_F.

Step2: $FS \rightarrow FU: Hmac(K_{FU-FS}, ID_{FU}||\ ID_F\ ||\ n_3) = T_4$

FS Computes tag T_4 where tag T_4= $Hmac(K_{FU-FS}, ID_{FU}||\ ID_F\ ||\ n_3))$ and sends tag T_4 to *FU*. Upon receiving the received tag T_4 from *FS*.The *FU* calculates and if the tag T_4 is valid *FU* generates current timestamp T_S and calculate the session key by using *KDF*. $kdf\ (\ ID_{FS}\ ||\ ID_F||\ K_{FU-FS}||T_S) = K_S.$

Step 3: $FU \rightarrow FS: Hmac(\ K_{FU-FS}\ , T_S) =\ T_5\ .$

FU sends the tag T_5 to *FS*, where the tag T_5= $(\ K_{FU-FS}\ ||T_S)$. *FS* calculates T_5' and verifies tag T_5. If the tag is valid *FS* calculates the session key K_S' such that $kdf\ (\ ID_{FS}\ ||\ ID_F||\ K_{FU-FS}||T_S) = KS'$ and store the session key for commutation between fog users and fog servers.

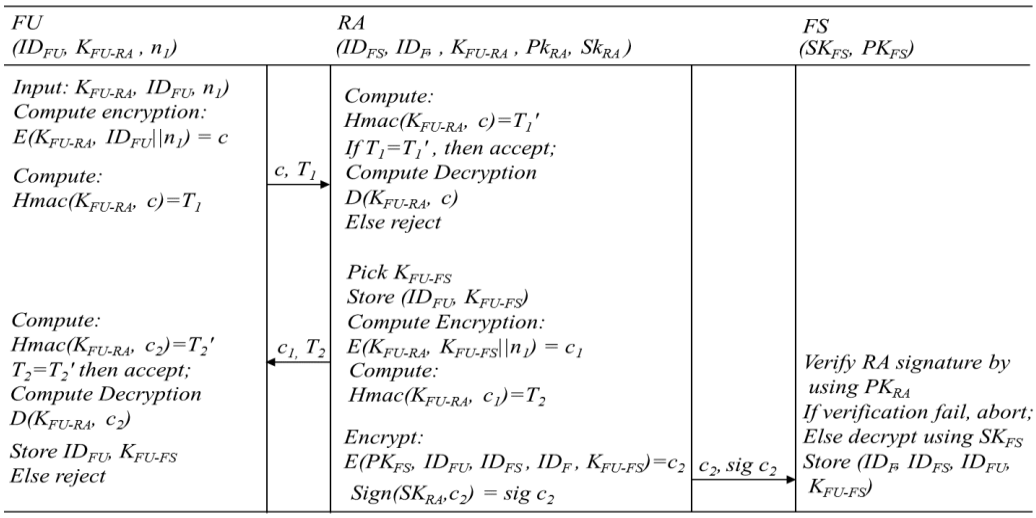


Fig. 3. Registration Phase of the Proposed Scheme.

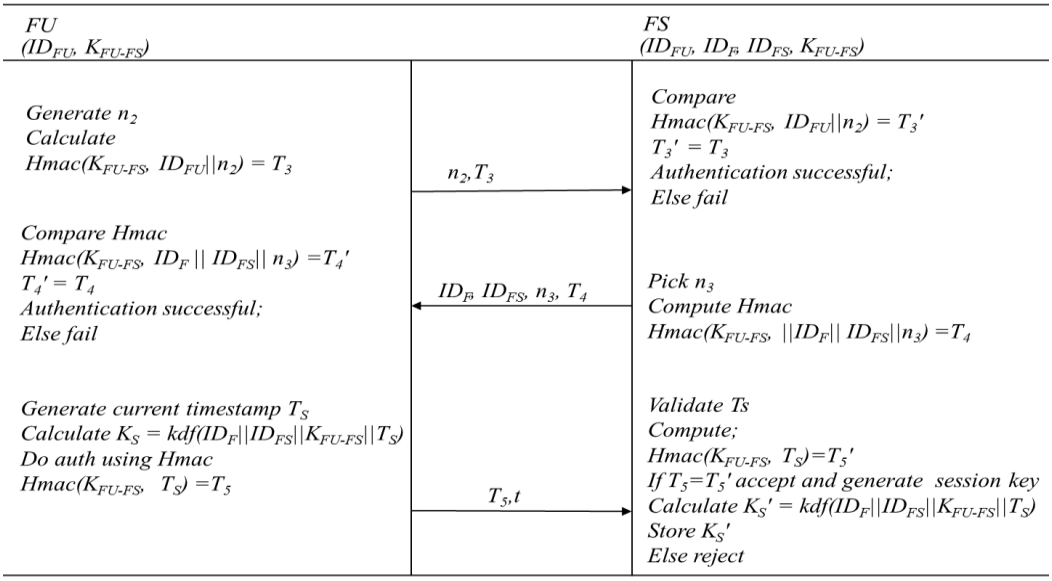


Fig. 4. Mutual Authentication Phase of the Proposed Scheme.

V. Security Services of the Proposed Scheme

A. Secure Mutual Authentication and Key Establishment

In our proposed scheme, the mutual authentication among fog users and fog servers is ensured based on hash message authentication code *Hmac*. During the registration phase, the fog user using random number encrypted with the identity of fog user. Similarly, *Hmac* tag is sent to the registration authority with cipher text, the attacker cannot deduce the valid tag. Thus the proposed scheme ensured the mutual authentication among fog servers and fog users. Also for mutual authentication phase the adversary may not be able to compute the valid tag because the adversary has no knowledge about the *Hmac* key. Thus in our proposed scheme, fog user securely start communication with the fog server using session key.

B. Resistance to Man in the Middle Attack (MitM)

In MitM attack, the active attacker attacks in the middle when two entities (*FU* and *FS*) exchange secret information. The attacker secretly forwards and receives secret parameters and messages. Suppose an attacker intercept the communication between the *FU* and the *RA*. The *FU* and *RA* using the message c, T_1. The aim of the attacker is to modify this message and make another valid message for example c' , T_1' the attacker is not able to compute the valid tag $T_1{}'$ because the attacker does not know the Hmac secret key K_{FU-RA}. Similarly, the attacker also cannot modify other messages such like message 2, c_1, T_2 , message 3 , c_2,sig_{C2} message 4, T_3, message 5, T_4 and message 6 T_6. It is then clear that our improved mutual authentication scheme resist (MitM).

C. Anonymity

In our proposed scheme, the identity of *FU* and *FS* is transmitted in secure channel. During the registration phase of the *FU*, the identity is encrypted by using symmetric key K_{FU-RA} . While mutual authentication phase, the identity of *FU* and *FS* are associated with message authentication code (*Hmac)*. The attacker cannot recognize the transmitted identity during the registration phase because the attacker does not know the secret key used by the *FU* and *RA*. While during mutual authentication phase the attacker cannot extract the valid identity from the associated *Hmac* value. Therefore, our proposed scheme ensures the anonymity of the *FU*.

VI. Security Proof based on Ban Logic

To prove that our modified mutual authentication scheme fulfil the requirement of mutual authentication, we are using the Ban logic postulates (BAN) [27]. The notation of Ban logic is given in Table II.

A. Ban Logic Postulate

The logical postulates (rules) of the BAN logic are giving below.

1) Message meaning rules

$$\frac{A| \equiv A \overset{K}{\leftrightarrow} B \lhd \{C\}K}{A| \equiv B| \sim C}$$

2) The nonce verification rules

$$\frac{A| \equiv \#(C), A| \equiv B| \sim C}{A| \equiv B| \equiv C}$$

3) The belief rules

$$\frac{A| \equiv B| \equiv C, A| \equiv (C, D)}{A| \equiv C}$$

4) The freshness rules

$$\frac{A| \equiv \#(C)}{A| \equiv \#(C, D)}$$

5) The jurisdiction rules

$$\frac{A| \equiv B \Rightarrow C, A| \equiv B| \equiv C}{A| \equiv C}$$

To conduct the security of the modified mutual authentication scheme, the following process should be performed. First, we make the idealized messages of our improved mutual authentication scheme in the formal logic.

Second, we present the initial assumption based on the initial status of the improved mutual authentication scheme. Third, we set goals based on the improved mutual authentication scheme. Fourth, we prove the mentioned goals by using Ban logic postulate.

B. Idealized Messages of the Proposed Scheme

$$Message_1: FU \rightarrow RA: \{n_1, FU \overset{ID_{FU}}{\longleftrightarrow} RA\}K^{FU\text{-}RA}, (FU \overset{HMAC(K_{FU-RA,C})}{\longleftrightarrow} RA)$$

$$Message_2: RA \rightarrow FU: \left\{n_1, RA \overset{K_{FU\text{-}FS}}{\longleftrightarrow} FU\right\} K_{FU\text{-}RA}, (RA \overset{HMAC(K_{FU\text{-}RA,C_1}}{\longleftrightarrow} FU)$$

$$Message_3: RA \rightarrow FS: (\{ID_{FU}, ID_F, ID_{FS}, K_{FU\text{-}FS}\}PK_{FS})SignSk_{RA}$$

$$Message_5: FS \rightarrow FU, (FS \overset{HMAC(K_{FU\text{-}FS,ID_F,ID_{FS},n_3})}{\longleftrightarrow} FU)$$

$$Message_6: FU \rightarrow FS, (FU \overset{HMAC(K_{FU\text{-}FS,T_S})}{\longleftrightarrow} FS)$$

TABLE. II. Banlogic Notations

Notation	Descriptions
A, B	Principals
C, D	Formulas
$A\| \equiv C$	A believes C
$\#C$	C is fresh
$A\| \sim C$	A one said C
$A \Rightarrow C$	A control over C
$A \lhd C$	A sees C
$A \overset{K}{\leftrightarrow} C$	A and B use the shared key K to communicate
$\{C\}K$	C is encrypted under K key
K_S	Represent Session key

C. Assumptions

Second the following assumption are made about the initial state of the scheme to analyse the proposed protocol where we have five assumption (*A1-A5*) which related with the public and private keys shared between the registration authority *RA*, fog user *FU* and fog server *FS*. The assumptions *A6-A9* shows that fog server *FS* and fog user *FU* believe that timestamp and random generated numbers elsewhere are fresh. This indicates that the protocol trusts heavily on the use of in time clocks. The final assumption *A10* indicates that both party have control the session keys shared between fog user and fog server.

$A1: FU \mid\equiv \left(FU \overset{K_{FU-RA}}{\longleftrightarrow} RA\right)$

$A2: RA \mid\equiv \overset{PK_{FS}}{\longrightarrow} FS$

$A3: FS \mid\equiv \overset{PK_{RA}}{\longrightarrow} RA$

$A4: FU \mid\equiv \left(FU \overset{K_{FU-FS}}{\longleftrightarrow} FS\right)$

$A5: FS \mid\equiv \left(FU \overset{K_{FU-FS}}{\longleftrightarrow} FS\right)$

$A6: FU \mid\equiv \#(n_1))$

$A7: FU \mid\equiv \#(n_2)$

$A8: FS \mid\equiv \#(n_3)$

$A9: FU \mid\equiv \#(T_S)$

$A10: FS \mid\equiv FU \Rightarrow \left(FU \overset{KS}{\leftrightarrow} FS\right), \#\left(FU \overset{KS}{\leftrightarrow} FS\right)$

D. Goals

Third, we set four goals for our proposed scheme as shown below. The goals of our improved mutual authentication scheme contributes to exchange shared secret key K_{FU-FS} , session key K_S and identity of fog user ID_{FU} between two entities. Both entities have to believe that the other entity also believe on secret key K_{FU-FS} , session key K_S and identity of fog user ID_{FU}. The goal *G1* and *G3* is related to the identity of fog user *FU*. *G2* is related to the shared secret key between fog user *FU* and fog server *FS*. *G4* is related to session key shared between two parties (*FU*, *FS*).

$G1: FU \mid\equiv RA \mid\equiv \left(FU \overset{ID_{FU}}{\longleftrightarrow} RA\right)$

$G2: FU \mid\equiv FS \mid\equiv \left(FU \overset{K_{FU\text{-}FS}}{\longleftrightarrow} FS\right)$

$G3: FU \mid\equiv FS \mid\equiv \left(FU \overset{ID_{FU}}{\longleftrightarrow} FS\right)$

$G4: FS \mid\equiv FU \mid\equiv \left(FU \overset{K_S}{\leftrightarrow} FS\right), \#\left(FU \overset{KS}{\leftrightarrow} FS\right)$

Fourth, the idealized form of the proposed protocol is analysed based on the BAN logic postulate. The main proof is stated as follows.

Step1: According to message 1, assumption A1 and message meaning rule we obtain L1.

$$\text{L1:}\frac{RA \lhd \{ID_{FU}, n_1\}K_{FURA}, (FU \overset{HMAC(K_{FU-RA,\,c})}{\longleftrightarrow} RA}{RA \mid\equiv FU \mid\sim (ID_{FU}, n_1), (FU \overset{HMAC(K_{FU\text{-}RA,\,C})}{\longleftrightarrow} RA}$$

Step2: According to message 2, assumption A1, and message meaning rules we obtain L2.

$$\text{L2:}\frac{FU \lhd, \{n1,\ K_{FUFS}\}K_{FURA}\ RA \overset{HMAC(K_{FU\text{-}RA,\,c_1})}{\longleftrightarrow} FU\Big)}{FU \mid\equiv RA \mid\sim \{n1,\ K_{FUFS}\}K_{FURA}, RA \overset{HAMAC(K_{FU\text{-}RA,c_1})}{\longleftrightarrow} FU}$$

Step3: According to message 3, assumption A2, A3, and public key-shared rules we obtain L3.

$$FS \mid\equiv \left(\overset{PK_{FS}}{\longrightarrow} RA\right),$$

$$\text{L3:}\frac{FS \lhd (\{ID_F, ID_{FS}, ID_{FU}, K_{FU\text{-}FS}\}PK_{FS})SignSk_{RA}}{FS \mid\equiv RA \mid\sim (ID_F, ID_{FS}, ID_{FU}, K_{FU\text{-}FS})SignSk_{RA}}$$

Step4: From derivation of L2, L3, Assumptions A4, A5 and belief rules, we obtain L4.

$$\text{L4:}\frac{FU \equiv FS \mid\equiv \left(FU \overset{K_{FU\text{-}FS}}{\longleftrightarrow} FS\right), ID_{FU,} ID_{FS,ID_{F,}}}{FU \equiv FS \mid\equiv \left(FU \overset{K_{FU\text{-}FS}}{\longleftrightarrow} FS\right)}$$

This satisfied *G2*

Step5: According to assumption, A6 and freshness rule we obtain L5.

$$\text{L5:}\frac{FU \mid\equiv \#(n_1)}{FU \mid\equiv \#(ID_{FU}, n_1)}$$

Step6: From derivation of L1, L4 and belief rule we obtain L6.

$$\text{L6:}\frac{FU \mid\equiv RA \mid\equiv (ID_{FU}, n_1)}{FU \mid\equiv RA \mid\equiv (ID_{FU})}$$

This satisfied *G1*

Step7: According to message 4, assumption A4 and message meaning rule we obtain L7.

$$\text{L7:}\frac{FS \mid\equiv \left(FU \overset{K_{FU\text{-}FS}}{\longleftrightarrow} FS\right), FS \lhd (FU \overset{HMAC(K_{FU\text{-}FS,\,ID_{FU}||n2})}{\longleftrightarrow} FS)}{FS \mid\equiv FU \mid\sim HMAC(K_{FU\text{-}FS,\,ID_{FU}||n2})}$$

Step8: According to message5 and message meaning rule we obtain L8.

$$\text{L8:}\frac{FU \lhd (FU \overset{HMAC(K_{FU\text{-}FS,ID_F,ID_{FS},n_3})}{\longleftrightarrow} FS)}{FU \mid\equiv FS \mid\sim (FU \overset{HMA(K_{FU\text{-}FS,\,ID_F,ID_{FS},n_3})}{\longleftrightarrow} FS)}$$

Step9: From derivation of L7, L8 and nonce verification rules, we obtain L9.

$$\text{L9:}\frac{FU \mid\equiv \#(n_2, n_3),\ FS \mid\sim (ID_F, ID_{FS}, n_3)}{FU \mid\equiv\ FS \mid\sim (ID_F, ID_{FS}, n_3)}$$

Step10: From derivation of L9, assumption A7, A8 and freshness rules we obtain L10, L11.

L10: $\dfrac{FU \mid\equiv \#(n_2)}{FU \mid\equiv \#(n_2, n_3)}$

L11: $\dfrac{FS \mid\equiv \#(n_3)}{FS \mid\equiv \#(n_{2,} n_3)}$

Step11: From derivation of L7 and belief rule, we obtain L12.

L12: $\dfrac{FU \mid\equiv FS \mid\equiv (ID_{FU}, n_2)}{FU \mid\equiv FS \mid\equiv (ID_{FU})}$

This satisfied G3.

Step12: According to message 6 and message meaning rule we obtain L13.

L13: $\dfrac{FS \lhd (FU \overset{HMAC(K_{FU\text{-}FS}, T_S)}{\longleftrightarrow} FS)}{FS \mid\equiv FU \mid\sim (FU \overset{HMAC(K_{FU\text{-}FS}, T_S)}{\longleftrightarrow} FS)}$

Step13: According Assumption 9 and freshness rule we obtain L14.

L14: $\dfrac{FU \mid\equiv \#(T_s)}{FU \mid\equiv \#(K_s, T_s)}$

Step14: According to assumption A10 and jurisdiction rules we obtain L15.

L15: $\dfrac{FS \mid\equiv FU \Rightarrow (FU \overset{KS}{\leftrightarrow} FS), FS \mid\equiv FU \mid\equiv (FU \overset{KS}{\leftrightarrow} FS)}{FS \mid\equiv (FU \overset{KS}{\leftrightarrow} FS)}$

This satisfied G4.

Therefore goals *G1-G4* prove that our improved scheme achieve the mutual authentication between fog users and fog servers.

VII. Security Performance of the Proposed Scheme

Table III shows the security services of the existing schemes and our proposed scheme. The Maged Hamada Ibrahim scheme fail to provide the security services such as mutual authentication, user anonymity and cannot resist to the MitM. While the Amor model fulfils the security services, except for the lightweightedness. This model is computationally complex as compared to our scheme. It uses public-key encryption and decryption on the fog user, which is not lightweight for IoT devices In contrast; our proposed scheme provides the security services such as confidentiality, mutual authentication, message integrity, and user anonymity. We proved that our scheme can strongly resist to man in the middle attack using BAN logic as shown in section 6. Our proposed scheme and the Maged Hamada Ibrahim scheme are lightweight as they are using symmetric encryption, which requires less power for its operations and less power for its functioning.

TABLE. III. Security Performance of the Proposed Scheme

Scheme / **Services**	**Maged scheme**	**Amor Scheme**	**Our Scheme**	
Mutual Authentication	no	yes	yes	
User anonymity	no	yes	yes	
man in the middle attack	no	yes	yes	
Message integrity	no	yes	yes	
Lightweight	yes	No	yes	

VIII. Computation Comparison of the Proposed Scheme with Closely Related Schemes

The computation comparison of the proposed scheme with other closely related schemes is shown in Table IV. After thorough analysis, our proposed scheme during the registration phase, the fog user only needs one encryption and decryption operations and two *Hmac* operations, while the registration authority performs one symmetric encryption and decryption operation and one asymmetric encryption. During the mutual authentication phase, the fog user needs two *Hmac* operations while on the fog server-side only two *Hmac* operation is performed. Amor model performed asymmetric encryption and decryption operations between the communication of fog user side and registration authority Thirdly the Maged Hamada Ibrahim scheme performed an asymmetric operation and one hash invocation during the registration phase, while in the authentication phase the model performed one symmetric operation and one hash invocation on the fog user. While in the fog server-side encryption and decryption process is executed. The conclusion, our scheme provides better security and less computationally as compared with the others closely related schemes.

TABLE. IV. Computation Comparison of the Proposed Scheme with Closely Related Schemes

Scheme	**Maged scheme[17]**			**Amor scheme[18]**			**Our scheme**		
Phase	***Initialization***	***Registration***	***Authentication***	***Initialization***	***Registration***	***Authentication***	***Initialization***	***Registration***	***Authentication***
FU	-		*1 sym enc* *1 sym dec* *1 hash*	-	*1 asy enc* *1 asy dec*	*asym enc*	-	*1 sym enc* *1 sym dec* *2Hmac*	*2Hmac* *1 KDF*
FS	*1 sign verif*	*1 sign verif* *1 asym dec*	*1 sym enc* *1 sym dec*	-	*1 sign verif* *1 asym dec*	*asym enc*	-	*1 sign verify* *1 asym dec*	*2Hmac* *1 KDF*
RA	*1 sign gen* *1 hash*	*1 asym enc* *1 hash* *1 sign verif*	-	-	*1 sign* *1 asy dec* *2 asym enc*		-	*1 sym enc* *1 sym dec* *2 Hmac*	

IX. Conclusion

In this paper, we have demonstrated that Maged Hamada Ibrahim scheme is vulnerable to (MitM) and it cannot provide mutual authentication and anonymity of the fog user. To overcome the weaknesses of Maged Hamada Ibrahim scheme we have proposed an improved mutual authentication security scheme for fog computing. Our modified improved scheme have resistance to various types of attacks. We proved that our modified improved scheme provides mutual authentication between fog user and fog server using BAN logic. We compared the modified mutual authentication scheme with the existing schemes and it shown that our modified mutual authentication scheme performs well in term of security and functionality requirements.

Acknowledgment

The authors would like to thank Information Security Interest Group (ISIG), Faculty of Computer Science and Information Technology (FSKTM) University Tun Hussein Onn Malaysia (UTHM) for recurrent assistance in executing this research study and, RMC UTHM, for financially supporting this Research under research Fund E15501.

References

[1] S. B. Nath, H. Gupta, S. Chakraborty, and S. K. Ghosh, "A survey of fog computing and communication current researches and future directions," arXiv preprint arXiv, pp. 1-47, 2018.

[2] J. Ni, k. Zhang, X. Lin, and X. S. Shen, "Securing fog computing for internet of things applications: challenges and solutions," IEEE Communications Surveys & Tutorials, vol. 20, no. 1, pp. 601-628, 2017.

[3] N. Abubaker, L. Dervishi, And E. Ayday, "Privacy-preserving fog computing paradigm," IEEE Conf. on Communications and Network Security, pp. 502-509, October 2017.

[4] H. Atlam, R. Walters, and G. Wills, "Fog computing and the internet of things: a review," big data and cognitive computing, vol. 2, no. 2, pp.10, 2018.

[5] M. Verma, N. Bhardwaj, and A. K. Yadav, "Real time efficient scheduling algorithm for load balancing in fog computing environment'. Int. J. Inf. Technol. Comput. Sci, vol. 8, no. 4, pp.1-10, 2016.

[6] B. Z. Abbasi, and M. A Shah, "Fog computing: Security issues, solutions and robust practices," Int. Conf. on Automation and Computing (ICAC), 2017, pp. 1-6, September 2016.

[7] P. Kumar, N. Zaidi, and T. Choudhury, "Fog computing: Common security issues and proposed countermeasures," 2016 Inte. Conf. System Modeling & advancement in research trends, pp. 311-315, November 2016

[8] Y. Sun, F. Lin, and N. Zhang, N "A security mechanism based on evolutionary game in fog computing," Saudi journal of biological sciences, 2018, vol. 25, no. 2, pp. 237-241, 2018.

[9] C. Li, Z. Qin, E. Novak, and Q. Li, "Securing SDN infrastructure of IoT–fog networks from MitM attacks," IEEE Internet of Things Journal, vol. 4, no.5, pp. 1156-1164, 2017.

[10] J. Ni, K. Zhang, X. Lin, and X. S. Shen, "Securing fog computing for internet of things applications: Challenges and solutions," IEEE Communications Surveys & Tutorials, vol. 20, no. 1, pp. 601-628, 2017.

[11] G.Rahman, and C.C. Wen, "Fog computing, applications security and challenges, Review," International Journal of Engineering & Technology, vol.7, no. 3, pp.1615-1621, 2018.

[12] .Y. Desmedt, "Man-in-the-middle attack. Encyclopedia of cryptography and security," pp. 759-759, 2011.

[13] G.Rahman, and C. C. Wen, "Man in the Middle Attack Prevention for edg-fog, mutual authentication scheme," International Journal of Recent Technology and Engineering (IJRTE), vol.8, no 2s2, July 2019.

[14] B. N. Ekanayake, M. N. Halgamuge, and A. Syed, "Security and privacy issues of fog Computing for the internet of things (IoT)," In Cognitive Computing for Big Data Systems Over IoT, Springer, Cham, , pp. 139-174, 2018.

[15] S. Khan, S. Parkinson, and y. Qin, "Fog computing security: a review of current applications and security solutions', Journal of Cloud Computing, vol. 6 no. 1, pp.19, 2017.

[16] Z. Chen, S. Guo, R. Duan, and S. Wang, "Security analysis on mutual authentication against man-in-the-middle attack," First Inte. Conf. on Information Science and Engineering, pp. 1855-1858, 2009.

[17] H. M. Ibrahim, "Octopus: An edge-fog mutual authentication scheme, IJ Network Security, vol. 18, no. 6 pp.1089-1101,2016.

[18] A.B. Amor, M. Abid, and A. Meddeb, "A Privacy- Preserving Authentication Scheme in an Edge-Fog Environment," In IEEE/ACS 14th Int. Conf. on Computer Systems and Applications, , pp. 1225-1231, October 2017.

[19] A. Albakri, M. Maddumala, and L. Harn, "Hierarchical polynomial-based key management scheme in Fog Computing,". 17th IEEE Inte.Conf. On Trust, Security and privacy in computing and communications, pp. 1593-1597, august 2018.

[20] C. Dsouza, G. J. Ahn, and M. Taguinod, "Policy-driven security management for fog computing: Preliminary framework and a case study," In Proc. Int. Conf. on information Reuse and integration , pp. 16-23, august 2014.

[21] Y. S. Jimoh, and M. Abdulhamid, "Dragonfly algorithm-based detection technique for man-in-the-middle attack in fog Computing Environment: A Conceptual Framework," Proceedings of the 1st national communication engineering conference , pp.1-6, 2018.

[22] 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, 2017.

[23] B. P. Nikkhah, H. H. S. Javadi, T. Dargahi, and A. Dehghantanha, "A hierarchical key predistribution scheme for fog networks," Computing, vol. 6, no.1 pp 1-14,2017.

[24] K. Lee, D. Kim, D. Ha, U.Rajput, and H. Oh, "On security and privacy issues of fog computing supported Internet of Things environment," 6th Int. Conf. on the network of the future, pp. 1-3, September 2015.

[25] R. Lu, K. Heung, A. H. Lashkari, and A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT," IEEE Access, vol. 5, pp. 3302-3312, 2017.

[26] I. Stojmenovic, and S. Wen, S, "The fog computing paradigm: Scenarios and security issues," In Federated Conf. on Computer Science and Information Systems pp. 1-8, 2014.

[27] M. Burrows, M. Abadi, and R. Needham,R "A logic of authentication', 1990, ACM Trans. Comput. Syst , vol. 8, no. 1, pp. 18-36, 1990.

Milk Purity Recognition Software through Image Processing

Alvarado-Díaz Witman[1], Meneses-Claudio Brian[2], Roman-Gonzalez Avid[3]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades (UCH), Lima, Peru

***Abstract*—Currently in Peru, there is a per capita milk consumption of 87 kg per year; however, the Food and Agriculture Organization of the United Nations (FAO) recommends a consumption of 120 kg per person; the industry, when the milk is acquired from small livestock suppliers, does not analyze the milk before buying it, which there is a high risk that the milk is adulterated with water, in this sense, it proposes an alternative way of preliminary detection of the presence of water in milk, only through a laser a photograph, which greatly reduces the costs of milk analysis. Milk contains different nutrients, vitamins and minerals, which are beneficial for people, so it is very known if it is adulterated or not, that way to prevent diseases. In this document, the reader will read an alternative to the existing methods for the analysis of milk, for the presented method the application of Matlab Classification Learner and the fine K-Nearest Neighbors (KNN) algorithm were used, in which a success rate of 95.4% was obtained.**

Keywords*—Milk; adulterated milk; milk with water; milk analysis; image processing; classification learner; image processing

I. Introduction

One of the objectives of the Ministerio de Agricultura y Riego (MINAGRI) for the year 2021, is to achieve a national production of raw milk of 2.7 million tons, with an average yield of 9.8 kg per day per cow and a per capita consumption of 120 kg per person per year. Milk is an essential food, since it contains proteins of high biological value, between 3 and 5%, it also provides an important percentage of fundamental elements such as potassium, phosphorus, Vitamins A, B, B2, D, Niacin, necessary fats and minerals such as magnesium, potassium, zinc and phosphorus.

There are different works to verify if there are adulterated material in the milk, in [1], [2], [3] and [4] it is carried out through the study of the impedance, using different techniques such as the use of MSNC (Minimum Square Nonlinear Complex), KNN (K-Nearest Neighbors), LDA (Linear Discriminant Analysis), respectively, with papers of this type, adulterated milk has been identified with an approximate efficiency of 94.9%.

In the research [5] they mention that in countries like India, adulteration in milk implies health problems, and even corruption, they also mention that there are compounds that can adulterate milk such as: water, flour, starch and even urea in undetectable measures. In the work, a low-cost optofluidic microviscometer is described, whose operation is based on the linear relationship between the dynamic viscosity and the width of the channel derived from the flow of two immiscible fluids within a channel, based on the operating principle of the Hagen-Poiseuille flow equation. In the study, about 60 milk samples with various adulteration ratios were analyzed. It was finally found that the device was accurate enough to measure the entire range of adulteration ratios with 95% accuracy.

In [6], they present a hardware platform based on Near Infrared Spectroscopy (NIR) used to determine the concentration of substances in solid and liquid samples, which an optical condenser with fixed lenses was developed in order to increase the Signal-Noise relation, also LEDs were used as a light source and InGaAsSb sensors, to detect diffusedly reflected light. It mentions that the instrument described in the article was tested in samples of milk and water mixtures, achieving an efficiency of approximately 99%. In [7] they use a similar method and mention hat spectroscopy is an alternative solution to detect the addition of water in milk, the method proposed is based on diffuse reflectance, the electronic device created has a reflectance of 88%, using infrared LEDs and a photodiode In-Ga-As-Sb.

In [8] they mention that milk analysis is an important role in determining its quality; therefore, in the article they describe the pH behavior of milk, in addition to its density. The impact of the addition of water on the milk produces a change in the pH electrode potential by approximately 1 mV and the density varies by 0.004 kg / m3.

In the research [9] a system was designed to measure milk adulteration, the system shows an accuracy of approximately 95% with a deviation of 1%; The system was tested to measure the percentage of fats and water; another addition to the system is the heater interface, which it seeks to perform tests for the ultrasonic sensor by transporting the cold samples to environmental temperature, to obtain greater precision.

In [10] there is a need for new and more robust techniques for the detection of adulterated milk; therefore, they propose a prototype of a digital, microcontrolled and portable photometer, which uses three infrared LEDs, with this equipment, they measure the transmittance of infrared radiation in milk samples, which evaluate whether it has water or not. The team demonstrated an average absolute error less than 1%.

In [11] they mention that the electrical conductivity of milk is frequently used for diagnostic tests of milk quality, in addition the electrical conductivity of milk can be an indicator of mastitis, which is a disease of cattle, and as a method to

measure the amount of water, present in milk. Therefore, its objective is to design a system that allows to detect the percentage of water in milk and subsequently predict the presence of mastitis.

In [12], a non-contact method for the detection of milk adulteration is proposed, considering approximately 500 samples of milk in three categories, cow's milk, buffalo milk, skim milk; the parameter involved in the detection was the refractive index, which changes as the amount of water present in the milk changes. According to the results shown, the accuracy was approximately 95%.

Section II presents the methodology that has been followed for the research work. Section III contains the preliminary results obtained. In Section IV are the discussions and finally in the section, the conclusions of the research is presented.

II. Methodology

For the present work, the methodology to follow is schematized in the block diagram shown in Fig. 1.

A. Acquisition of Images

For the image acquisition stage, it makes mixtures of fresh milk and water according to Table I, mixtures are made in a 100ml beaker, which is placed on a dark surface, pointed to the liquid with a laser and a photo is captured, it should be taken into account that the photos are taken at an approximate distance of 7 cm from the glass; in total, 675 images were acquired for each mixture of milk and water.

From Table I, it obtains 11 classes for the processing and classification stage which are named as: M100W00, M90W10, M80W20, M70W30, M60W40, M50W50, M40W60, M30W70, M20W80, M10W90 and M00W100, these names correspond to the labels that will be placed on each image.

B. Image Processing

For this stage the Thresholder color tool was used as the reader can see in Fig. 2, in which it selected four areas of interest, in which variations of the color are presented, due to the interaction of the laser with the mixture, which produces a superficial diffuse reflection, producing in the photographs zones differentiated by color; a mask is created for each zone and the code generated by the application is imported.

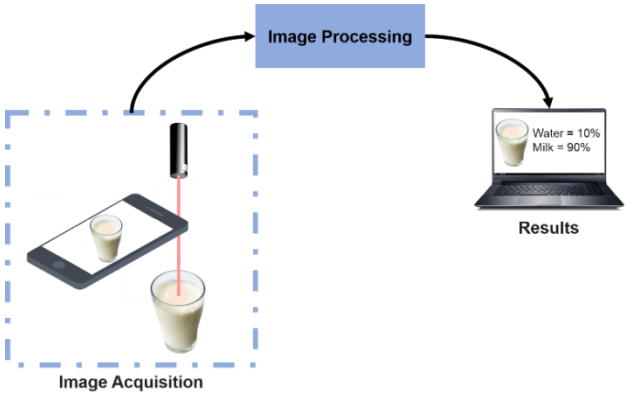


Fig. 1. Work Scheme.

TABLE. I. Chart of Milk and Water Proportions

Milk (ml)	Water (ml)		Milk (ml)	Water (ml)
100	0		40	60
90	10		30	70
80	20		20	80
70	30		10	90
60	40		0	100
50	50			

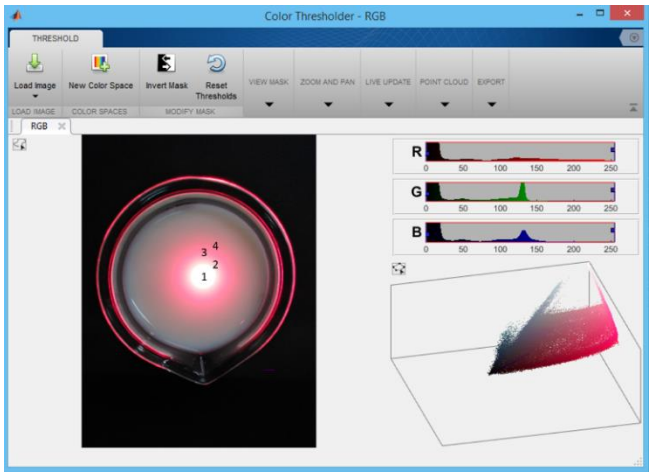


Fig. 2. Thresholder Color Tool.

In order to perform the same procedure on all images, an algorithm was created, of which it can see its general scheme in Fig. 3.

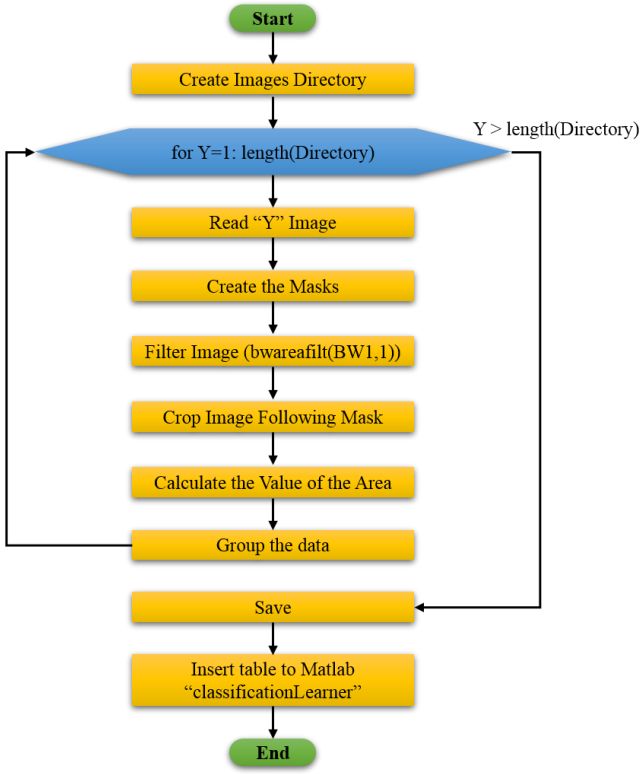


Fig. 3. Scheme of Data Processing.

Fig. 3 shows the procedure to create masks, filter and crop the image, for the four zones mentioned above, of which the corresponding area is calculated, which will be a characteristic for the classification stage with the Classification Learner application.

III. Results

In Fig. 4(a), it can see an original image which is the milk concentration equivalent to 30% that is 30ml of milk and 70ml of water, in Fig. 4(b), 4(c), 4(d) and 4(e) are cropped images corresponding to the four characteristics previously missioned.

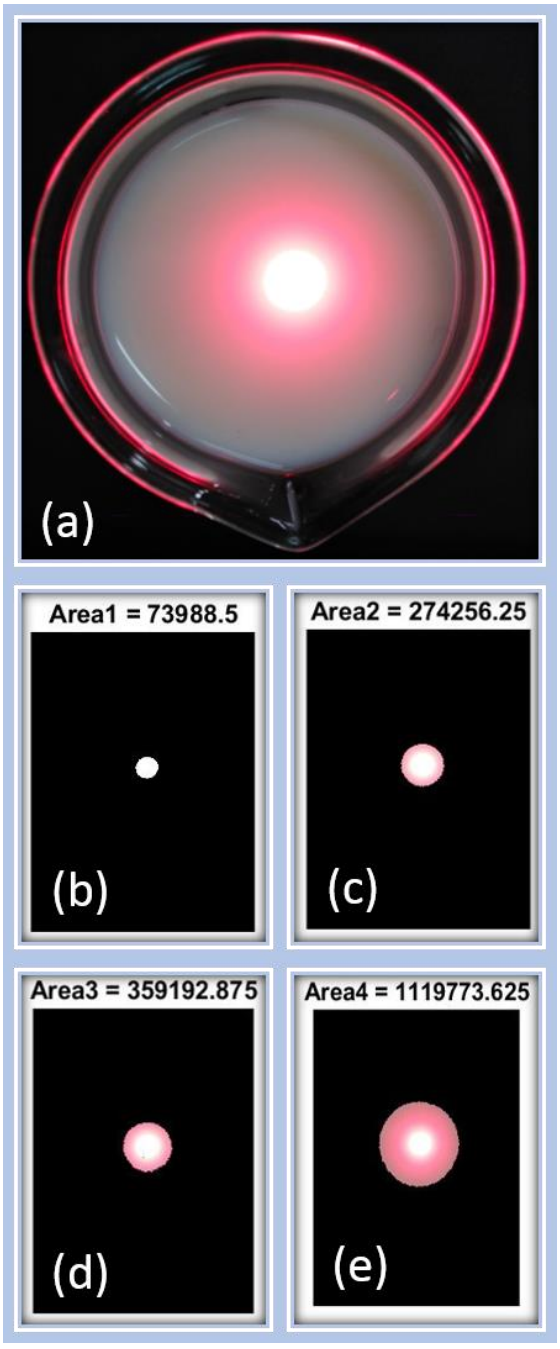


Fig. 4. (a) Photograph of Adulterated Milk; (b) Cropping of the Central Zone; (c) Cropping the Second Zone (d) Cropping the Third Zone (e) Cropping the Fourth Zone.

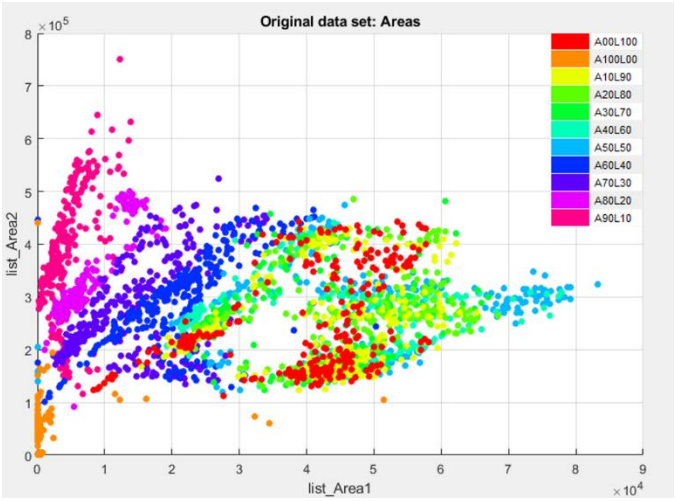


Fig. 5. Graphing the Data in the Classification Learner Application.

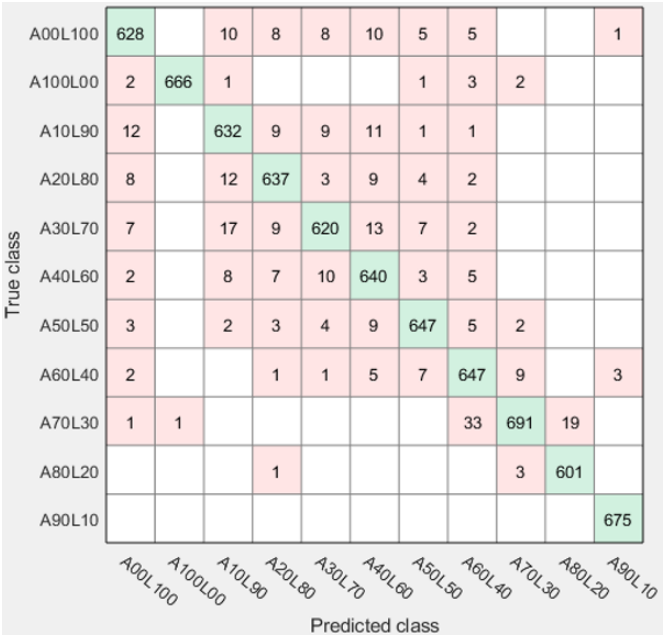


Fig. 6. Confusion Matrix Generated by the Classification Learner Application.

All the file of characteristics because of the algorithm described in the previous section, are introduced in a Matlab tool called Classification Learner, which generates a graph of the data as it shows in Fig. 5. The fine k-nearest neighbors (KNN) algorithm was chosen to classify the data.

Using the aforementioned algorithm, the application generates an approximate success rate of 95.4%, also generate a confusion matrix as the reader can see in Fig. 6, which the performance of the chosen algorithm is evaluated.

IV. Discussion and Evaluation

For the detection of adulterated milk with water, there are many jobs as in [5] where an efficiency of 95% was obtained through the study of the density of the sample; as well as in [6], where a 99% success rate was obtained using sensors and infrared light; in [9] they show an accuracy of approximately

95% with a deviation of 1%; there are research as in [2] that by carrying out a study of the electrical impedance they obtain an efficiency of 94.9% applying the KNN algorithm; in [12] a distinction is made between milk of three categories, cow's milk, buffalo milk, skim milk, obtaining an accuracy of approximately 95%; finally in [13] they mention that they have obtained an efficiency percentage of 83% using neural networks; in our case, it obtains similar results by applying the KNN method, image processing and Classification Learner application, which it obtains efficiency in the classification of 95.4%.

It is important to mention that no papers related to the study of milk adulteration were found, through image analysis, references obtained and compared in this section, are papers that at least are related to the algorithm applied for the classification stage.

V. Conclusions

Image treatment is widely applied nowadays in different industries, in this document it demonstrates that it can apply image treatment to detect the amount of water presented in a 100ml sample of milk, only by pointing to the sample with a laser and taking a photography, it is shown that the procedure used has a very good efficiency, equivalent to other methods of detecting milk adulteration.

Future work, the main action is in improving the efficiency and the proposed method, studying whether the laser power influences the calculations performed as well as the classification; the main difficulty of this paper is that it cannot detect adulterated milk with substances other than water, for this reason the possibility of conducting studies that identify compounds other than water using a method equal or similar to that proposed in this work remains open for future works.

References

[1] S. Das, M. Sivaramakrishna, B. Goswami, and K. Biswas, “Study of Electrical Equivalent Model of The PMMA Coated Probe Dipped In Milk And Milk Adulterated With Tap Water,” Int. Conf. Ind. Electron. Control Robot., pp. 53–58, 2010.

[2] G. Durante, W. Becari, F. A. S. Lima, and H. E. M. Peres, “Electrical Impedance Sensor for Real-Time Detection of Bovine Milk Adulteration,” IEEE Sens. J., vol. 16, no. 4, pp. 861–865, 2016.

[3] T. K. Mondal and S. Das, “Discrimination of various Milk adulterated sample using Linear Discrimination Analysis(LDA),” 11th IEEE India Conf. Emerg. Trends Innov. Technol. INDICON 2014, 2014.

[4] M. Sude and K. Ghodinde, “Electrical impedance Sensor for Real-Time Detection of Urea and Starch in Milk,” 2019 3rd Int. Conf. Trends Electron. Informatics, no. Icoei, pp. 431–434, 2019.

[5] P. S. Venkateswaran, A. Sharma, S. Dubey, A. Agarwal, and S. Goel, “Rapid and Automated Measurement of Milk Adulteration Using a 3D Printed Optofluidic Microviscometer (OMV),” IEEE Sens. J., vol. 16, no. 9, pp. 3000–3007, 2016.

[6] L. De Souza Ribeiro, F. A. Gentilin, J. A. De França, A. L. De Souza Madureira Felício, and M. B. M. De França, “Development of a hardware platform for detection of milk adulteration based on near-infrared diffuse reflection,” IEEE Trans. Instrum. Meas., vol. 65, no. 7, pp. 1698–1706, 2016.

[7] L. Da Silva Dias, J. C. Da Silva, A. L. De Souza Maudeira Felicio, and J. A. De Franca, “A NIR Photometer Prototype with Integrating Sphere for the Detection of Added Water in Raw Milk,” IEEE Trans. Instrum. Meas., vol. 67, no. 12, pp. 2812–2819, 2018.

[8] R. B. Somasagar and A. Kuasagur, “Potentiometrie and pH based electronic method for dilution detection in milk,” 2017 Int. Conf. Intell. Comput. Instrum. Control Technol. ICICICT 2017, vol. 2018-Janua, pp. 1762–1765, 2018.

[9] A. Dave, D. Banwari, S. Mansinghani, S. Srivastava, and S. Sadistap, “Ultrasonic sensing system for detecting water adulteration in milk,” IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, pp. 2228–2231, 2016.

[10] M. Moreira et al., “A Low-Cost NIR Digital Photometer Based on InGaAs Sensors for the Detection of Milk Adulterations with Water,” IEEE Sens. J., vol. 16, no. 10, pp. 3653–3663, 2016.

[11] L. M., Alonso V., and A. M. M., “Design and construction of a system for measuring the concentration of water in milk,” Proc. - Electron. Robot. Automot. Mech. Conf. CERMA 2006, vol. 2, pp. 47–51, 2006.

[12] A. Dave, D. Banwari, S. Srivastava, and S. Sadistap, “Optical Sensing System for Detecting Water Adulteration in Milk,” IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, pp. 2228–2231, 2016.

[13] M. A. Putra, M. Rivai, and A. Arifin, “Milk Assessment using Potentiometric and Gas Sensors in Conjunction with Neural Network,” Proceeding - 2018 Int. Semin. Intell. Technol. Its Appl. ISITIA 2018, pp. 409–412, 2018.

Time and Frequency Analysis of Heart Rate Variability Data in Heart Failure Patients

Galya N. Georgieva-Tsaneva
Institute of Robotics, Bulgarian Academy of Sciences
Sofia, Bulgaria

***Abstract*—The paper presents a mathematically based analysis of heart rate variability of two groups cardiology records: healthy individuals and patients diagnosed with heart failure. The main objective of the study is to perform a parametric evaluation of the cardiovascular system of the human body using time domain and frequency domain analysis of heart rate variability. Making distinguish between diseased individuals and healthy individuals is an interesting challenge that contemporary researchers are working on. Cardiologic records obtained through continuous Holter monitoring (24 hours) were used to address the issues in this study. The obtained results show significantly reduced values of most of the studied parameters in the time domain and the frequency domain in patients with heart failure compared to healthy individuals. The low values of the studied parameters indicate low variability of heart rate and poor overall health status. The graphical results of two study groups are shown when applying the modified Welch periodogram. These graphical results give a visual idea of the variability of time series in healthy and diseased individuals. The obtained numeric and graphical results show that heart failure patients can be distinguished from healthy individuals. The applied mathematical methods in heart rate variability studying can be used as an aid in the cardiology practice of doctors.**

Keywords*—Heart rate variability; time domain analysis; frequency domain analysis; Welch periodogram; heart failure

I. Introduction

Today's non-invasive methods for analyzing the functional activity of the human cardiovascular system are preferred methods of investigation because of the almost imperceptible nature of data capture. Today, the use of technological tools, methods of analysis and diagnostics that extract maximum information about the condition of the individual with minimal impact on his body is coming to the fore.

Electrocardiography (ECG) and Holter monitoring are the best methods for recording the electrical activity of the heart, providing data to measure the time intervals between heartbeats. The sampling rate of the data recording system is an important factor [1] in the accuracy in calculating cardiac parameters. Electrocardiographic data are very specific to each individual [2] and can be used in systems for analysis, diagnosis, treatment, and recognition.

The measure of the number of variations in the intervals between successive heartbeats (RR intervals) over time is known in the scientific literature as heart rate variability (HRV) [3]. An expression "RR interval variability" also is used to indicate these changes because it takes into account the variability in successive heart rate intervals. In healthy human organisms, high heart rate variability is normally reported. Even at rest, in healthy subjects have frequent changes in the length of consecutive cardiac intervals. The duration of cardiac intervals is influenced by the activity of the sympathetic nervous system, by the activity of the parasympathetic nervous system, and by the influence of various humoral factors. In healthy human organisms, the heart rhythm is regulated by the neurons of the autonomic nervous system (ANS) and is affected by the hormonal control balance. HRV takes into account the influence in the human body of many factors such as metabolism, respiratory process, hormones, nutrition, physical activity, emotional state, baroreceptor reflex, cycles day/ night and sleep-wake cycle, stress and others.

Heart rate variability is a very useful and essential method in the study of diseases that have cardiovascular, neurological, endocrine and other pathology. This method is used in the description of autonomic dysfunctions of patients, monitoring the natural fluctuations of the autonomic function of the nervous system, assessing changes occurring in the human body after various interventions, and to make a prognosis for the development of the disease. Heart rate variability parameters have diagnostic importance for diseases such as hypertension, thyroid pathology, neurological disorders, brain tumors, multiple sclerosis, and many other diseases.

HRV may be a predictor of death in patients with acute myocardial infarction [4]. Several researchers have studied the relationship between autonomic regulation of cardiac activity and cardiac mortality. A study by the North American Society of Pacing and Electrophysiology, including patients with past myocardial infarction (MI), found that low heart rate variability was more correlated with the risk of sudden death than other commonly accepted clinical indicators development of the patient's disease. This methodology gained popularity and in 1996 the Standards of measurement, physiological interpretation, and clinical use [5] were developed and adopted by the European Society of Cardiology and the North American Society of Pacing and Electrophysiology.

The rest of the paper is summarized as follows: Overview of heart failure research studies was presented in Section II. Research Background was shown in Section III. Heart Rate Variability Time Domain Analysis was presented in Section IV and Frequency Domain Analysis - in Sections V. Data Preprocessing was described in Section VI. The experimental results were presented and discussed in Section VII. Conclusions are presented in Section VIII.

II. Review of Heart Failure Studies

Heart failure is a serious disease of a person's cardiovascular system, which has a close relationship with the condition of the individual's autonomic nervous system. Usually, heart failure corresponds to the presence of abnormal activity of the autonomic nervous system. In most cases, this disease is accompanied by increased activity of the sympathetic nervous system and decreased the activity of the parasympathetic nervous system. The normal variability of cardiac intervals corresponds to a good sympathetic balance in the human body. In patients with heart failure, the reduced or missing low-frequency spectral component is observed. According to [6], the missing low-frequency component in heart failure individuals reflects impaired baroreceptor function. Therefore, the ability to make a sufficiently accurate assessment of heart rate variability is important for the detection, monitoring, and prevention of heart failure. The non-invasive method, which studies the variability of the heart rate, can evaluate the activity of the two parts of the nervous system, the sympathetic activity of the nervous system and the activity of the parasympathetic part. These non-invasive studies are not complicated to administer; they may accompany the patient's treatment and be a means of monitoring the results of that treatment.

The heart failure disease is considered to be one of the most severe cardiac diseases that often have a poor prognosis for the patient. According to scientific research [7], up to 50% of individuals who have this disease die from sudden cardiac death. According to a study [8], the risk of sudden cardiac death in such individuals is five times higher than in healthy individuals. Author in [9] presents studies on patients with heart failure disease by determining the parameters of HRV.

III. Research Background

Frequency analysis of HRV is an effective tool for evaluating the cardiovascular autonomic activity of the heart. The following methods have been used in the scientific literature for frequency analysis of HRV data: Fast Fourier transform (the method is fast and with low computational cost); autoregressive method (performed on a small block of data and it does not need data interpolation [10] but is complex and in some cases produces erroneous results [11], [12]); wavelet analysis and other classical techniques.

The authors of [13] apply an autoregressive spectral approach to assess circadian modulation in hypertensive patients.

In research [14] to evaluate the spectral power in the cardiac data of newborn premature infants, the authors use Fast Fourier Transform.

The Burg method [15] is based on the idea of minimizing errors, has good resolution, and produces reliable spectral results and works well with different input series.

Lomb Periodogram Method - not a commonly used method, no need for input data series interpolation; but some authors [16] declare receiving insufficiently correct data.

The authors of [17] use the YuleWalker method and declare a distortion of the obtained results.

In the research [18], authors apply the Welch method to evaluate cardiac data in patients with insomnia and use a consistent estimator by averaging periodogram from overlapping intervals.

In their study [19] the authors use the classic and the Welch periodogram (implemented with Hann window) for spectral analysis of cardiac data.

The work [20] presents a temporal statistical analysis of changes in the duration of consecutive RR intervals originating from sinus rhythm.

In recent years, researchers have been developing new methods for spectral analysis. For example, trigonometric regressive spectral analysis [21] - uses trigonometric regression functions to examine variations in cardiac series.

IV. HRV Time Domain Analysis

Fig. 1 presents the model the input data recording, preprocessing, interpolation and mathematical estimation of cardiac time interval data.

Cardio data can be obtained by an electrocardiograph (5-20 minute) or Holter (long-term monitoring device - from 24/72 / 2 weeks). If the input data is in a compressed format, then it is decompressed. Performing preprocessing of cardio data: denoising [22], QRS complexes detection [23] and the RR interval obtained (R - a point corresponding to the peak cardio wave). The next step is to exclude the ectopic intervals in the resulting time series and formation of the normal-to-normal (NN) intervals. The NN time series is interpolated and downsampled. Analysis of the resulting HRV data included HRV Time Domain Estimation and Spectral Estimation.

Statistical parameters. The statistical time analysis of rhythmograms investigates two types of parameters: the duration of the NN intervals and the difference in the duration of the adjacent NN intervals.

The following indicators (Table I) are used to analyze the duration of NN intervals: SDNN, SDANN, SDNN index. To evaluate the variability of the adjacent NN intervals, the following parameters are calculated: NN50, pNN50, RMSSD. The SDNN (ms) parameter reflects the standard deviation of all NN intervals in cardiac monitoring; characterizes the state of regulation mechanisms; indicates the total effect of influence on the sinus node of sympathetic and parasympathetic parts of the autonomic nervous system. SDANN (ms) reflects the standard deviation of the duration of average NN intervals every 5 minutes from the record; SDNN index (ms) - mean of standard deviations for all 5-minute blocks of the observation period.

NN50 is a number of pairs of consecutive NN intervals that differ by more than 50 ms over the entire observation period; pNN50 (%) is a percentage of consecutive R – R intervals, the difference between which exceeds 50 ms; RMSSD (ms) - a square root of the sum of squares of differences in the values of consecutive pairs of NN intervals (the activity index of the parasympathetic system of autonomic regulation) [20].

TABLE. I. TIME DOMAIN PARAMETERS

Parameter [units]	Formula
SDNN [ms]	$SDNN = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(RR_i - \overline{RR})}$
SDANN [ms]	$SDANN = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{RR_i} - \overline{\overline{RR}})^2}$
RMSSD [ms]	$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - \overline{RR_i})^2}$
SDNNindex [ms]	$SDNN_{index} = \frac{1}{N}\sum_{i=N}^{N} SDNN_i$
pNN50 [%]	$pNN50 = \frac{NN50}{NN}.100\%$
HRVTi [-]	$HRVTi = \frac{\sum_{i=1}^{N_b} b(t_i)}{max_i b(t_i)} = \frac{N-1}{max_i b(t_i)}$

Geometric parameters: HRVTi - triangular index (the proportion of all accepted RR intervals to their modal measurement at a 1/128s bins discrete scale; TINN - triangular interpolation of cardio interval histogram.

V. HRV FREQUENCY DOMAIN ANALYSIS

Numerous methods for spectral analysis of HRV have been developed in the scientific literature, the most popular of which are based on traditional Fourier transform (e.g. fast Fourier transform) [24] and wavelet theory. The spectral analysis makes it possible to quantify the different frequencies contained in the investigated signal and to study the action of the regulatory systems of the human body.

HRV analysis is performed in four frequency bands - high frequency range (HF), low frequency (LF), very low frequency (VLF) and ultra low frequency (ULF) bands [25] (Table II). Usually, the analysis is performed on 5-minute segments, not on the entire cardiology record. Using a five-minute block for frequency analysis is one of the recommendations [5] of the Task Force (1996) heart rate variability standard.

Two of the ranges are important for cardiac clinical practice - LF and HF bands. The HF component in range 0.15-0.40 Hz is associated with respiratory sinus arrhythmia (RSA) and parasympathetic activity. Approximately 1 minute of cardiac recording is sufficient to evaluate the HF components of HRV. The LF component (0.04-0.15 Hz) is not yet sufficiently studied; it is related to the sympathetic activity of the nervous system and represents both sympathetic and vagal influences. It takes at least 4 minutes [26] to obtain correct values for the power in the LF range. For this reason, the determination of spectral power generally requires the use of a five-minute cardiac record.

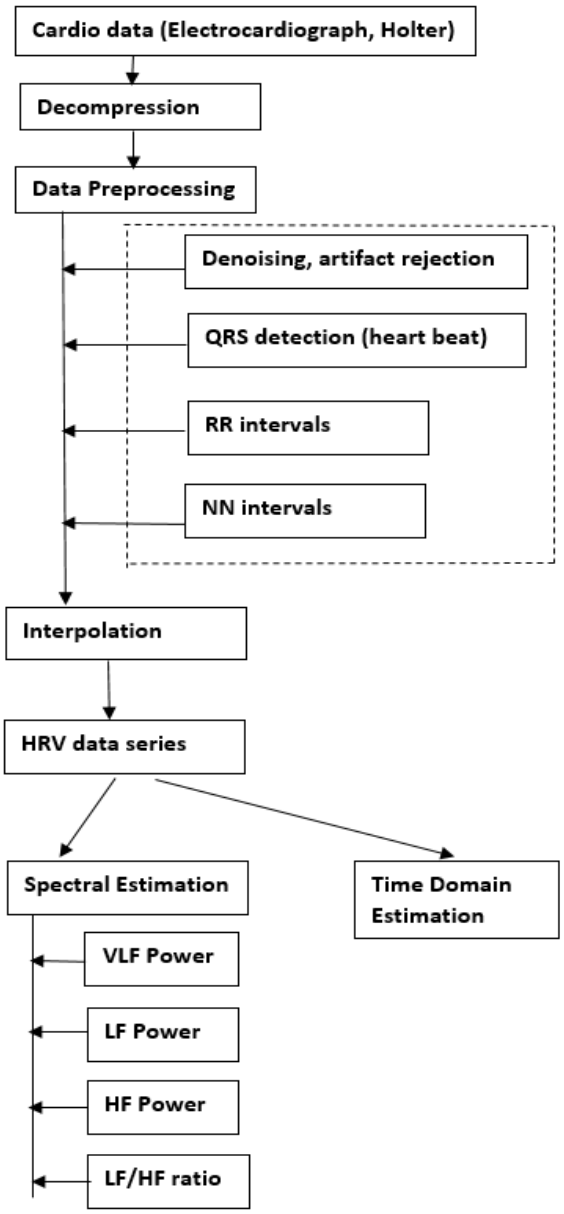


Fig. 1. Data Processing.

Spectral analysis parameters are used to evaluate the influence of the autonomic nervous system [27]. The parasympathetic branch of the autonomic nervous system is responsible for relaxation, rest, preservation of vital energy, lowers heart rate and lowers blood pressure.

The sympathetic branch of the autonomic nervous system is responsible for mobilizing the internal resources of the human body and manifests itself in conditions of physical activity, stress; it expresses acceleration of cardiac activity, increases blood pressure, accelerates breathing, raises the temperature and others.

Another important parameter - LF/HF ratio (known as sympathovagal balance) indicates the balance between sympathetic and vagal tones.

TABLE. II. FREQUENCY RANGE

Power [ms^2]	Frequency [Hz]	Cause
ULF - ultra low frequency	0-0.003	day/night cycle
VLF - very low frequency	0.003-0.04	sympathetic activity
LF - low frequency	0.04-0.15	sympathetic and parasympathetic activity
HF - high frequency	0.15-0.4	respiratory sinus arrhythmia; parasympathetic activity

A method with modified Welch Periodogram: With the Welch Periodogram method, the data series is split into blocks that can overlap. The data from each block is processed by the weighted window function and then the calculations are performed. The software calculates the periodograms in each Welch block [28], thereafter it starts to average for reducing variance.

The next step is to calculate the modified Welch Periodogram [29] in each block, using the following formula:

$$Welch_{ModifiedP}(f) = \frac{1}{U.K}.\left|\sum_{j=0}^{K-1} x_i\,(j).w(j).e^{-2j\pi fn}\right|^2,$$

Where:

U – Normalization factor;

K - Elements in each block;

N – Number of blocks that was overlap;

$w(j)-$ Window function;

$x_i(j)$ – input data;

f - frequency;

$n = 0{,}1 \ldots K-1;$

$i = 0{,}1 \ldots N.$

The obtained average Periodogram for all blocks is determined by:

$$Welch_{Periodogram}(f) = \frac{1}{U.K.N}\sum_{i=0}^{N-1}\left|\sum_{j=0}^{K-1} x_i\,(j).w(j).e^{-2j\pi fn}\right|^2.$$

VI. DATA PREPROCESSING

Before performing time and spectral analysis, the input data is subjected to preprocessing. This treatment involves sampling, digitizing, artifact identification and rejection, input data editing, RR interval rejection. The device that captures the input data must be able to recorded data at a frequency greater than 200 Hz [10] for the cardiac complexes to be correctly identified in the next steps. In the artifact identification and rejection step, all artifacts (technical and other) and arrhythmic events must be removed for correct mathematical analysis.

Obtaining NN interval series. The NN interval sequence is derived from the RR interval sequence excluding the extrasystoles (the HRV studies are based on the normal sinus rhythm, and it excludes the involvement of extrasystoles). Extrasystoles are extraordinary cardiac contractions that do not originate from the sinus node. A healthy person can also have extrasystoles, but their amount is negligible (e.g. 30 per hour, which is 0.5 extrasystoles per minute). In case of abnormal occurrence, extrasystoles may reach one in each second.

Interpolation. The choice of the type of interpolation depends on the specific methods of data capture depends on, the selected methods of mathematical analysis, the quality of the input cardiology records, the type of extrasystoles, also the characteristics of the cardiac signal reflecting the type of disease. It is carried out when it is necessary to fulfill the requirement for uniformity of the studied interval time series.

VII. RESULTS

A. DataBase

Preprocessing and selection of records. The process of cardiac data recording is very sensitive to various factors: network disturbances, patient's respiratory process, quality of contact between the electrodes and the patient's skin, patient attention to the proper attachment of the electrodes to his body, etc. For this reason, the obtained data are carefully inspected for damage to the records and where such sections are found and the record is removed and it will be not an object of study (in case of major damage therein).

B. Subjects

The cardio data used in this paper were obtained with the help of Holter monitoring from the Varna Medical University, the Republic of Bulgaria. The cardio data is continuous electrocardiographic Holter monitoring records, second Lead. The control group of healthy people (volunteers) is in the same age group. All studied people provided informed written consent. Therefore all identified data were removed. All data used in this study is anonymous to protect the personal information of all Volunteers and Patients.

Two groups of HRV data records were chosen: 22 healthy individuals and 24 patients diagnosed with heart failure. Holter monitoring was performed on all for 24 hours.

C. Characteristics of the Subjects

Table III present the demographic characteristic (the age and gender distribution, the mean age of the individuals). The analyzed patient records are about 24 individuals aged 35-55 years including 13 males and 11 females. The analyzed healthy records are about 22 individuals aged 34-52 years including 12 males and 10 females. Values are expressed as mean ± standard deviation (SD) or in percent (%). No significant difference between the different groups according to demographic characteristics.

TABLE. III. DEMOGRAPHIC CHARACTERISTICS

Parameter	Heart failure N=24	Healthy N=22	P value
Gender, Men %	54.16	54.54	NS (0.9116)
Age ± SD	47.83± 4.37	47.50±4.64	NS (0.803)

Fig. 2 shows the RR intervals data series from a healthy individual. The data are distinguished by a wide amplitude of RR intervals values (from 0.38 to 1.6 seconds). Fig. 3 shows the RR intervals data series for a patient diagnosed with heart failure disease. The lengths of the time intervals are clustered/grouped around the value of 0.5 sec. This figure graphically presents a low variability in cardiac interval values.

D. Time Domain Results

Table IV presents the obtained results of the HRV analysis in Time Domain obtained from a study of two groups of Individuals (total number 46): Healthy Individuals and Heart Failure Individuals.

The calculations show that the main value of MainRR is 832.98 ms, which is much higher than the main value of MeanRR (632.44 ms) for patients with heart failure. The MeanHR of the healthy people (72.03 bpm) is much lower than the MeanHR (94.87 ms) in the studied patients' group. The assessments of the mean value of RR intervals and the mean heart rate (HR) indicate that these parameters have statistically significant ($p<0.005$ for Mean RR and $p<0.01$ for Mean HR).

The parameter SDNN for Heart failure patients (mean value 114.06 ms) is not that high that the SDNN calculated in the group of the healthy people (mean value 142.18 ms). The same thing is observed with the SDANN parameter (92.87 ms of heart failure group versus 122.16 of healthy). The parameter RMSSD is slightly reduced in heart failure people compared to healthy people (26.73 ms versus 30.41 ms). From time domain studied, the values of SDNN ($p<0.0001$), SDANN ($p=0.0013$), RMSSD ($p=0.0373$) and pNN50 ($p=0.0494$) have statistical significance ($p<0.05$). The SDindex (60.37 in sick versus 64.08 in healthy) hasn't statistical significance ($p>0.05$).

For geometric parameters, the mean value of HRVTi (18.31) is lower in sick people than in the healthy (24.11). This parameter has statistical significance ($p=0.0261$). The TINN (433.54 in sick versus 518.91 in healthy) hasn't statistical significance ($p>0.05$).

In conclusion in the group of patients diagnosed with heart failure, the values of almost all studied parameters in the time domain are significantly lower compared to the Healthy individuals.

Many of the time domain parameters in patients have larger SD (Table IV). This is most likely due to the larger variations in these parameters in different patients, depending on the degree of their disease.

E. Frequency Domain Results

Spectral method realization. In this study, the Welch periodogram was implemented using a Hamming window function, overlapping at 50% was performed. RR intervals data is interpolated with a cubic spline basis and then sampled at 4 Hz.

Table V presents the results of the HRV analysis in Frequency Domain obtained from the study of two groups studied. Signal power values in the low frequency range (absolute and normal values) and the high frequency range (absolute and normal values) were investigated and LF/HF ratio. The results show that all spectral parameters are significantly lower in patients diagnosed with heart failure. When examining the spectrum in absolute value in LF and HF range Pvalue has values <0.0001. When examining the spectrum in normal units in the LF and HF range Pvalue has values <0.05. The sympathetic balance index has a value of 1.16, which is significantly lower ($p<0.05$) in the heart failure group than the value of this index 1.56 (included in the normal values recommended by the standard for variability) in healthy individuals.

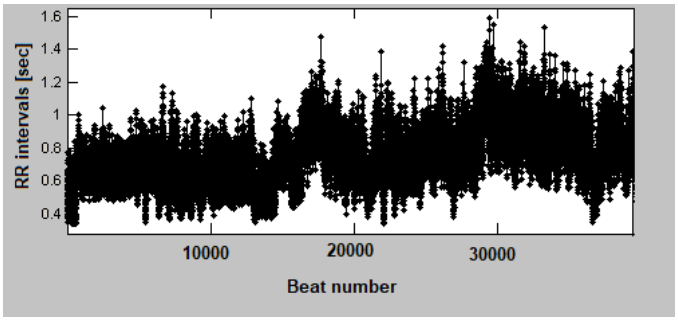


Fig. 2. RR Intervals of a Healthy Individual.

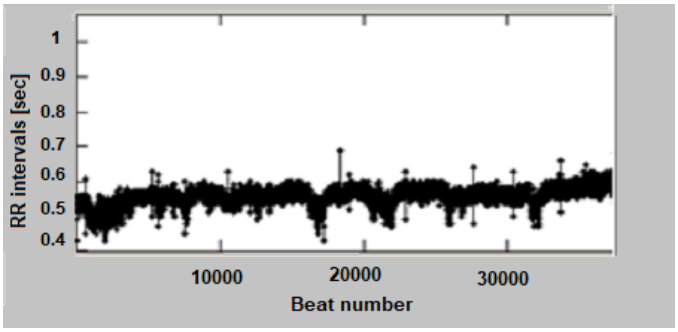


Fig. 3. RR Intervals of Heart Failure Individual.

TABLE. IV. PARAMETERS IN THE TIME DOMAIN

Para-meter	Heart failure N=24	Healthy N=22 (mean± sd)	P value mean± sd
MeanRR [ms]	632.44 ± 263.71	832.98± 144.32	<0.005 (0.0029)
MeanHR [bpm]	94.87± 36.29	72.03± 8.91	<0.01 (0.0063)
SDNN [ms]	114.06± 18.12	142.18± 20.31	<0.0001
SDANN [ms]	92.71± 26.84	122.16± 31.08	<0.005 (0.0013)
SDindex [ms]	60.37± 28.94	64.08± 15.62	NS (0.5961)
RMSSD [ms]	26.73± 10.95	30.41± 6.08	<0.05 (0.0373)
pNN50 [%]	11.04± 8.33	14.92± 3.56	<0.05 (0.0494)
HRVTi	18.31±9.62	24.11± 7.16	<0.05 (0.0261)
TINN [ms]	433.54± 172.98	518.91± 131.47	NS 0.068

TABLE. V. PARAMETERS IN THE FREQUENCY DOMAIN

Para-meter	Heart failure N=24	Healthy N=22 (mean± sd)	P value mean± sd
LF [ms^2]	663.83± 102.03	1204.67± 247.53	<0.0001
LF (nu)	0.5379± 0.1108	0.6094± 0.1207	<0.05 (0.042)
HF [ms^2]	570.31± 108.12	772.18± 209.31	<0.0001
HF (nu)	0.4621± 0.1084	0.3906± 0.1108	<0.05 (0.0323)
LF/HF	1.16±0.62	1.56±0.47	<0.05 (0.0184)

The studied groups were also compared using a graphical method of the Welch periodogram.

Fig. 4 shows a PSD of a healthy individual obtained by the Welch Periodogram method. The PSD values are with high values in the three tested ranges: VLF, LF, and HF.

Fig. 5 shows the PSD graphical presentation of heart failure individuals. The PSD in VLF, LF and HF area are with small values. This shows a low heart rate variability of RR interval series, the predictor of serious cardiac disease.

The use of the Welch Periodogram method shows significant differences between the spectral parameters in healthy people and heart failure individuals.

F. Statistical Analysis

Descriptive statistics of data are presented as mean ± standard deviation (SD). For statistical analysis, the T-test has been used. The p-value of <0.05 was considered as statistically significant.

There are some limitations to performing with the analysis above. First, the study included a limited number of patients and healthy people (22 healthy individuals and 24 patients with heart failure). Secondly, the study uses time and frequency domain analysis. These limitations are imposed by the stage of work on the study of heart rate variability in healthy people and patients with heart disease. This study demonstrates the ability to distinguish the diseased patients from healthy individuals by using time domain and frequency domain heart rate variability analysis.

G. Discussion

Differentiating healthy people from diseased individuals is extremely useful in the treatment of cardiac diseases. The use of interdisciplinary approaches to addressing health problems would lead to an improvement in the quality of healthcare and an increase in the health status of the population. For these reasons, the in-depth penetration of mathematical technologies into the study of phenomena such as heart rate variability may prove particularly useful and may enter clinical practice soon.

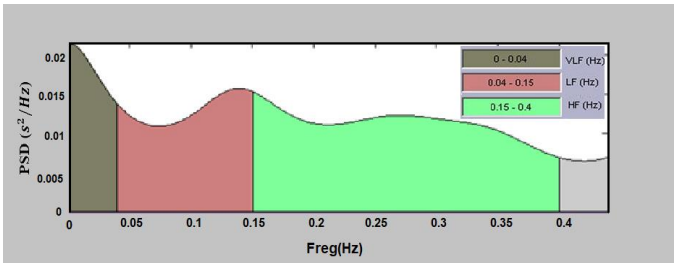


Fig. 4. PSD of a Healthy Individual.

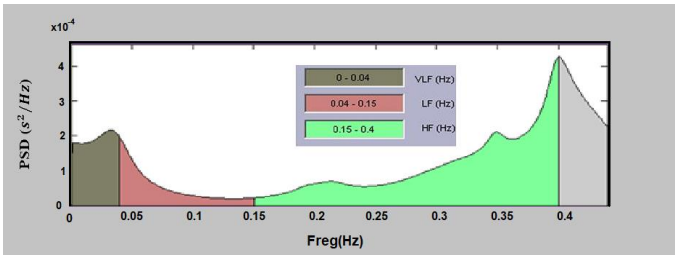


Fig. 5. PSD of Heart Failure Individual.

VIII. CONCLUSION

The paper presents effective mathematical methods for studying the parameters of heart rate variability in the time and frequency domain. Examination of cardiac record parameters of patients diagnosed with heart failure shows significant differences from certain values of the same parameters in healthy individuals. HRV parameters in the time domain and frequency domain are significantly reduced in heart failure people, and in some cases parameters (such as the LF frequency parameter) are times smaller than in healthy people. This shows extremely low variability in the time series of cardiac intervals, risky low health status and in some cases risk to the patient's life. The parameters studied can be determined during the treatment of patients and the change (in a positive or negative direction) of their values can be monitored. This method can be used for predictive purposes to determine the course of disease development during the treatment of patients.

IX. FUTURE WORK

Analysis of the parameters of HRV will be performed with other mathematical methods: nonlinear methods, fractal and wavelet analysis. It is planned to investigate the parameters in patients with other cardiac diseases and to evaluate the impact of these diseases on values of heart rate variability.

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] T. Kuusela, "Methodological aspects of heart rate variability analysis, In Heart Rate Variability (HRV) Signal Analysis: Clinical Applications" (eds M.V. Kamath, M.A. Watanabe, A.R.M. Upton) Boca Raton, FL: CRC Press; pp. 9–42, 2013.

[2] S. Hadiyoso, S. Aulia, A. Rizal, "One-Lead Electrocardiogram for Biometric Authentication using Time Series Analysis and Support Vector Machine", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, pp. 276-283, 2019.

[3] V. Vesterinen, K. Häkkinen, T. Laine, E. Hynynen, J. Mikkola, and A. Nummela. "Predictors of individual adaptation to high-volume or high-intensity endurance training in recreational endurance runners". Scand J Med Sci Sports, Vol. 26(8), pp. 885–93, 2016.

[4] S. Marchev, "Heart Rate Variability − Measurement Standards". Cardio − Vascular Diseases. 1998, Vol. 1, pp. 28-35 (in Bulgarian).

[5] 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.

[6] J.E. Sanderson, "Heart Rate Variability in Heart Failure", Heart Failure Reviews, Vol. 2, Issue 4, 1998 pp. 235–244, https://doi.org/10.1023/A:1009745814816.

[7] M. Cipriani, B. De Chiara, E. Ammirati, et al. "Right ventricular dysfunction in advanced heart failure". Giornale Italiano Cardiologia. Vol. 15 (7-8), pp. 430–40, 2014.

[8] A. Shah, B. Claggett, N. Sweitzer et al. "Cardiac Structure and Function and Prognosis in Heart Failure With Preserved Ejection Fraction: Findings From the Echocardiographic Study of the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT)", Trial. Circ. Heart Fail., 2014; Vol. 7 (5), pp. 740–51.

[9] R. Mahajan, T. Viangteeravat, and O. Akbilgic. "Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics". International Journal of Medical Informatics, Vol. 108, pp. 55-63, 2017.

[10] K. Li, H. Rüdiger, T. Ziemssen, "Spectral Analysis of Heart Rate Variability: Time Window Matters", Frontiers in Neurology, 29 May 2019, https://doi.org/10.3389/fneur.2019.00545.

[11] GJ Silva, MR Ushizima, PS Lessa, L Cardoso, LF Drager, MM Atala, et al. "Critical analysis of autoregressive and fast Fourier transform markers of cardiovascular variability in rats and humans". Brazilian Journal of Medical and Biological Research, 2009, Vol. 42, pp.386–96, DOI: 10.1590/S0100-879X2009000400012.

[12] D Chemla, J Young, F Badilini, P Maison-Blanche, H Affres, Y Lecarpentier, et al. "Comparison of fast Fourier transform and autoregressive spectral analysis for the study of heart rate variability in diabetic patients". Int J Cardiol., 2005, Vol.104, pp.307–13. DOI: 10.1016/j.ijcard.2004.12.018.

[13] F Badilini, P Maison-Blanche, P Champomier, JC Provost, P Coumel, H Milon. "Frequency-domain heart rate variability in 24-hour Holter recordings: role of spectral method to assess circadian patterns and pharmacological autonomic modulation". Journal of Electrocardiology, 2000; Vol. 33, pp. 147-157.

[14] M Morin, S Marchand, L Couturier, S Nadeau, S Lafrenaye. "Long-Term Persistence of Abnormal Heart Rate Variability following Long NICU Stay and Surgery at Birth". Pain Research and Treatment, Vol. 2014, Article ID 121289, 7 pages; DOI:10.1155/2014/121289.

[15] K. K. Kim, J. S. Kim, B. H. Choi, G. S. Chung, H. B. Lee, Y. G. Lim, K. S. Park, "Comparison of HRV spectral analysis methods for unconstrainedly measured ECG", Proceedings of the Fifth IASTED International Conference Biomedical Engineering, 2007, Austria, pp. 365-369.

[16] D Fonseca, AA Netto, R Ferreira, AM de Sá, editors. "Lomb-scargle periodogram applied to heart rate variability study". In: 2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC), Rio de Janerio: IEEE, 2013.

[17] M De Hoon, T V. der Hagen, H Schoonewelle, H V Dam. "Why YuleWalker should not be used for autoregressive modeling". Ann Nucl Energy. 1996, Vol. 23, pp.1219–28. DOI: 10.1016/0306-4549(95)00126-3.

[18] R. T Krafty, M. Zhao, D. J. Buysse, J. F. Thayer, and M Hall. "Nonparametric spectral analysis of heart rate variability through penalized sum of squares". Statistics in medicine, Vol. 33(8), pp. 1383–1394, 2014. DOI:10.1002/sim.6038.

[19] M Estévez, C Machado, G Leisman, T Estévez-Hernández, A Arias-Morales, A Machado, and J Montes-Brown. " Spectral analysis of heart rate variability", International Journal on Disability and Human Development, 2015; DOI: 10.1515/ijdhd-2014-0025.

[20] F. Shaffer, and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms", Front Public Health. 2017, Vol. 5: 258, pp. 1-17, DOI: 10.3389/fpubh.2017.00258.

[21] K Li, H Rudiger, R Haase, T. Ziemssen, " An innovative technique to assess spontaneous baroreflex sensitivity with short data segments: multiple trigonometric regressive spectral analysis". Front Physiol. 2018, Vol.9:10. DOI: 10.3389/fphys.2018.00010.

[22] G. Georgieva-Tsaneva, "Wavelet based interval varying algorithm for optimal non-stationary signal denoising", Proceedings of the 20th International Conference on Computer Systems and Technologies, ACM International Conference Proceeding Series, New York, USA, pp.200-206, 2019, doi>10.1145/3345252.3345268.

[23] G. Georgieva-Tsaneva, "QRS detection algorithm for long term Holter records, Proceedings of the 14th International Conference on Computer Systems and Technologies", ACM International Conference Proceeding Series, New York, USA, 2013, pp. 112-119, doi>10.1145/2516775.2516811.

[24] P-C Lin, H-Y Hsu, C-C Chang, T-C Hsiao, "Frequency Domain Analysis for Assessing Fluid Responsiveness by Using Instantaneous Pulse Rate Variability", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016, pp. 229-233.

[25] G. Ernst, Heart Rate Variability, Springer-Verlag London, 2014.

[26] A Serafi, "Heart Rate Variability (HRV) - Analysis And Clinical Significance", International Journal of Biology and Biotechnology, Vol. 15 (2), pp. 193-199, 2018.

[27] E. Gospodinova, M. Gospodinov, N. Dey, I. Domuschiev, A. Ashour, S. Balas, T. Olariu. "Specialized Software System for Heart Rate Variability Analysis: An Implementation of Nonlinear Graphical Methods". In: V. Balas, L. Jain, M. Balas (eds) Soft Computing Applications. SOFA 2016. Advances in Intelligent Systems and Computing, Vol. 633. Springer, Cham, 2018.

[28] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". IEEE Transactions on audio and electroacoustics. 1967; 15(2):70–73G.

[29] S. A. Akar, S. Kara, F. Latifoglu, V. Bilgic. "Spectral Analysis of Photoplethysmographic Signals: The Importance of Preprocessing", Biomedical Signal Processing and Control, Vol.8 (1), 2013, pp. 16-22.

E-learning Benchmarking Adoption: A Case Study of Sur University College

Saleem Issa Al Zoubi[1]
Computing Technology and Control System Department
Irbid National University, Amman, Jordan

Ahmad Issa Alzoubi[2]
Information Systems and Technology (IST) Sur University College, Muscat, Sur

***Abstract*—As an integral tool nowadays, e-learning presents fresh paradigm in the fields of education and management. The effectiveness of e-learning in the improvement of learning and teaching method has been proven. Accordingly, there have been countless works carried out to gain comprehension on the adoption of e-learning particularly in terms of its use extent. Furthermore, a new model of e-learning Success based on McLean & DeLone Information System Success (IS Success) and Diffusion of Innovation (DOI) and the effects of e-learning on student's performance are comprehensively highlighted. Appositely, e-learning benchmarking adoption among students is characterized and measured in this study with the integration of DOI with IS Success attributed with such adoption. In this quantitative study, data were gathered from the students enrolled in SUC. The results indicate that and the variables of relative advantage, complexity, system quality, information quality and service quality appear to have significant linkage to the adoption of e-learning as well as net benefit is significantly correlated with the adoption of e-learning. In regards to the used methods for examining the adoption of e-learning particularly, future studies could benefit from the use of quantitative and qualitative methods in combination.**

***Keywords*—e-learning; DOI; IS success; net benefit; extent of usage; LMS; benchmarking; quality assurance**

I. Introduction

The exponential evolution of Information Communication Technologies (ICT) and Internet are influencing the delivery of learning and teaching process as it offers an innovative paradigm for the domains education and management today. E-learning concept carries the most comprehensive meaning towards the notion of openness. In the context higher education, e-learning adoption is highly essential as it facilitates the significant transformation in the learning and teaching process. The verity of factors namely how students and instructors are supported in its use, the integration of technology within the student learning experience and the availability of technology are the most significant criteria for successful adoption and usage of e-learning within the entire organization [1, 2].

Learning style transformation has been facing challenges. Among these challenges are continuous development of users' technological skills and shifts in cultural expectations. These challenges cannot be neglected; they have to be addressed. Effectively implementation of e-learning is crucial as this will guarantee enrichment of users' experiences. For this reason, fitting usage of technological combination has to be assured as this will result in fruitful ICT application in the context of education. The usage of E-learning encompasses usage of ICT in various education domains, and such usage in the context of higher education institutions imparts user with learning quality. ICT technology usage as the conventional classrooms' add-ons, online learning or combination of both modes, are among such examples [3, 4].

E-learning has been proven to be beneficial. For instance, e-learning opens more doors to learners towards education at lower cost and not only that learner could also engage in learning at all times and venues. Besides that, geographical hindrance is solved with e-learning as the technology allows learner to gain access at their convenience and this capacity enhances the quality of education. Meanwhile, with e-learning, interactions between learners and teachers or between learners and learners are improved via the model of learning network model. Also, increase can also be seen on the level of IT literacy through e-learning, while decision making becomes faster [5, 6, 7]. For working people with interest in pursuing their studies, e-learning offers opportunities for part-time studies. In the past, part-time studies were not as available and those with interest would need to quit their job in order to attend classes. Comparatively, e-learning allows people to further their studies at select institutions while still working [8]. The presence of e-learning can be seen intertwined with the conventional learning. Not only that, e-learning has become the solution to the problems associated with the conventional learning.

In this globalization era, the significance of education in the improvement of government's human capital is recognized by governments including that of Oman. In Oman, e-learning is deemed important because it resolves the challenges presently obstructing learning in the country. Still, the adoption of e-learning in this country is rather low, just in other developing countries. Countless social transformations have been faced by the Sultanate of Oman and such have changed the needs of the Omani society. The Internet, in addition to other types of e-Learning platforms, has been integrated into the system of education in Oman. Such integration has compelled teachers and faculty members go online. Moodle, WebCT, and e-portals learning management systems are among those platforms utilized in educational institutions [9].

In the context of Oman, e-learning is quickly turning into a crucial component in the process of education in this country. E-learning is the next big thing after the printing press came into the scene. As Oman has swiftly expanding workforce

consisting of graduates that are flexible and well educated, it could greatly contribute to e-Learning within the region. Being an ambitious project launched by the Ministry of Higher Education, e-learning system provides the best imaginable e-learning solutions to both the students and staff. In Oman, e-learning is developed and implemented to encourage lifelong learning and satisfy the demand for unremitting professional development in the country. In Oman and the neighbouring countries (Arab countries), e-learning is deemed as a novel learning and teaching approach.

Accordingly, this study will examine the extent of e-learning adoption among undergraduates in Oman, particularly those enrolled in Sur University College (SUC). Further, the gathered information is compared with the benchmarking of e-learning used by international education agencies. The benchmarking to use for SUC will also be proposed. Previous empirical studies that conducted on e-learning in Oman have examined the introduction, use and the potential impact of the ICT on learning. For example, a study conducted by [10] to investigate the barriers of e-learning adoption in higher education in Oman: academics' perspectives. [11] carried out a study on impact learning management system has done on students and how effectively LMS has influenced students on their academic activities. These studies did not explain further factors that might control e-learning adoption from the students' perception. In addition, there is no evidence of studies done on factors affecting adoption of e-learning in the context of Oman as well as integrated Rogers DOI with D&M IS model. Thus, the Higher Education in Oman needs to have some guidelines to ensure the successful and efficient implementation of e-learning in higher learning institutions. In case of Oman as a developing country, till now, there is little published evidence on what is the extent of e-learning adoption.

The rest of this research is organized as follows. The next section provides a statement of the problem, significance of the study and scope of the study. Literature review related to the theories that explain diffusion of innovation and adoption will be argued. Section 3 will describe in detail the methodology adopted in this research. At the end, discussion on the results and conclusion is presented.

A. Statement of the Problem

Several problems and challenges have been frequently faced by students during their e-learning activities. Since e-learning lacks face-to-face interaction between learner and instructor, students are obliged to have more independence and responsibility towards their own process of learning [12]. This may compel students to change their ways of thinking, behaviour and habits to succeed in their studies. However, as reported in [13] there are students who are showing disinterest in technology utilization. These students are more comfortable with the traditional learning mode (i.e., face-to-face interaction in physical classrooms). Moreover, reported the lack of knowledge, skills, training as well as the negative attitudes towards technology usage as the primary factors of students' resistance towards e-Learning usage in their education. Meanwhile, [14] emphasised the need for students to cultivate skills and confidence in e-Learning usage all through their education. In particular, there are few studies that have aimed to build a theoretical framework to examine e-learning adoption among learners. Thus, there is a need for a theoretical framework that can be used to solve many problems associated with e-learning adoption [15].

As generator of standard and reference point, benchmark is generally described as the criterion employed in measuring, scoring or judging something. As reported in [16] [8] for e-learning, internationally established benchmarking is available. Still, in the context of Oman, benchmarking of e-learning is still far from perfect. However, among the benchmarking initiatives include benchmarking of virtual campuses in Europe and Chiron. This benchmarking is linked to the project on state-of-the-art technological solutions for ubiquitous learning. This study attempts to employ a set of indicators to benchmark e-learning in order to facilitate the examination of e-learning adoption in higher education institutions in Oman, and in this study in particular, in SUC.

In Oman, various programmes have been initiated by the government in promoting e-learning adoption. Nonetheless, only a handful of works were devoted to the characterization of e-learning adoption and the scrutiny of e-learning adoption behaviour in this country. In the context of developing countries, Oman included e-learning applications, particularly with respect to the adopted applications type and their usage denoting their patterns of adoption patterns need more exploration. Additionally, empirical studies presenting an innovative model of e-learning Success grounded upon IS success and DOI in addition to the effects of e-learning in both developing and developed countries, creating a gap, which this study will attempt to fill.

B. Significance of the Study

Studies that explore the use of e-Learning among students at Omani universities, as well as the skills of students towards e-learning application in their education are still too few. Accordingly, the significance of this study is factored by several reasons. Firstly, the accessible information to demonstrate e-learning adoption in accordance with the benchmarking set selected is still limited. In addition, there is prospective in taking upon benchmarking for the improvement of the content and delivery development, change management and IT planning for uptake of e-learning. Also, this study offers universities in Oman information pertaining to use of e-learning usage, which will motivate additional works on students' knowledge regarding e-learning and their skills toward this tool in learning and teaching arena. Besides that, from this study, the presidents of universities could be imparted with future plans regarding their needs for computers and new comprehension and attitudes toward e-learning. Ultimately, this study presents a new model of e-learning Success following the IS Success Model and the Roger's DOI for the purpose of assisting education scholars in the application of e-learning in their field which leads to more development in e-learning utilization in higher education [8].

It is crucial to have clearer comprehension of the factors that positively or negatively impact the adoption of e-learning among students in order to assure success of the e-learning implementation itself [17]. For this reason, it is important that this study is conducted, in Oman particularly.

C. Scope of the Study

E-learning comprises computer technology or ICT usage in giving instruction to students and in allowing them to complete their credit courses online. E-learning comes in countless of modes including video-conferencing, video, television, blended learning, just to name a few. In this study, e-learning adoption scope covers the application of web-enabled LMS for the purpose of improving the learning quality and flexibility to all undergraduates. In the context of SUR, its learning portal is known Moodle and all students of SUC can have access to it. Flexible and qualitative learning is offered by these learning portal applications to all students but for the purpose of this study, only undergraduate students will be the focal point.

II. Literature Review

A. Definitions of Electronic-Learning

Definitions for e-learning are various. Also, terms used in delineating e-learning vary, including online education, Learning Management System (LMS), web-based learning online learning, distance education, and distance learning [18, 19] but generally, e-learning is viewed as the application of information and communication technology (ICT) in education, computer support instruction, online education or computer-aided domains for the purpose of improving education [20, 21].

As described in [22] and [23] the modes of e-learning generally falls within two categories namely Asynchronous e-learning and Synchronous e-learning. Asynchronous e-learning generally includes the use of media including e-mail and discussion boards, supports work interactions between learners and teachers even without the online presence of participants, whereas Synchronous e-learning usually includes the application of media including Videoconferencing and chat, which could support e-learners in establishing the communities of learning. According to learners and teachers, synchronous e-learning is perceived to be more social, and as it allows real-time questioning and answering session, frustration can be prevented. Some institutions use both modes of e-learning (Asynchronous e-learning and Synchronous e-learning) particularly those offering courses of e-learning to their remote learners [22, 24].

In [25] e-learning is viewed as a mode of learning that allows the fastest possible information delivery to learner. In [26], e-learning is described as a learning mode that is supported by ICT and this learning mode enhances teaching and learning in terms of quality. Author in [27] views e-Learning as the effective addition to supplement the conventional teaching and learning. As perceived by [28] e-Learning is not bound by place; it can happen anywhere such as university halls, home, and work, just to name a few.

The term e-learning is a general one and it can describe the application of all electronic means in education domain. This learning mode was in fact dubbed as internet enabled learning in [29]. Somehow, considering that the focal point of this study is students, e-learning is therefore referred as the application of electronic media for the delivery of adaptable learning. In this context, e-learning will encompass access to, downloading and utilization of web, CD or computer learning resources either in classroom or at home. In addition, access to and partaking in course activities including group discourses and evaluation activities are also included. Furthermore, in the context of this study, e-learning is interchangeably referred as learning management system (LMS), and also as SUC Moodle.

B. Theoretical and Models of Technology Acceptance

Adoption and Diffusion of Innovation (DOI) theory was first introduced in the work of [30]. The purpose of this theory was to delineate the diffusion of innovation and adoption of novel technology. Accordingly, the term diffusion was defined by the author as the process whereby an innovation is conveyed among members of a social system via specified channels over time. An innovation comprises a fresh idea, object, or practice as perceived by the related adoption unit [31], and for an individual, the newness of the innovation, as he perceives, will dictate how he reacts towards it.

The attributes impacting the innovation adoption and diffusion namely relative advantage, compatibility, complexity, trialability and observability were also highlighted in [30]. In this regard, diffusion of innovations model by Rogers is deemed the most essential adoption and adaption model due to its appropriateness as a research framework while also the most commonly tested model among [32, 33, 34, 2, 15].

As e-learning is an innovative learning mode, innovation diffusion theory is appropriate for e-learning adoption researches [35, 4, 2]. The theory is also fitting in examining how the beliefs of potential adopters regarding the innovation characteristics affect the adoption of e-learning. In this regard, [36] mentioned [30] model as among the most significant models in delineating the characteristics of innovation.

Among the most commonly cited models of IS success is that of DeLone & McLean [37, 38]. This implies the capacity of a systematic blend of specific measures from categories of IS success in generating an all-inclusive instrument of measurement. The DeLone and McLean model comprises 6 dimensions of IS success as follows: (1) system quality, (2) information quality, (3) use, (4) user satisfaction, (5) individual impact and (6) organizational impact.

In comprehending the success of IS, the use of the IS greatly contributes in two ways: offering a system for the categorization of variety of measures of IS success as reported the literature, and the recommendation of a model that illustrates the temporal and causal interdependencies between the categories [27, 15]. Several empirical works have been carried out since early 90s regarding the multidimensional connexions among IS success measures, for instance, [39, 2, 38, 38] for instance, part of the model was tested and for the purpose, the authors employed structural equation model (SEM). Such study has expanded the model via the inclusion of new dimensions including intention to use, service quality, and Net benefits (as substitutes to the construct of organisation impact) [40]. Approximately 38% of research paper have used D&M IS model [41, 42]. A total of 300 articles were used this model in the range of 1992 to 2003 [34, 43]. In between 1992 to 2014 the total of 3500 articles have used these models. This demonstrates how strong these models are, in information systems and e-learning systems and updated model in

particular which is DeLone and McLean. The aim of this study is to integrate Rogers DOI with IS model for examining e-learning adoption similar to the studies applied by [43, 15]. One theoretical model for evaluating e-learning adoption successfully is not sufficient [44]. Due to, Rogers DOI and IS model have emerged and each has been the basis for studies examining systems adoption. These two models streams have numerous parallels as suggested by [45, 39, 15] and other researchers.

Briefly stated, this study primarily attempts to characterise and measure the adoption of e-learning amongst learner, and determine the factors impacting these learners' adoption of e-learning. Notably, this study pioneers the development of a success model for e-Learning grounded upon DOI theory by Rogers and the IS success model proposed by Delone and McLean.

C. Technology Factors

Innovations carry attributes affecting the decision to adopt, implement and utilize the innovations [30]. Accordingly, [30] presented five characteristics of innovation that affect the attitude of individuals in the process of adoption, namely: relative advantage, compatibility, complexity, trialability, and observability. Each is elaborated as follows:

Relative advantage relates to the level to which an innovation is viewed to be superior as compared to the notion it succeeds. Compatibility relates to the level to which an innovation is viewed to have consistency with the values at present time and experiences in the past. Meanwhile, complexity relates to the level to which an innovation is viewed to be challenging in comprehension and utilization, whereas trialability relates to the level to which an innovation may be trialled. As for observability, it is associated with the level to which the outcomes of an innovation are noticeable to others.

In analysing innovation characteristics proposed by Rogers across a vast range of types of innovation, [46] reported compatibility, relative advantage, and complexity as having the significant linkage to adoption. Accordingly, this study will employ 3 innovations characteristics of e-learning adoption among learners, namely relative advantages, compatibility and complexity (e.g., [47, 46]) with the aim of integrating the innovation characteristics that Rogers had proposed namely compatibility, relative advantage, and complexity, alongside the Delone and McLean IS success model which comprises information quality, system quality and service quality, in order to evaluate the adoption of e-learning. System quality relates to the perceived ease of use of system and it measures by adaptability, availability, reliability, response time and usability [40]. The significance of system quality on adoption intention has been highlighted in [43] in their discussion on Internet usage. Information quality relates to the degree to which users are confident that the information is timely, complete, accurate, and has relevance [48]. The service quality of the support received by system users is provided by the IS department and IT support personnel, and it encompasses the following: responsiveness, accuracy, reliability, technical competence, and empathy of the personnel staff. Indeed, information quality demonstrates the content attributes that a message carries. System quality, service quality, and information quality that impact the adoption and usage of e-learning among users must be understood in order to achieve better e-learning adoption and implementation [49, 50].

Both the innovation characteristics by Rogers and Delone and McLean IS success model underpin this study's investigation of systems adoption. The past works on both appear to be comparable to one another [42, 51]. For instance, DOI theory by Rogers stipulates that the effort made by an organization toward the diffusion of suitable IT inside a user community is comparable to Delone and McLean IS success model [52]. Accordingly, there are 6 factors proposed in the measurement of e-learning adoption among learners as follows: compatibility, relative advantage, complexity, information quality, system quality and service quality. The ensuing section will present the steps involved in the Delone and McLean IS success model.

D. The Outcome of e-Learning on Students

Author in [40] added the dimension of system impact or net benefit in examining the result of a given Information System evaluation. Pertinently, [53] proposed moving in reverse direction and assembling all measures of impact into a class of singular impact or benefit which they refer as "net benefits" (NB). As presumed in past work, system impact or NB is appropriate in all settings of education. NB denotes the manner in which the system can bring benefit to individual user, a group of users, an organization or the industry in its entirety. NB has linkage to job performance of users and also to the change in work activity and enhanced productivity. Hence, evaluation can be made to Net Benefits of individuals via job effects, efficiency, error reduction, decision quality, as well as effectiveness [40, 13]. Meanwhile, in the evaluation of learning management system (LMS), net benefits highlight the balance of positive and negative effects on the behavior of users, and this involves system developers, instructor and learner, administrator, or all of the stakeholders of distance learning.

III. Methodology

This quantitative study employs survey questionnaires which were manually distributed to respondents. For this study, the determined population encompasses degree students enrolled in SUC. Accordingly, a sample frame which encompasses a list of population elements from which a sample can be drawn was attained from SUC Students. Meanwhile, the identification of respondents was performed using the approach of random sampling. This study received assistance from the representatives of several student residential halls in identifying the respondents. Respondents were given two weeks to complete and return the questionnaires.

A. Questionnaire Design

There are three major components contained within this study's questionnaire. These include items for eliciting the demographic information of the respondents as well as those that provide indications in benchmarking the uptake of e-learning. For e-learning Benchmarking adoption, the respective items were obtained from SUC E-learning Moodle. There are two menus: Main Menu and Course Menu. Through adoption

and usage, the framework measures e-learning adoption, and the measures follow four-point ordinal measures. The measurement scale which denotes the respondent's volume and sophistication of usage ranges from 'Not Using' to 'Use all the time.' These measures present the benchmarking for the adoption of LMS. The final section is to identify the factors impacting the e-learning adoption benchmarking while also measuring the effect of learner's e-learning adoption.

IV. Findings

A. e-Learning Adoption and Identifying Adoption Groups

The application of the framework in characterizing and measuring SUC students' e-learning adoption is discussed in this section. In this study, the respondents were asked to identify the types of e-learning applications that they were using and for each application, they were using, they were asked to determine the extent of usage. The adoption level provided denotes the present state of e-learning applications used by respondents, in this context, the students, and it also addresses the questions regarding the usage of these applications among students.

B. Level of Adoption

The adoption level describes the present state of e-learning applications that the students were adopting and it presents an initial depiction regarding the types of applications that Omani students were using. Accordingly, the e-learning applications distributions of these students can be viewed in Fig. 1. As shown by the figure, online registration appears to be the most popularly adopted among students at 96.7%, followed by online grades at 96.2. Meanwhile, online save documents were adopted most by 90.7 % of learners, while attaching was adopted by 86.1% of learners, whereas 78.1% learners have adopted online search. With respect to online courses material, it was adopted by 71.6% of learners. Assessment and Mic.word were adopted by about half (50.2%) of the learners. The use of Email, online library, excel, powerpoint, PDF and electronic journals was in the range of 44% to 3.8% of the respondents. Four applications were not used as follows: chat room, video or audio, discussion group, and testing.

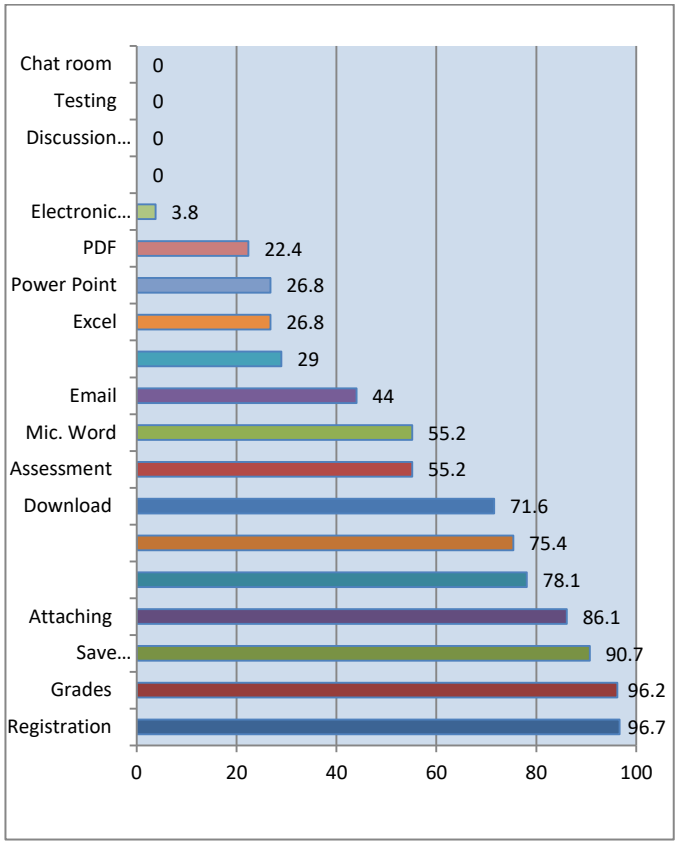


Fig. 1. Level of Adoption.

C. Extent of Usage

The use extent illustrates on the present situation of e-learning application among learners. The four categories (not using, used sometime, used most of the time, and used all the time) were used to determine the usage level of adoption of e-learning [8, 2]. It also addresses the questions regarding the usage of these applications among students.

As Fig. 2 is showing, none of the students used video or audio, discussion group, testing, or chat room. On the other hand, 37.7% of students who was adopted assessment and Mic. Word stated that they used the applications sometime, whereas 12% of them stated that they used both applications most of the time, and just 5% stated that both applications were used all the time. Less than half (45%) of the students had never used both applications. Registration and grades were used 'sometime' by 27.8% of students, and were used 'most of the time' by 38% of the students. These applications were used 'all the time' by 30% of the students, and only 3.2% of the students stated that they never used them.

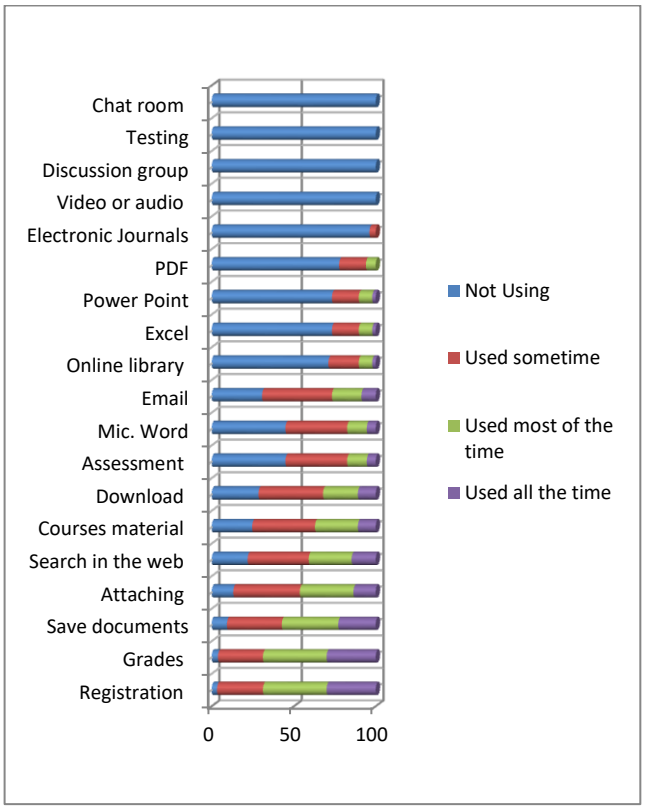


Fig. 2. Extent of usage.

Electronic journals were used 'most of the time' by 3.8% of the students while the majority of student (96.1%) stated non-usage of this application. Online library was not used by a large majority of students at 71%, while only 2.1% of the students stated that they used this application all the time, whereas approximately 8.1% stated that they used this most of the time. Further, online library was used 'sometimes' by 18.1% of the students. About 30.6% of students stated that they had never adopted the use of email while 18% expressed that they used it most of the time. Meanwhile, 8% used email all the time and 42.6% used this application sometimes.

With respect to the use of Search in the web, roughly 21.8% of students stated non-usage of it, while 26% stated that they used this application most of the time, whereas 14.7% indicated usage of the application all the time. About 37.1% of students stated that they used this application sometimes. In terms of the application of Save documents, 34.4% of students stated usage of 'most of the time', while 22.9% stated usage of 'all the time' whereas 33.3% stated usage of 'sometimes.' Those that never used this application accounted for 9.2%. The application of Attaching was used 'most of the time' by 32.7% of the students, 'all time' by 13.6% of students, and 'sometimes' by 40.4% of students. Attaching application was never used by 13.1% of the students.

The applications of Excel and PowerPoint were never used by 73% of the students, while the remaining 27% is represented by usage of 'sometimes,' 'most of the time' and 'all the time' in combination. Online Courses material was used 'sometimes' by 38.2% of students, 'most of the time' by 26.2% of students, and 'all the time' by 10.9% of students. Those who never used this application accounted for 24.5%. Download was used 'most of the time' by 21.3% of students, 'all the time' by 10.9% of students, and 'sometimes' by 39.3% of students. On the other hand, 28.4% stated non-usage of this application. Non-usage of PDF was expressed by most students (77.5%), while 16.3% stated that they used this application sometimes, and PDF was used most of the time by 6% of students.

The outcomes indicate that the applications of Assessment, Attaching, Chat room, Courses material, Discussion group, Download, Electronic Journals, Email, Excel, Grades, Mic. Word, Online library, PDF, PowerPoint, Registration, Save documents, Search in the web, Testing, and Video or audio were used 'sometimes' by the majority of the students. Also, among these applications, Registration and Grades were used the most while live audio and live video showed the lowest level of usage.

D. Identifying Adoption Groups

The approach used in this study demonstrates the effectiveness in identifying the patterns of adoption present in the sample, and accordingly, students were grouped in accordance with the patterns of adoption. Consequently, Table I highlights the cluster members and the finalised three adoption groups obtained from the matrix table. These groups encompass low-adopters represented by 63.9% of students, moderate-adopters represented by 36% of students, and high-adopters represented by 10.9% of students.

TABLE. I. DERIVATION OF FINAL GROUPS AND CLUSTER MEMBERS

Group No	Preliminary Group	Final Group
1	76	117 (Low-adopters)
2	41	
3	37	66 (Moderate-adopters)
4	29	
5	20	20 (High Moderate - adopters)
Total	183	183

E. Research Conceptual Framework Evaluation

Table II presents the summarized results obtained from the multinomial logistics regressions. This table highlights 6 predictor variables, and the variables of relative advantage, complexity, system quality, information quality and service quality appear to have significant linkage to the adoption of e-learning.

As shown by the results, there are three groups of distinctive adopters namely low, moderate and high adopters. Also, the dimensions of relative advantage, complexity, system quality, information quality and service quality positively affect the probability of e-learning adoption in the context of low-adopters vs. high adopters and moderate-adopters vs. low-adopters as opposed to moderate-adopters vs. high adopters. In addition, the dimensions of relative advantage, complexity, system quality, information quality and service quality have significant linkage to e-learning adoption amongst students. These outcomes are in agreement with those of past works particularly with respect to the significance of the dimensions of relative advantage, complexity, system quality, information quality and service quality on students' adoption of e-learning.

TABLE. II. MULTINOMIAL LOGISTIC REGRESSION

	Low-adopters vs. High-adopters		Moderate-adopters vs. High-adopters		Moderate-adopters vs. Low-adopters	
	Coff (β)	Wald p-value	Coff (β)	Wald p-value	Coff (β)	Wald p-value
Relative Advantage	.601	.011	5.884	.281	3.710	.413
Compatibility	.291	.188	.236	.656	.364	.169
Complexity	.743	.015	.092	.525	.477	.000
System Quality	.339	.041	.125	.562	.171	.416
Information Quality	.688	.021	. 482	.482	.134	.033
Service Quality	.013	.969	.528	.236	.161	.005

F. Outcome of E-Learning on Students

The analysis of correlation was carried out in this paper to ascertain if the degree of e-learning adoption and the impact measure are related. As can be seen in Table III, net benefit is significantly correlated with the adoption of e-learning, with the positive correlation of .196**.

TABLE. III. E-LEARNING ADOPTION AND NET BENEFITS MEASURES

Impact / EL Adoption	Job Performance	
	Sig.	R
Adoption	.196**	.000

V. CONCLUSION

This research mainly contributes to the theoretical knowledge particularly with respect to innovation adoption and IS success theory. The research proposes a new model that describes a concrete set of factors that higher education policy makers and universities' managers have to concern about to facilitate learners to adopt and use e-learning in their teaching and learning process. The importance of the theoretical framework is to help the decision and policy makers to determine which factors need support and which need treatment to encourage the learners to adopt e-learning. This paper presented an e-learning adoption framework which is an alternative framework which elucidates innovation adoption. This involves two dimensions namely the level of the adopted e-learning application and the usage degree of each e-learning application. The scrutiny of the usage degree generated three adoption groups which also have linkage to factors relating to technology and e-learning outcome. At the individual-level of innovation adoption, the proposed framework presents a more expansive comprehension and more universal sense for innovation adoption. For this reason, the framework is applicable in the adoption of different IS/IT innovations or the e-learning system adoption within diverse industries. In the context of IS innovation adoption, the framework proposed in this study contributes in the adoption level as it presents information regarding the types adopted application, as well as the usage degree of each e-learning application as opposed to the utilization of a single perspective (adopted or not adopted). In addition, this framework improves DOI theory through the inclusion of factors of system quality, information quality and service quality alongside the outcome of e-learning contexts to the theory. Within the domain of innovation diffusion, a conceptual framework was theoretically established and empirically assessed using factors of technology and the outcome of e-learning which allows the evaluation of the use and outcome of e-learning among students. On the other hand, this study is a valuable addition to the Arab world as it presents information regarding e-learning adoption; it also offers fresh information for other comparable developing countries. Perusing this study, researchers could also carry out additional works on e-learning in the context of higher education students.

In terms of limitations of the study, the study had come across are worthy of mention as well. In fact, this study's strong point lies in the recognition of its limitations, as these limitations can pave new way for future works while also providing clarification to the theoretical implications. The merit of this study lies in its scrutiny on the quintessence of e-learning adoption among Omani students, and considering that e-learning implementation in Omani academia is still rudimentary, there should be more studies done in this context. Hence, for future works on e-learning adoption, the use of other groups of students or participants is recommended as this may generate new significant findings. Apart from the use of other groups of students or participants, the scrutiny of other factors is also recommended in order to solve the issues of responsiveness and personalization.

ACKNOWLEDGMENT

The authors wish to acknowledge Irbid National University support, that without which this report would not have been accomplished.

REFERENCES

[1] T. Anderson and T. Elloumi, "The Theory and Practice of Online Learning," Edmonton: AU Press, 2011.

[2] S. I. Alzoubi and A. I. Alzoubi, "Investigating the Extent of Learning Management System Adoption in a Jordanian," International Journal of Business and Social Science, vol. 8, no. 1, pp. 208-216, 2017.

[3] OECD, " (2005) E-learning in tertiary education. [Online]..pdf (Accessed 21st April 2011).," 2005. [Online]. Available: Available at http://www.cumex.org.mx/archivos/ACERVO/ElearningPolicybriefenglish. [Accessed 21 April 2018].

[4] A. Al-adwan and J. Smedley, "Implementing e-learning in the Jordanian Higher Education System: Factors affecting impact.," International Journal of Education and Development using Information and Communication Technology, vol. 1, no. 8, pp. 121-135, 2012.

[5] S. Alexander, "E-learning Developments and Experiences.," Education and Training, vol. 43, no. 4, pp. 240-248, 2001.

[6] M. Y. Cheng, "Antecedents and consequences of e-learning acceptance," Inf. Syst. J, vol. 21, no. 1, p. 269–299, 2011.

[7] J. L. Chen, "The effects of education compatibility and technological expectancy on e-learning acceptance," Comput. Educ, vol. 57, no. 1, 2011.

[8] S. L. Thi, M. N. Sharif, A. Darus and T. S. Lai, "Evaluating E-Learning Uptake in a Malaysian Higher Education Institution," International Conference on Information Systems for Business Competitiveness, pp. 1-8, 2015.

[9] R. Boateng, A. S. Mbrokoh, B. Lovia, P. K. Senyo and E. Ansong, "Determinants of e-learning adoption among students of developing countries," The International Journal of Information and Learning, vol. 44, no. 4, pp. 248 - 262, 2016.

[10] W. A. A. Rakhyoot, "Institutional and individual barriers of e-learning adoption in higher education in Oman : academics' perspectives," ETHOS, vol. 8, 2017.

[11] N. Sarachandran and R. Patil, "A Study on the Impact of Learning Management Systems on Students of a University College in Sultanate of Oman," International Journal of Computer Science Issues, vol. 9, no. 1, 2012.

[12] M. Hatcher and M. Yen, "Using theory of constraints in E-learning for overcoming internalexternal, cultural, and international constraints," Journal of the Academy of Business and Economics, vol. 5, no. 3, pp. 611-617, 2005.

[13] Ishtaiwa, "Factors influencing faculty participation in e-learning: the case of Jordan," AGFDS, vol. 2, no. 2, pp. 12-25, 2006.

[14] A. Macpherson, M. Elliot, I. Harris and G. Homan, "E-learning: Reflections and evaluation of corporate programs," Human Resource Development International, vol. 7, no. 3, pp. 295-313, 2004.

[15] I. S. Alzoubi and I. A. Alzoubi, "An Investigation of Factors affecting E-Marketing Customers' Behavioral intention to use Telecommunication industry in Jordan; Vol. 11, No. 3, Published by International Journal of Marketing Studies.," 2019.

[16] E. Ossiannilsson, "Benchmarking E-Learning inHigher Education," Lessons Learned from International Project. Unpublished Ph. D. Dissertation, University of Oulu, Finland, 2012.

[17] R. G. Saadé and D. Kira, "Computer anxiety in e-learning: the effect of computer self-efficacy," Journal of Information Technology Education, pp. 177-191, 2009.

[18] J. Hayen, J. Cappel and L. Roger, "Evaluating E-Learning: A Case Study," Ournal of Computer Information Systems, vol. 44, pp. 49-56, 2004.

[19] L. A. Halawi, S. Pires and Mccarthy, "An Evaluation of E-Learning on the Basisof Bloom's Taxonomy: An Exploratory Study," Education for Business, vol. 84, no. 6, pp. 374-380, 2009.

[20] S. M. Mutula, "E-Learning Initiative at the University of Botswana: Challenges and Opportunities," Campus-Wide Information Systems, pp. 99-109, 2002.

[21] C. Fallon and S. Brown, "E-learning Standards: A Guide to Purchasing," Developing, and Deploying Standards-Conformant E-Learning Press.: Delray Beach, , 2003.

[22] Moore, Deane and Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," Internet and Higher Education, no. 14, p. 129–135, 2011.

[23] I. A. Al-Zoubi and Mohamad, "An Integrated Success Adoption Model for Examining E-Learning Among Adult Workers in Jordan. 978-1-4799-0059-6/13/31.00 ©2014 IEEE".

[24] K. Behling, J. Orczyk and J. Jenkins, "Live distance learning delivery of Master of Science courses in building construction management," 37th ASEE/IEEE Frontiers in Education Conference, 2007.

[25] B. H. Khan, "Managing e-learning: Design, delivery, implementation and evaluation," pp. Hershey, PA: Information Science Publishing, 2005.

[26] Begičević, Divjak and Hunjak, "Imaginative acquisition of knowledge - strategic planning of e-learning," ITI, pp. Cavtat, Croatia, 2006.

[27] M. McPherson and M. B. Nunes, "Organizational Issues for ELearning: Critical Success Factors as Identified by HE Practitioners," International Journal of Educational Management, vol. 20, no. 1, p. 542 – 558, 2002.

[28] P. Hadjiathanasiou, "The e-Learning readiness of Cyprus primary teachers ahead of dias system integration into Cyprus schools," European Journal of Open, Distance and E-Learning, pp. 34-51, 2009.

[29] A. Gunasekaran, R. D. McNeil and D. Shaul, "E-learning research and application.," Industrial and Commercial Training, vol. 32, no. 3, pp. 44-53, 2002.

[30] E. M. Rogers, Diffusion of innovations, New York: The Free Press. (Original publication 1962) , 1995.

[31] G. Zaltman, R. Duncan and J. Holbek, Innovations and organizations, vol. 1973, New York: Wiley , 1973.

[32] J. F. Enge, R. D. Blackwell and P. W. Miniard, Consumer Behaviour, 8th Ed, Chicago: Dryden Press, 1995.

[33] G. Salter and S. Hansen, "Facilitating web-based staff development in higher education In L Richardson & J. Kidstone (Eds.)," ASET/HERDSA 2000 Flexible Learning for a Flexible Society, 2-5 July 2000, pp. 612-617, 2001.

[34] A. I. Al Zoubib and Z. Jali, "An integrated success adoption model for examining E-learning among adult workers in Jordan," in International Conference on Computer and Information Sciences (ICCOINS), 2014.

[35] Y. a. Duan, "A study on e-learning takeup intention from an innovation adoption perspective: A case in Chin," Computers & Education, vol. 55, pp. 237-246., 2010.

[36] G. C. Moore and I. Benbasat, "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," Information System Research, pp. 192-222, 1991.

[37] D. X. Ding, P. H. Hu and O. R. Sheng, "e-SELFQUAL: A scale for measuring online self-service quality," Journal of Business Research, Article in Press, 2010.

[38] H. H. Chang, Y. H. Wang and W. Y. Yang, "The impact of e-service quality, customer satisfaction and loyalty on e-mrketing: moderating effect of perceived value," Total Quality Management, vol. 20, no. 4, pp. 423-443, 2009.

[39] S. Alaa, "Digital storytelling: a meaningful technology-integrated approach for engaged student learning," Education Tech Research Dev, p. 487–506, 2008.

[40] W. H. DeLone and E. R. McLean, "The Delone and Mclean model of information systems success: A ten-year update," Journal of Management Information Systems, vol. 19, no. 4, p. 9–30, 2003.

[41] E. L. Lwoga, "Critical success factors for adoption of web-based learning management systems in Tanzania", International Journal of Education and Development using Information and Communication Technology, Vol.10, No.1,2014, pp.4-21."

[42] A. Ozkan and R. Koseler, "Multi-dimensional Students' Evaluation of E-learning Systems in the Higher Education Context: An Empirical Investigation", Computers & Education, Vol.53, No.4, pp.1285-1296.," 2009.

[43] C. J. Lin and C. Lu, "Towards an understanding of the behavioural intention to use a web site," International Journal of Information Management, vol. 20, pp. 197-208, 2000.

[44] S. H. Liu, H. L. Liao and J. A. Pratt, "Impact of Media Richness and Flow on E-Learning Technology Acceptance. Computers & Education, 52(3), 599-607," 2009.

[45] M. B. Prescott and S. A. Conger, " Information technology innovations: a classification by IT locus of impact and research approach," The Data Base for Advances in Information Systems, p. 20–41, 1995.

[46] L. G. Tornatzky and R. G. Klein, "Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings",," in IEEE Transactions on Engineering Management EM-29: , 1982.

[47] B. Chokri, "Factors influencing the adoption of the e-learning technology in teaching and learning by students of a University class," European Scientific Journal, vol. 28, no. 8, p. 1165–1190, 2012.

[48] Y. W. Lee, D. M. Strong, B. K. Kahn and Wan, "AIMQ: a methodology for information quality assessment," Information & Management, p. 133–146, 2002.

[49] S. Alkhalaf, S. Drew, R. AlGhamdi and O. Alfarraj, "E-Learning system on higher education institutions in KSA: attitudes and perceptions of faculty members," Procedia - Social and Behavioral Sciences, p. 1199 – 1205, 2012.

[50] T. Ramayah and J. C. Lee, "System characteristics, satisfaction and e-learning usage: A structural equation model (SEM)," TOJET: The Turkish Online Journal of Educational Technology, vol. 11, no. 2, pp. 26-28, 2012.

[51] R. G. Fichman, "Going Beyond the Dominant Paradigm for IT Innovation Research: Emerging Concepts and Methods," Journal of the Association for Information Systems, pp. 314-355, 2004.

[52] R. B. Cooper and R. W. Zmud, "Information Technology Implementation Research: A Technological Diffusion Approach," Management Science, vol. 36, no. 2, p. 123–139, 1990.

[53] W. H. DeLone and E. R. McLean, "Information system success: the quest for the dependent variable," Information System Research, vol. 3, no. 1, pp. 60-95, 1992.

Development of Warning Device in Risk Situations for Children with Hearing Impairment at Low Cost

Kevin Rodriguez-Villarreal[1], Zumaeta-Mori Jhon[2], Alva Mantari Alicia[3], Roman-Gonzalez Avid[4]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—Hearing impairment is the partial or total loss of hearing. There are approximately 34 million hearing impaired children in the world. The equipment used as a means of communication to improve interaction with society is very expensive, so in this study was built an electronic device with the ability to recognize some words configured as an emergency message. This equipment will be used as a basic means of communication for hearing impaired children at a very low price. The equipment consists of a transmitter and a receiver, which communicate over Wi-Fi 802.11 at distances between 0m and 95m using low-power electronic devices and recent technology such as WeMos D1 mini Lite cards. This device was tested on approximately 20 people caring for hearing impaired children, obtaining a measure of approval of approximately 74%. This is the first step in research that we plan to continue to reduce health gaps and improve communication for children with disabilities. Our group works for preventive health, reducing health gaps in the most vulnerable population.**

Keywords*—Hearing disability; device; low cost; algorithm; communication

I. Introduction

In the world, there are approximately 466 million people with a hearing-impairment (in adults over 40dB and in children at 30dB) of which 34 million children, according to the World Health Organization [1]. Hearing impairment in children is that difficulty or inability to listen, due to partial or total loss of hearing ability, which can occur in one or both ears. Some of the causes of hearing impairment in children can be trauma, otitis, meningitis, excessive exposure to noise or strong medications that could cause some impact on the auditory nerve [2]. Hearing impairment in children is the third most frequent disability in Peru [3], and the index increases more and more, because a diagnosis is not made at an early age, so this condition can be aggravated in all infants who they present it [4]. In Peru 1 to 6 Peruvians are born every day with severe to profound hearing loss, there are no centers with a good infrastructure to help them. Despite the efforts made by the National Program for Early Management of Hearing Loss, it is not enough to care for all children [5]. Hearing-impairment in children in Peru has only been diagnosed in 30%. Hearing-impairment in children represents a severe public health problem, which the State should have as a priority, since progress can be dangerous throughout the country, despite having the hearing loss program that has granted about 500 implants cochlear, much more is needed to provide optimal care to all children who require it. This situation has been caused by the lack of medical instrumentation, which allows this condition to be diagnosed, in addition to lacking an institution with the necessary spaces and conditions to perform patient surgeries. It is estimated that only 30% of children with sensory hearing loss on both sides have been diagnosed and received a surgical intervention with a cochlear implant, through the aforementioned program [6]. About 300 children with hearing disabilities are born each year, of which only about 90 patients would be benefiting, while another 210 are not receiving treatment or diagnosis [7]. The solution for some of them is the cochlear implant, also known as the bionic ear, is that device that is implanted through surgery, helping to improve hearing- impairment in children, helping them overcome these problems originating in the inner ear or cochlea. This procedure is recommended for those children who have sensorineural hearing loss, allowing them to recover their hearing by 85%, preventing the problem from getting worse and becoming deaf, but this is a very expensive procedure. After undergoing this operation, children should receive language therapies to complete their communication process and be able to acquire learning in an optimal way, since hearing-impairment often causes difficulties in the acquisition of reading skills, because Communication is run over by this condition. This research seeks to provide an alternative communication and alert for children in a state of emergency through a device for non-visual communication. That is, the child can interpret the vibrations or visual stimuli activated by words of the person in charge of his care. The equipment works in such a way that it converts the message sent by the tutor into vibrations produced by the motors or into light emitted by the LEDs.

Translated with www.DeepL.com/Translator (free version).

II. Population and Sample

A. Sample

For the development of the communication device, one of the most important steps was the choice of the most relevant keywords. Through a survey was built in which 44 people over the age of 18 participated, who chose 7 words from a group of 19 words taken from emergency manuals for children in general. The result of this survey can be seen in Table I, where you can see that the selected words are:

- Help (56.8%)
- Stay with me (52.3%)

- Stay calm (52.3%)
- Emergency (43.2%)
- Stop (43.2%)
- Watch out (40.9%)
- SOS (36.4%)

TABLE. I. TABLE OF PHRASES USED IN EMERGENCY SITUATIONS

Consulted Phrases			
Consulted Phrases	***Frequency***	***Relative Percentage***	***Absolute Percentage***
SOS	16.00	36.4	5.63
Emergency	19.00	43.2	6.69
Urgent	14.00	31.8	4.93
Stop	19.00	43.2	6.69
Run	12.00	27.3	4.23
Help	25.00	56.8	8.80
Attentive	13.00	29.5	4.58
Help me	14.00	31.8	4.93
Do you feel any pain?	12.00	27.3	4.23
I love you	9.00	20.5	3.17
Stay with me	23.00	52.3	8.10
Watch out	18.00	40.9	6.34
Look at me	13.00	29.5	4.58
protect yourself	13.00	29.5	4.58
get moving	3.00	6.8	1.06
Where are you?	12.00	27.3	4.23
Do you feel any discomfort?	15.00	34.1	5.28
Slowly	11.00	25.0	3.87
Stay calm	23.00	52.3	8.10

III. METHODOLOGY

The work is based on the design of two electronic devices, capable of communicating with each other over distances greater than 80m using Wi-Fi transmission. Applying low-power devices that are wireless microcontrollers with Wi-Fi communication, specifically 802.11. In this sense, the application of the device consists of sending voice commands through a microphone connected to a voice recognition module, which recognizes the pre-recorded words and send them via the WeMos D1 mini Pro transmitter device to the receiver (Fig. 1). This receiver converts the data sent by the transmitter into sensitive stimuli that are manifested through dc motors and RGB LEDs.

A. Description of the main Components

1) Wemos D1 mini Lite: This low-power and highly functional device is a new technology, so there is not much information about it. This card is controlled by an ESP8266 and a Wi-Fi chip, which allows it to transmit over distances greater than 80m (Fig. 2).

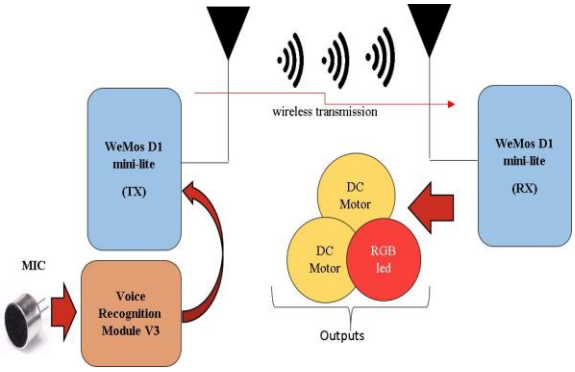


Fig. 1. Circuit Block Diagram.

Fig. 2. WeMos D1 mini Lite Development Board.

There are variants of the ESP8286 with more functionalities and other aggregates in the hardware, either with greater number of pins, greater memory capacity and greater range in your Wi-Fi antenna [8] (Details Table II).

2) Voice recognition module v3: The voice recognition module used for the recording of the words that will cause an action on the receiving device, communicates through the serial port Tx and Rx, in order to record the words to be used, in addition to this the module only recognizes the voice of the person who recorded it according to the frequency of his voice, thus preventing outsiders from manipulating him. The parameters that this module handles are those shown in Table III (Fig. 3). For the recording of the words, 5 of a list that was distributed as a survey were initially used. The module only picks up 7 voice commands at the same time, and if you want to record more they must be loaded in groups of 7, it supports a maximum of 80 commands in total with a duration of 1500ms. However, if almost 100% operation is required, the environmental conditions must be ideal, that is, without noise in the environment.

3) Micro vibrator CEBEK C-6070: The purpose of using vibrators and non-motors, is due to the application that this device has, for people with low hearing sensitivity the vibrating of the vibrator next to the ear is more perceived than a movement of a conventional DC motor, the details are those shown in Table IV. The size of the micro vibrator allowed it to be placed on both sides of the headset and thus activated with the voltage of the WeMos D1 mini Lite [9] (Fig. 4).

B. Operation of the Complete System

The development of the prototype and its understanding are detailed in the following flowchart, in which the following logic is observed, from the creation part of the algorithm to its implementation.

TABLE. II. CHARACTERISTICS OF THE WEMOS D1 MINI ESP8285

Parameter	Value
Input Voltage	3.3v
Flash memory	1MB
Communication Support	I2C
Digital Inputs	11
Wi-Fi Protocol	802.11 b/g/n/e/i
Operating band	ISM 2.4GHz.
Input Port	Micro USB
Analog Pin	A0 (1)
SPI Pins	D5,D6,D7,D8 (4)
Current	500 mAh
Compatibility	Arduino

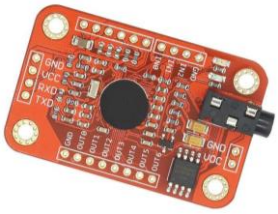


Fig. 3. Voice Recognition Module v3.

TABLE. III. CHARACTERISTICS OF THE VOICE RECOGNITION MODULE V3

Parameter	Value
Input Voltage	4.5-5.5V
Current	<40mA
Digital Interface	5V TTL level for UART interface and GPIO
Analog Interface	3.5mm mono-channel microphone connector
Recognition accuracy	99% (under ideal environment)

Fig. 4. Micro Vibrator Cebek C-6070.

TABLE. IV. CHARACTERISTICS OF THE CEBEK C-6070 MICRO VIBRATOR

Parameter	Value
Input Voltage	1.3v
Average speed	8000 rpm
Average current	75mA
Resistance	11,7 W ±10%
Approximate weight	1.8g

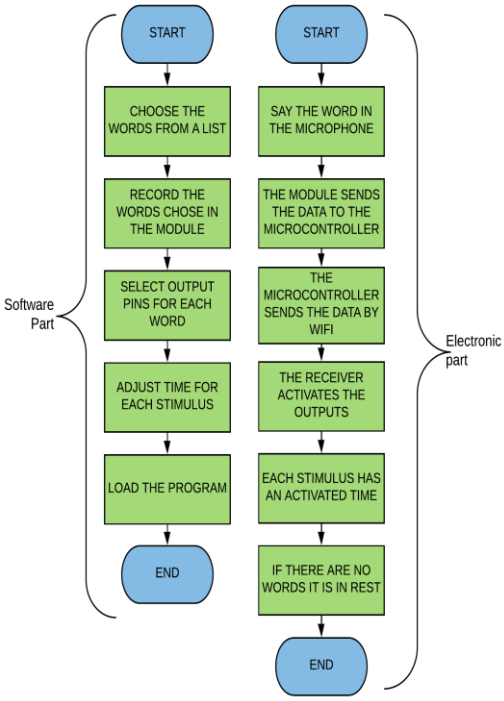


Fig. 5. Circuit Flow Diagram.

In the flowchart of Fig. 5. It is observed, the software part with the steps for the configuration of the voice recognition module through the Arduino IDE, and also the procedure of the hardware part, such as sending the message and What happens once this message is picked up by the receiver activating the different sensory devices attached to the headphones.

C. Limitations

The development of the device could be improved by including more words in the algorithm, currently we only have 7 words for identification. In addition to having been tested for usability and acceptability by the target population (parents, teachers and hearing-impaired children). At the same time, see how to improve performance and set a minimum acceptable age in the population for the application of the device.

IV. RESULTS

The development of this prototype is intended to provide an alternative to problems of hearing impairment, through applied electronics. We verify the operation of the devices created, both the transmitter and the receiver. For this, it was decided to adapt the transmitter circuit in a perforated Bakelite, along with some straps to use on the wrist as a bracelet (Fig. 6).

However, the design of the receiver was adapted to wireless headphones, since the purpose of this is to be able to provide freedom in the user's hands so that it moves with total comfort, also having it in a part that is not uncomfortable for the person who will use it.

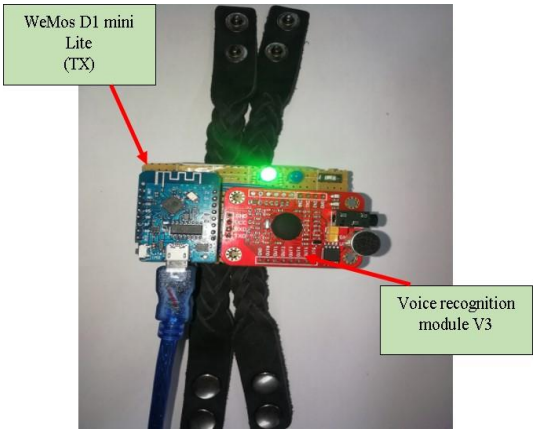


Fig. 6. Transmitter Bracelet.

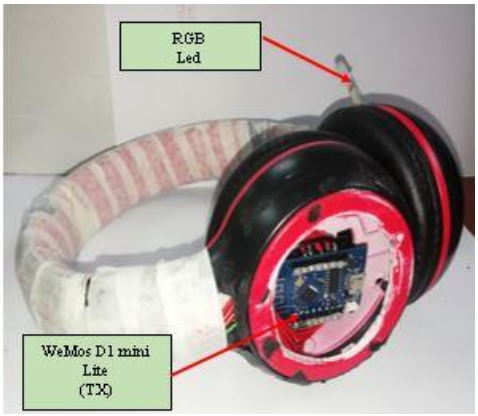


Fig. 7. Receiver Headphones.

As seen in Fig. 7, on one side of the headphones, an RGB LED was placed, which turns red, green or blue, depending on the word sent by the transmitter bracelet. The word already pre-recorded can activate a visual stimulus for a while and then turn it off or you can activate it indefinitely until the same message sent is said again. On the other hand, on both sides of the headset, there is an unbalanced DC motor on its axis, which has the function of improving its vibration, so that the stimulus feels more efficiently through the earpiece protectors.

The prototype was tested and has a range from 0 meters to 95 meters from the transmitter (bracelet used by the caregiver) with the receiver (headphones used by the child).

A small pilot was carried out in 20 carers of hearing-impaired children, obtaining a satisfaction of 7.4 out of a maximum of 10 points, it is worth mentioning that the participants did not have another device with these characteristics and their primary means of communication was sign language.

V. Conclusions

A low-cost device for alternate communication between hearing impaired people was developed at low cost (approx. $85). Within the device was included a word identification algorithm that allowed a communication addressed to people of scarce resources. The adaptation of the device for the initial pilot is based on the training of the child and the tutor. For this reason, a child with a minimum age of 9 years is needed. The reception of such a device in the families was positive due to the consequences of its use in the inclusion of the child in the community.

The development of this prototype is intended to be an alternative and is the first phase of an investigation that includes the construction of the equipment, its use and implementation in families with hearing impaired children. There is no such equipment on the market [9]. We offer the possibility of non-visual communication to these children. This work allows to open a whole branch of research related to the inclusion and the improvement of the quality of life in an economic way.

VI. Future Work

A next step in this study corresponds to the improvement of the equipment, to improve the distance of transmission of the message and at the same time to be able to store more quantity of words that allow to have a better communication. and that is not so limited. The inclusion of more people with moderate to severe hearing impairment and who do not have other options to improve communication at a low cost due to the inaccessibility of instruments for this purpose.

Acknowledgment

We want to thank the Universidad de Ciencias y Humanidades for the support in carrying out this study. We continue to work for preventive health, and to reduce health gaps throughout our country.

References

[1] K. R. White, I. Forsman, J. Eichwald, and K. Munoz, “The evolution of early hearing detection and intervention programs in the United States,” Semin. Perinatol., vol. 34, no. 2, pp. 170–179, Apr. 2010.

[2] P. M. Watkin, M. Baldwin, and G. McEnery, “Neonatal at risk screening and the identification of deafness.,” Arch. Dis. Child., vol. 66, no. 10 Spec No, pp. 1130–1135, Oct. 1991.

[3] “PERÚ - INEI: Perú: Resultados Definitivos de los Censos Nacionales 2017.” [Online]. Available: https://www.inei.gob.pe/media/MenuRecursivo/publicacionesdigitales/Est/Lib1544/. [Accessed: 06-Aug-2019].

[4] A. Erenberg, J. Lemons, C. Sia, D. Trunkel, and P. Ziring, “Newborn and infant hearing loss: detection and intervention.American Academy of Pediatrics. Task Force on Newborn and Infant Hearing, 1998- 1999,” Pediatrics, vol. 103, no. 2, pp. 527–530, Feb. 1999.

[5] J. Rado-Triveño and J. Alen-Ayca, “Evaluación de las otoemisiones acústicas en relación a los potenciales evocados auditivos de tronco cerebral en niños,” Rev. Peru. Med. Exp. Salud Pública, vol. 33, pp. 706–712, Dec. 2016.

[6] O. Hernández Fernández, L. Zamora Rodríguez, and V. Abreu García, “El implante coclear: una opción para el niño cubano con discapacidad auditiva,” Medicentro Electrónica, vol. 19, no. 3, pp. 167–170, Sep. 2015.

[7] C. Cáceres and S. Rosario, “Estudio epidemiológico del implante coclear en el Perú de 1999 al 2009,” Pontif. Univ. Católica Perú, Sep. 2018.

[8] A. K. Evans and K. Kazahaya, “Canal atresia: ‘surgery or implantable hearing devices? The expert’s question is revisited,’” Int. J. Pediatr. Otorhinolaryngol., vol. 71, no. 3, pp. 367–374, Mar. 2007.

[9] “American National Standard Methods of Measurement of Compatibility between Wireless Communications Devices and Hearing Aids,” C6319-2011Revision ANSI C6319-2007 - Redline, pp. 1–184, May 2011.

A Deep-Learning Model for Predicting and Visualizing the Risk of Road Traffic Accidents in Saudi Arabia: A Tutorial Approach

Maram Alrajhi[1], Mahmoud Kamel[2]
Information Systems Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Around the world, road traffic accidents (RTAs) cause significant concerns for decision makers and researchers on traffic safety. The diversity, rarity, and interconnectivity of historical data on factors causing car accidents point to the need for more focused studies for analyzing, predicting, and visualizing the risk of accidents over the short and long term for preventive purposes. There are many techniques and tools applied to analyze, forecast, and visualize risk. Most RTA studies have applied linear time-series methods to forecasting the risk with limited studies applying machine-learning and deep-learning techniques, especially in Saudi Arabia. Recently, many global studies have applied long short-term memory (LSTM) networks, which can be used to automatically learn the temporal dependence structures for challenging time-series forecasting problems. This paper displays a tutorial for designing a prototype of an interactive analytical tool based on a multivariate LSTM model for time-series data to predict future car accidents, fatalities, and injuries in the Kingdom of Saudi Arabia (KSA). This interactive tool visualizes the real data with the predicted values regionally in a web browser with Python. The tutorial represents the annual data of the period between 1417 (1996) and 1433 (2013), then uses the data with some contributing factors, such as population, gender, nationality, number of vehicles, and length of road, to generate the input data and predict the future values of accidents, fatalities, and injuries up to the year 1452 (2030). After that the real and predicted values are visualized regionally on an interactive map that represents the degree of risk. Finally, the paper discusses the evaluation and utilization of the proposed prototype in the future in the field of road safety.**

Keywords*—*LSTM for time-series forecasting; deep learning; RTA; data visualization; interactive map; Saudi Arabia

I. INTRODUCTION

Predicting road traffic accidents (RTAs) can improve road safety for both travelers and road-safety administrators. Nowadays, the accelerated development in data-collection techniques facilitates the availability of big datasets. Several studies have discussed the prediction of RTA risk regarding multiple influencing factors. Some have considered human factors [1], and others, vehicle factors [2]; still others have concentrated on road and environmental factors [3]. The rest have combined all these factors to predict risk [4].

Two approaches exist for estimating the time-series forecasting of RTAs. The first approach is a regression problem that forecasts the number of accidents based on the attributes of the accident dataset [5]. The second is a classification problem that predicts the severity of crashes based on the crash dataset. The autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) are two basic prediction methods that are notable for time-series prediction models based on regression [6] Most statistical models have difficulty dealing with complexity, randomness, irregularity, and the nonlinearity of real data, limiting the precision of their predictions; machine-learning models as in [7] can obtain more accurate predictions when compared with the traditional models because of repeated training iterations and learning approximation mechanisms [8]. However, in the case of time-series prediction, a lack of efficient processing of sequence dependencies between input variables is a problem in some machine-learning methods [9].

Recently, specific patterns of deep learning, especially recurrent neural networks (RNNs) [10] and long short-term memory (LSTM) algorithms [11], have been successfully applied to time-series forecasting tasks. RNNs are the most efficient method [12] as they are a kind of artificial neural network (ANN) containing connected nodes in a loop form. Looping allows the internal state of the neural network to manifest dynamic timing behavior. However, some problems, such as gradient disappearance, often occur during the training of RNNs with the increase in the length of the processing time series, especially in networks using conventional activation functions, such as sigmoid or tanh functions; this limits their prediction accuracy.

The LSTM method [13] is generally one of the most efficient approaches for dealing with time-series forecasting problems; here, a simple RNN is employed with some multithreshold gates for resolving and forgetting memory problems with long period of time. The structure of LSTM is inspired by cognitive neuroscience; some researchers have even introduced attention mechanisms into the coding–decoding framework and encoded semantics in long-term memory to improve the selection of input sequences and the information-processing capabilities of the neural multiorder network [14]. Recently, LSTM with different mechanisms has been broadly applied and performed well in diverse types of deep-learning tasks, such as image captioning [15], visual question answering [16], traffic flow [17], road accidents [18], and speech recognition [19].

For visualizing the volume of accident risk, many data visualization techniques, such as interactive maps, bar charts,

bubble charts, and combinations of more than one tool [20], aim to facilitate user understanding of information interactively in a web browser [21]. Several web-programming languages, such as Python and JavaScript, contribute to designing these interactive tools [22]. Some studies have analyzed the time-series problem of RTAs regarding different factors, even focusing on classification or regression. Previous studies have noted some machine-learning techniques that can be used for forecasting RTA data. The present study aims to predict and visualize RTA data by applying LSTM with Python.

This paper is presented as follows: Related work is described in Section II. Then, the whole process of the proposed experiment is considered in Section III. The evaluation of the experiment's results is given in Section IV. Section V discusses the utilization and improvement of the proposed prototype. Finally, the conclusion is provided in Section VI.

II. Related Work

A. Time-Series Forecasting with Long Short-Term Memory

At one time, statistical methods—especially error, trend, and seasonal ETS models and ARIMA models—performed better than other methods in terms of time-series forecasting [23]. However, traditional time-series forecasting has been affected by the advancements in the field of deep learning for time-series prediction. In recent years, there has been attention paid to RNNs and LSTM, with their deep-learning applications for many disciplines, including natural language processing, finance, and computer vision. Deep-learning methods have the capability of identifying patterns and structures of data, such as complexity and nonlinearity, in time-series predicting [24]. LSTM is used in such predictive research for time-series data in different fields. The experiment in [25] solves the location prediction problem using time-series analysis. The data are transformed into a multivariate time series, and this is predicted using RNN and its variants (bidirectional RNN [BRNN] and LSTM). These developed models learn the movement behavior of a specific vehicle. Analyzing the learning of many different vehicles, it is observed that vehicles that travel across larger geographic areas, and thus have a more complex movement pattern, also have a higher prediction error. The problem of high variance is approached by using regularization and dropouts. The study shows from the experiments that the LSTMs have better predictions when compared with simple RNNs and BRNNs in analysis and implementation of the near-time future prediction of truck locations.

Another study [26] in the finance field compared LSTM and support vector regression (SVR) for implementing a robust forecasting model to facilitate predicting phone prices in European stores. The study applied LSTM for its architecture to consider the problems that are not addressed by classic RNNs. In addition, the study utilized the support vector machine (SVM) method for both classification and regression because of its ability as a potent and accurate machine-learning method for univariate models, as shown in variant studies. In terms of introducing more variables for the multivariate approach, the models exhibited better prediction performance. By comparing the results, it was found that the SVR model can forecast the next-day price with a root mean squared error (RMSE) value of 33.43 euros for the univariate model. Nevertheless, applying multivariate models, the LSTM RNN gives the most accurate forecasting result for the next day's price, with an RMSE of 23.640 euros.

In the case of model flexibility to deal with missing values, [17] discussed developing the LSTM model, the model masking with vector M to perform two functions in predicting the traffic flow, called LSTM-M. This advanced approach can assume traffic flow data despite the missing values. The traffic flow dataset used in the study covered the time interval between April 4th, 2015 and January 3rd, 2016, including spring and winter seasons, by considering all-weather fluctuations for rainy, snowy, and sunny days. The study area relied on data collected from different locations of Hangzhou elevated road. In the data gathering step, the study faced missing data related to describing the working condition of the emergency detection and road equipment. The study applied multiscale temporal smoothing to infer the missing data and determine the prediction for the residual values. However, the time interval affected the rate of missing data. In a particular period, there was a small occurrence probability of overall data loss appearing with the larger time interval. It is unreasonable to have a large time interval when the traffic situation should be reflected accurately in a short period. As a result, the LSTM-M's superiority exceeds that of LSTM, producing a gain of over 1.5% in the average accuracy. When the traffic flow is massive, this implies a large mean absolute error (MAE) value, meaning that a greater presence of vehicles on the road may give rise to traffic accidents or bursts, and the traffic flow series may lead to fluctuations.

Author in [26] introduced two forecasting tasks based on the LSTM model. The first task was recognizing the unique features of the traffic data by applying deep LSTM at the peak hour for forecasting traffic. The other task was dealing with traffic datasets collected from 2,018 loop detectors, located along the Los Angeles arterial streets and highways. These loop detectors showed the datasets as large-scale and real-world datasets because they covered 5,400 miles cumulatively for intervals between May 19th and June 30th, 2012. The second task utilized a mixture of deep LSTM architecture that consolidated deep LSTM with a stacked autoencoder. The loop detectors picked three essential types of accidents, which were as follows: minor injuries, rolling, and major injuries. Approximately 1,650 sensors detected 6,811 accidents. There were different attributes associated with each accident, like the affected traffic direction and downstream post mile. The framework for the two situations proved the significant performance improvement in large-scale real-world traffic data. The suggested approach showed 30–50% superiority over the baseline, which was achieved by training end to end with proper regularization. In addition, the study described a novel technique to illustrate the model with signal stimulation and engaging summaries noted from the trained neural network (NN).

Regarding prediction of the frequency of traffic accidents using spatial data, [18]represented the risk of traffic accidents based on frequency, constructing a deep-learning model relying on LSTM for capturing the regional spatial and temporal correlation patterns. The studied dataset considered

traffic accidents based on different features, such as time, and the global positioning system coordinates of Beijing between 2016 and 2017. The model performance relied on comparing RMSE values, with the predicted risk map illustrating the effectiveness and accuracy of the proposed model. The study showed that the influences of temporal–spatial features, big traffic accident data, and the deep RNN on gaining accurate traffic accident risk prediction. The traffic accident warning system utilized in the study approach can help people avoid traffic accidents by choosing safer regions. However, the study had some limitations in that it utilized the traffic accident data for prediction, without other related data, such as human mobility, traffic flow, special events, and road characteristics, which may also be vital for traffic accident risk forecasting. The prediction results were coarse grained, and therefore, could not afford the level risk forecast of the road accident. Nevertheless, the approach can be efficiently utilized for road network–based forecasting. Therefore, future research should be conducted to combine the urban road network structure and comprehensive features related to traffic accidents to generate more reliable prediction results. Indeed, the LSTM approach's ability to generate predictions using the temporal features of traffic has been illustrated. However, LSTM algorithms fail to capture the spatial features and represent them on the map [27]. For this reason, the present study isolated the spatial features from prediction and took only their benefit to represent real and predicted accidents, fatalities, and injuries on an interactive map.

B. Visualizing Risk

Data visualization is a successful way to encode information so that our eyes can observe it and our brains can comprehend it. This is much more a science than an art, and we can only achieve successful results by studying human thought. The purpose is to translate abstract information into visual representations that can accurately, efficiently, easily, and meaningfully decode information. Data visualization is used to preset awareness of multiple risks in the world or make a decision concerning a specific risk. There are different purposes for using choropleth maps of a specific country or area. For example, maps are used to visualize diseases[28], flooding [29], traffic accident [11], and other risks and social issues. There are various tools for visualizing data; as discussed in [30] these include technical and nontechnical tools. Author in [20] applied bar and row charts with colored maps using a library called Dimensional Charting JavaScript (dc.js). This library allows highly efficient exploration of large data on accidents extracted from newspapers by the text-mining process. The dc.js library facilitates visualizing data and analyzing it in an interactive way on mobile devices or in a web browser. The map also has bar charts with color variations to visualize the statistics of accidents, with the degree of risk according to accident locations. To create online interactive maps with charts, plotly and folium are powerful and famous libraries for visualizing data by python mentioned in[30].

Folium is a python wrapper for Leaflet.js; this library facilitates the binding of the data to the map. The tool was developed to take in the trajectory of paths and plot them on a map. In one work, the folium library, with the help of Java code, contributed to plotting big datasets into an interactive geotemporal map of Morocco for assessing the air quality affected by significant correlations between emissions spread in air; this was done by incorporating the set of industrial activities associated with thermal power factories, plants, ports, and transportation[31]. Another study [32]applied folium to creating dynamic and interactive energy maps along with different informative charts allowing different stakeholders to view the characterization of the energy performance of buildings located in different areas in Italy, supported by different data and knowledge-visualization techniques. In addition [25], used folium with Matplotlib and plotly for visualizing the data in the near-time future location prediction of trucks' locations under location-based services to provide valuable services in the transportation sector to SCANIAs company customers in Taiwan.

C. Saudi Arabia as a Case Study

The RTA problem has a limited study history in traffic safety and traffic awareness literature in the Kingdom of Saudi Arabia (KSA). The total publication number for the traffic safety field and traffic awareness field is around 50 studies as identified in [33]. Author in [33] examined the causality relationship between the RTA, GDP, population, road miles, road vehicles, and number of driver's licenses in the KSA for 1971–2012, using a multivariate framework analysis. The study applied the autoregressive distributed lag (ARDL) model for cointegration in the KSA, employing the cointegration test. The results showed that the variables are cointegrated, meaning that there is a stable long-run relationship between RTA and its determinants, although there may be deviation in the short-run; thus, RTAs and the independent variables have bidirectional causality in the KSA.

After 2016, some studies discussed the RTA issue in relation to different fields, such as data mining for analyzing RTAs[34] using the internet of things (IoT) for classifying reasons for RTAs [35] employing the IoT to create e-awareness application against RTAs[35], and detecting RTAs via the IoT with ANNs[36]. In terms of prediction[37], performed the first study was conducted in 1995 and used three ARIMA models with a Box–Jenkins methodology for time-series data to predict the total traffic accidents, injuries, and fatalities in the KSA. The three models developed in this study showed the need for a quick revision of the current traffic safety programs in Saudi Arabia since no decrease was evident in the future predictions of the numbers of accidents, injuries, and fatalities. An improvement in traffic safety plans is urgently needed to reduce these numbers in the future. Author in [6] showed that the ARIMA model can forecast RTAs using time-series data. It introduced the applicant model for making RTA forecasts for up to 7 years. The prediction values of traffic accidents showed that there will be increasing deaths and injuries in the coming years.

Predicting RTAs is still a highly challenging problem because of the diversity of factors causing accidents from one place to another. In addition, data resources are restricted in some countries, limiting the quality of prediction results and visualization tools. This paper aims to elaborate on the concept of RTA prediction using deep-learning techniques, with the Kingdom of Saudi Arabia (KSA)—where such research is still lacking—receiving focus as a case study. This tutorial can be a

starting point for considering the issue more deeply in future research. The paper introduces the design of an online interactive map to visualize both real and predicted RTAs, fatalities, and injuries for all KSA regions in the years 1417 to 1452, based on the Hijri Calendar (AH), according to the annual historical data.

III. Designing the Process

The entire process of the experiment is an interactively advanced solution to visualize the actual and predicted numbers of accidents (As), fatalities (Fs), and injuries (Is). This process involves Python programs with five leading platforms—three for future prediction models and two for visualization. Therefore, the complete process can help address the global need for RTA prediction. First, it is necessary to prepare the historical data for visualizing the training dataset. The preparation step requires some preprocessing to prepare the data for prediction.

After predicting, the Python program interactively visualizes the number of road accidents, fatalities, and injuries in a web browser. There are different tools that can be utilized for visualization purposes, such as geographical maps and histograms. Visualizing data contributes to facilitating the understanding of complicated statistics in an easy, accurate way in a short time; due to these features, the concept of visualizing data has become popular on the internet. Fig. 1 illustrates the sequence of the experimental process. The explanation below identifies the steps in the figure, considering both directions.

A. Preparing Raw Data

The study dataset comprises numerical data obtained from reports of the KSA's General Department of Traffic, Ministry of Interior, Saudi Open Data, and General Authority for Statistics. The data contain the statistics on accident, injury, and fatality records classified in regional form between 1417 (1996) and 1433 (2013). In addition to the necessary information—that is, totals of "A," "I," and "F" per KSA region—the dataset also contains many valuable features related to accidents, such as vehicle numbers (N), road lengths (L), and population statistics (P) for males (Pm) and females (Pf) in each region. The dataset is arranged in the form of CSV/Excel files to be ready for the next step. Ultimately, we attain four files (accidents, injuries, fatalities, and factor attributes) based on the regional classification. The dataset requires some preprocessing operations to be ready for modeling. There is a need to perform a set of preprocessing tasks with the help of Python libraries to import, read, operate, clean, interpolate, extrapolate, and visualize data. We import Pandas to read, manipulate, and perform mathematical operations on the data in the form of CSV/Excel files. In addition, we use this library to handle the missing values via cleansing, extrapolating, and interpolating data. Other libraries used are Numpy to perform complex mathematical array operations and Matplotlib for data visualization in the form of line charts.

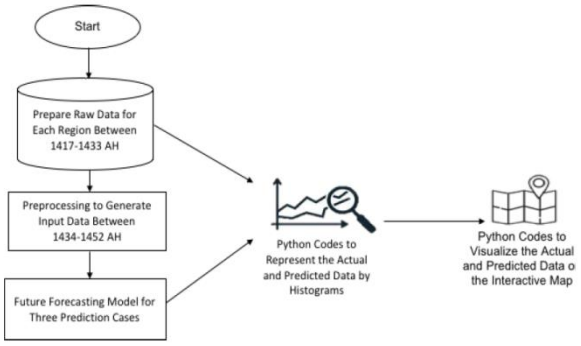


Fig. 1. The Steps Performed in the Tutorial.

B. Preprocessing to Generate Input Data

There is a need to populate all the data for 1434–1452 AH because there are no visible data for this period. This is done by calculating the average annual growth percentage for each input feature column per region after preprocessing the missing values. The four following features are used to make the required predictions:

- Male population (Pm).
- Female population (Pf).
- Number of registered vehicles (N).
- Length of roads in kilometers (L).

All the mentioned features with the original statistics on accidents, fatalities, and injuries can be accessed for each region in the project folder; Fig. 2 and 3 illustrate these data before and after preprocessing. Based on the provided dataset for 1417–1433, we build the generating input data for each region for the years 1434–1452, as illustrated in Fig. 3, to predict the following three quantities for every given region in the dataset: As, Fs, and Is. After generating the input data, the data must be normalized before regression by standardizing the column-wise data along the mean and dividing the column-wise data by the standard deviation for each column.

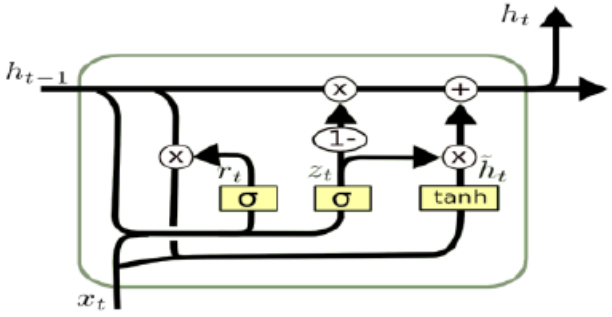


Fig. 2. Structure of the LSTM Cell that Describe (1), (2), (3), and (4).

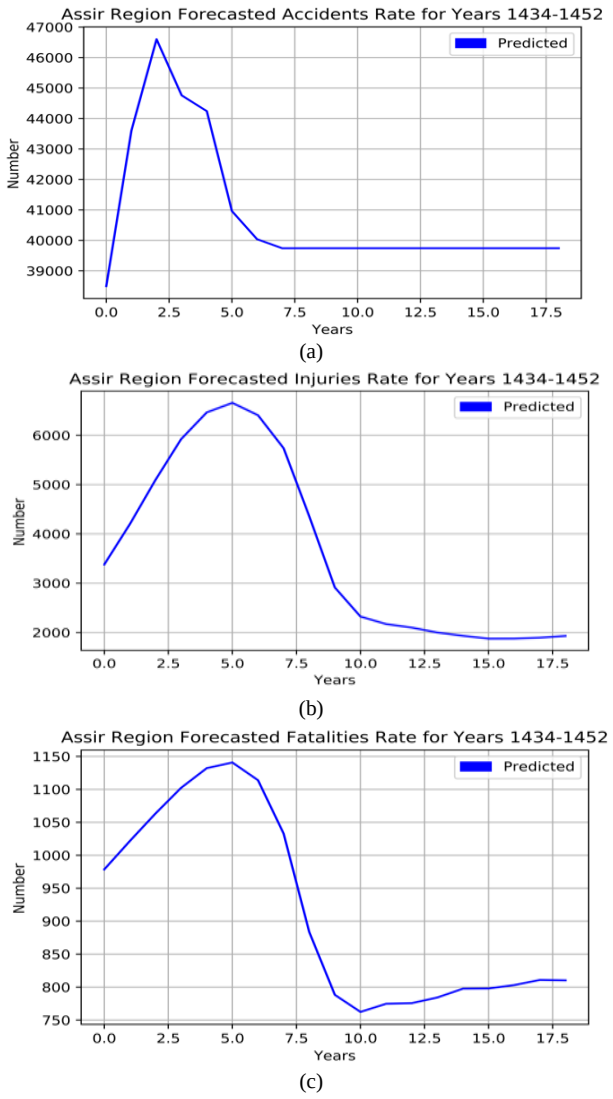


Fig. 3. (a). Future Accidents Prediction for the Assir Region. (b). Future Injuries Prediction for Assir Region. (c). Future Fatalities Prediction for Assir Region.

C. Future Forecasting Model

In this phase, the data are split into two groups, with the years 1417–1433 for training and 1434–1452 for testing. Then, the data are transformed into a sequence-to-sequence learning model. The data must pass through a set of tasks with the support of LSTM. LSTM is the most popular variant of RNN, representing a classic deep-learning method for handling sequence learning tasks. The Python libraries and Keras, which is based on Tensorflow, are used to import, read, operate, clean, interpolate, extrapolate, and build the models. The preprocessing returns a reframed data frame, which is first split into the training and test data based on the number of years, then split based on the inputs and outputs of both separated data frames. We use LSTM for the time-series data forecasting, which requires a specific three-dimensional input format of our dataset, where we need the size of input rows, timestep value, and feature length value. The model is multivariate, comprising (N, 1, M), with N representing the sample number, 1 the timestep, and M the number of multiple features depending on factors selected for the study. For each region; the multivariate model is constructed by 16 samples and one timestep with five features regarding each predicted quantity for 19 years. The model is executed based on one layer of four LSTM neurons, converging to one layer of a single dense neuron. The current models are run for 10 epochs each, with a batch size of 2, where the data are kept unshuffled to maintain the sequence. The model's loss function is the mean absolute error (MAE), with the Adam algorithm being the optimizer. The training data comprise a sequence of features that are passed into a layer of a number of LSTM neurons. Each LSTM neuron has the following inner workings [38]:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \quad (1)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]), \quad (2)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]), \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t, \quad (4)$$

where h_{t-1} is the output from the previous iteration of the same neuron, h_t denotes the output of the neuron for the same iteration, and x_t denotes the input for the current iteration. The adam optimizer is used for the weight optimization. One dense layer specifies the need to estimate a single value, that is, the number of accidents, injuries, or fatalities for a given instance of time. The loss function represents the difference between the real and predicted value, which helps in updating the weights, as shown in Fig. 2.

Finally, the model represents three graphs for each region, indicating the predicted values for the future As, Is, and Fs. The following figures illustrate the model in three cases for the Assir region sample—Fig. 3(a) for accidents, Fig. 3(b) for injuries, and 3c for fatalities. Table I shows the model values for each year based on the five features for the Assir region to predict A, I, and F.

TABLE. I. MODEL CASE RESULTS FOR THE ASSIR REGION SAMPLE

o. of Year	*Year*	*Predicted Accident Values*	*Predicted Injury Values*	*Predicted Fatality Values*
1	1434	38497	3379	978
2	1435	43603	4199	1021
3	1436	46601	5096	1063
4	1437	44758	5921	1102
5	1438	44240	6464	1132
6	1439	40959	6657	1140
7	1440	40038	6407	1113
8	1441	39742	5738	1032
9	1442	39743	4359	883
10	1443	39743	2913	788
11	1444	39743	2324	762
12	1445	39743	2171	774
13	1446	39743	2100	775
14	1447	39743	2000	784
15	1448	39743	1933	797
16	1449	39743	1876	797
17	1450	39743	1877	803
18	1451	39743	1895	810
19	1452	39743	1930	810

D. Representing Data by Histograms

The step of representing data using graphs has a Python code to create colorful histograms with legends to specify the degree of risk using Matplotlib and Plotly, which help stream the data to be analyzed and visualized. The histograms represent the real and predicted numbers of accidents, injuries, and fatalities, with colored legends to indicate the degree of risk in each region. Fig. 4(a) illustrates the histogram of real accidents for the Assir region in 1417–1433, while Fig. 4(b) shows the predicted values of accidents in 1434–1452 for the same region.

E. Representing Data on the Interactive Map

The phase of applying data visualization as an interactive map combines science and art. Using the map, the user can obtain information clearly and efficiently from information graphics constructed using meaningful information from data attributes or variables. The perfect template of data visualization will give effective results, aiding in data unification and analysis for the end user. The purpose of using data visualization is for promoting the eye's ability to differentiate extensive, diverse data and vast coherent information, serving a clear purpose in a small space. Web browser media support the designed information being displayed colorfully and interactively. The related information on road accidents, injuries, and fatalities is extracted to the interactive map, then visualized regionally according to the following entities:

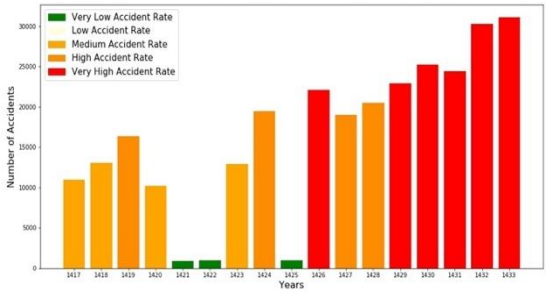
- The number of accidents, injuries, and fatalities for RTAs in 1417–1433.
- The number of accidents, injuries, and fatalities for RTAs in 1434–1452 (2030).

For data visualization, there are many different approaches and tools that can be employed. In this process, the Folium library (i.e., one of the Python libraries for interactive visualization) is used. The creation of an interactive map by Python, including a set of technical details, is illustrated in Fig. 5.

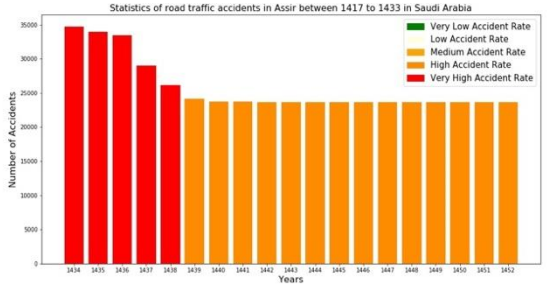
The visualizing phase represents the real and generated A, I, and F data and the results of predictions in the form of histograms on an interactive map using Python libraries like Folium. The interactive map is created using Python via the steps illustrated in Fig. 5. Folium is a Python library that allows visualizing the preprocessed data on Leaflet.js maps. This document covers and explains the process of making interactive choropleth maps from a given dataset on a map using the Folium library. Before visualizing the data on the KSA map, it is necessary to define the boundaries of the country and its regions to enable identifying the area on the world map being used. To obtain the GeoJSON file of a country, there are also a few steps that need to be taken:

- First, it is necessary to download the shape file of the geographical location that we are interested in. A shape file is a digital vector storage format for storing geometric location and associated attribute information. Country shape files can be downloaded from gadm.org.
- After downloading the shape file, it is necessary to convert it into a GeoJSON file for use with Folium to display all the region boundaries on the map. Converting the shape file to a GeoJSON file can be done through the MapShaper website. By uploading the shape zip file to the website, it is possible to export a GeoJSON file from the website of the same area/country.

The following screenshots of the KSA online map illustrate the related data visualizations for each region according to any of the following selected features: accidents, injuries, fatalities, future accidents, future injuries, and future fatalities; the related histogram appears after pressing the location icon of the region. In Fig. 6(a), the map represents the real numbers of accidents in the Assir region in 1417–1433. The same thing occurs when selecting the future accident case to represent accidents in Assir in 1434–1452, as shown in Fig. 6(b). As the samples show in Fig. 6(a) and 6(b), there is an identical histogram indicating the numbers of As, Is, and Fs separately for each region, covering the real and predicted cases. All these histograms are connected to the map internally to calculate the degree of risk, represented by the colored regions. The risk level for the map regions is based on the total numbers of As, Is, and Fs during the study years.



(a)



(b)

Fig. 4. (a). Risk Histogram of Real Accidents Statistics for Assir region in 1417-1434. (b). Risk Histogram of Predicted Accidents Statistics for Assir Region in 1417-1434.

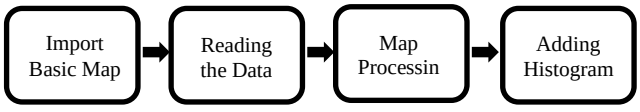
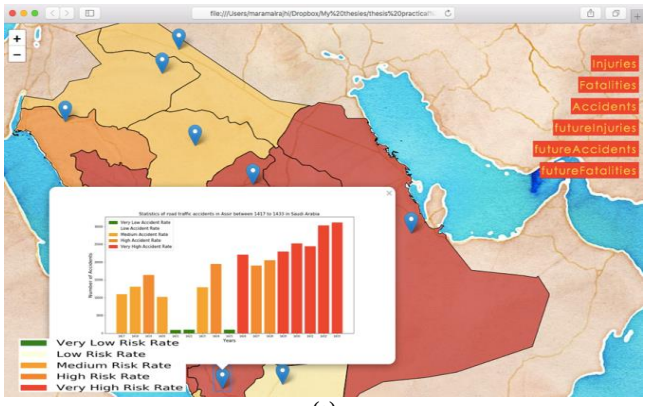
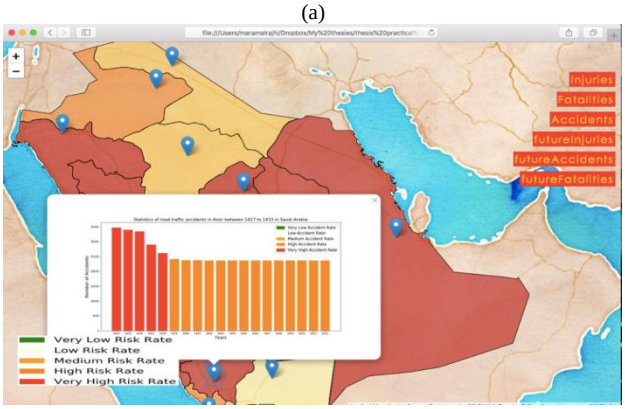


Fig. 5. Interactive Map Design Process.



(a)



(b)

Fig. 6. (a). Assir Case Illustrating the Data Visualization of Accidents Statistics in 1417–1433. (b). Assir Case Illustrates the Data Visualization of Predicting Accidents Statistics in 1434-1452.

IV. Evaluation

For representing information, it is important to maintain a high accuracy rate. This is crucial for evaluating the performance of the interactive map process. There are two approaches that are essential in evaluating the prototype process. The first approach involves measuring the accuracy of the predicted results. It is necessary to identify and estimate the efficiency and effectiveness of the LSTM with the dataset by calculating the mean squared error (MSE) or use other estimator tools for estimation purposes. Most of the evaluation metrics can be computed using Scikit-Learn in Python. The second approach is using feedback from decision makers or researchers as the end users of the whole process, focusing on how the interactive map can help for different purposes (improving decisions, use as an analytical reference, or for spreading awareness) by measuring criteria like perceived enjoyment, usefulness, and ease of use.

This paper checked the first approach by calculating the MAE, MSE, and root MSE (RMSE). Equations (5), (6), and (7) illustrate the syntax of the mentioned metrics:

$$MAE = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}, \quad (5)$$

$$MSE = \frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}, \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}}, \quad (7)$$

TABLE. II. Error Rates for All Cases of Future Model

Error Rate	*Accident Case*	*Fatalities Case*	*Injuries Case*
MAE	3736216722981	848	1757
MSE	1340508538934528620363776	1233102	4180412
RMSE	318260542776	1110	2044

where A_t indicates the real accident, injury, or fatality values in a certain year t, F_t indicates the corresponding forecasted value for each A_t, and n represents the total dataset years. For the study model's evaluation, Python calculates the mentioned metrics for each case as in Table II.

The study faced data limitations in that it was forced to deal with annual data; just 14 datapoints were collected for each region, representing the number of years differentiating the study data. This resulted in the establishment of a small dataset to construct the time-series problem to be solved by deep learning. For future works, it would be better to consider cities or parts of cities instead of regions. In addition, for obtaining a better result with LSTM, it is important to have a long period of time to include hundreds or thousands of datapoints. For 14 datapoints, it would be better to have extra datasets on a monthly, daily, or hourly basis than the annual standard to apply LSTM perfectly with time-series problems. Perhaps, for the study dataset, it would be a good idea to combine another tool like a Convolutional Networks (CNN) with LSTM, as accomplished in [39], to improve the performance of short-term forecasting via a proposed new method called CNN & LSTM traffic flow prediction method (CLTFP) for forecasting future traffic flow data.

V. Utilization and Improvement

The proposed LSTM process can be applied online, particularly for statistical presentations, providing in-depth details and interactive representation. This is because data visualization offers an easy template that facilitates a comparison of different parts of the data, displaying a huge amount of data in a small area and a lot of coherent information to serve a clear goal. The internet is a public medium, and displaying information via web browsers is highly useful. We can see some popular newspapers and statistical and weather organizations using visual data representation to present informative reports and predictions.

The study process can be further enhanced. The procedure is practiced as a prototype for historical data to analyze, predict, and visualize the numerical data on accidents, injuries, and fatalities related to RTAs. In this tutorial approach, limited data have been used in regional-based research. The results can be made more dynamic by automatically processing these statistical data from different official web resources. Several factors can be added to improve deep learning, such as the weather, time of occurrence, and spatial location of incidents in relation to Saudi cities. One important factor is that only information about road accidents, injuries, and fatalities is extracted and visualized because of limitations on resources in the KSA case study; yet, there are still many critical details that should be included to evaluate the risk more accurately.

Concerning KSA Vision 2030, it is essential—as the KSA is an economically influential country—to be aware of death statistics every day. This whole process can be adopted by decision makers to reduce the considerable losses in health, social, and economic life, and it can be configured from our social responsibility. Limitations in the available data cause problems in evaluating the best prediction method to analyze them and indicate that the results will be inaccurate. This issue can be reconstructed for further enhancement of this process by trying other methods and expanding the dataset in more depth. Adding many related factors can be extremely helpful for enhancing the interactive map. In addition, working with live and spatial data can provide more useful and creative elements.

VI. Conclusion

The design of an interactive map prototype was described in this paper. The work presented a tutorial approach to illustrate the analytical and practical process of generating a prototype. Predicting the future of accidents, fatalities, and injuries was done based on historical data collected from some KSA websites. The collected data were saved in CSV files and converted to the sequence-to-sequence problem solved by LSTM in Python by Keras. Then, the real and predicted data were represented in ranked, colored histograms to act on the map in the form of combining the art and science to allow easy observation of risk. In the future, this process can be extended and evaluated more deeply after the disposal of the data limitation problem. New trends in data science can be added to this prototype to attract the attention of decision makers and the public to the concern of road accidents in the KSA.

References

[1] J. J. Rolison, S. Regev, S. Moutari, and A. Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," Accid. Anal. Prev., vol. 115, pp. 11–24, February 2018.

[2] A. A. Mohammed, K. Bin Ambak, A. M. Mosa, and D. Syamsunur, "Classification of traffic accident prediction models: a review paper," Int. J. Adv. Sci. Eng. Technol., vol. ISSN, no. 6, pp. 2321–9009, 2018.

[3] M. M. Islam, M. Alharthi, and M. M. Alam, "The impacts of climate change on road traffic accidents in Saudi Arabia," Climate, vol. 7, no. 9, pp. 1–13, 2019.

[4] N. T. Ratrout, S. Chowdhury, U. Gazder, and S. M. Rahman, "Characterization of crash-prone drivers in Saudi Arabia—a multivariate analysis," Case Stud. Transp. Policy, vol. 5, no. 1, pp. 134–142, 2017.

[5] C. Liu and A. Sharma, "Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity," Anal. Methods Accid. Res., vol. 17, pp. 14–31, February 2018.

[6] M. Al-zyood, "Forecast car accident in Saudi Arabia with ARIMA models," Int. J. Soft Comput. Eng., vol. 7, no. 3, pp. 30–33, 2017.

[7] C. Voyant et al., "Machine learning methods for solar radiation forecasting: a review," Renew. Energy, vol. 105, pp. 569–582, May 2017.

[8] J. Chen et al., "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide," Environ. Int., vol. 130, September 2019.

[9] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," IJCAI Int. Jt. Conf. Artif. Intell., vol. 4, pp. 2627–2633, August 2017.

[10] M. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," Appl. Sci., vol. 7, no. 6, p. 476, 2017.

[11] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, pp. 984–992.

[12] DE Rumelhart, GE Hinton, and RJ Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533–536, 1968.

[13] G. P. Zhang, "Neural networks for time-series forecasting," Handb. Nat. Comput., vol. 4, pp. 461–477, 2012.

[14] J. Verdegaal, "External memory enhanced sequence-to-sequence dialogue systems," University of Amsterdam, 2018.

[15] K. Xu, H. Wang, and P. Tang, "Image captioning with deep LSTM based on sequential residual," in Proceedings—IEEE International Conference on Multimedia and Expo, no. 2017, pp. 361–366.

[16] C. Ma et al., "Visual question answering with memory-augmented networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 6975–6984.

[17] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "LSTM-based traffic flow prediction with missing data," Neurocomputing, vol. 318, pp. 297–305, November 2018.

[18] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, November 2018, pp. 3346–3351.

[19] R. Trianto, T. C. Tai, and J. C. Wang, "Fast-LSTM acoustic model for distant speech recognition," in 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, January 2018, no. 2, pp. 1–4.

[20] H. Akhter, "Information extraction and interactive visualization of road accident related news," Int. J. Comput. Appl., vol. 128, no. 5, pp. 37–40, 2015.

[21] S. Murray, Interactive Data Visualization for the Web, 2nd ed., Sebastopol CA, O'Reilly, 2017.

[22] S. H.-S. Brian Suda, "The 38 best tools for data visualization | Creative Bloq," Creative Blog Art and Design Inspiration, 2017. [Online]. Available: https://www.creativebloq.com/design-tools/data-visualization-712402. [Accessed: 22 February 2018].

[23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: concerns and ways forward," PLoS One, vol. 13, no. 3, pp. 1–26, 2018.

[24] H. P. Sascha Krstanovic, "Ensembles of recurrent neural networks for robust time series forecasting," in International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2017, pp. 34–46.

[25] A. Srinivasan, "Near-time predictions of future truck locations," Chalmers University of Technology, 2017.

[26] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning: a generic approach for extreme condition traffic forecasting," in Proceedings of the 2017 SIAM International Conference on Data Mining, 2017, pp. 777–785.

[27] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," Sensors (Switzerland), vol. 17, no. 7, 2017.

[28] N. A. Samat and L. W. Mey, "Malaria disease mapping in Malaysia based on Besag-York-Mollie (BYM) Model," J. Phys. Conf. Ser., vol. 890, no. 1, 2017.

[29] B. Elboshy, S. Kanae, M. Gamaleldin, H. Ayad, T. Osaragi, and W. Elbarki, "A framework for pluvial flood risk assessment in Alexandria considering the coping capacity," Environ. Syst. Decis., vol. 39, no. 1, pp. 77–94, 2019.

[30] E. G. Caldarola and A. M. Rinaldi, "Big data visualization tools: a survey: the new paradigms, methodologies and tools for large data sets visualization," in DATA 2017—Proceedings of the 6th International Conference on Data Science, Technology and Applications, 2017, pp. 296–305.

[31] B. B. Semlali and E. A. Chaker, "Towards remote sensing datasets collection and processing," Int. J. Embed. Real-Time Commun. Syst., vol. 10, no. 3, pp. 49–67, 2019.

[32] T. Cerquitelli et al., "Exploring energy performance certificates through visualization," in CEUR Workshop Proceedings, August 2019, vol. 2322.

[33] M. M. Ageli and A. M. Zaidan, "Road traffic accidents in Saudi Arabia: An ADRL approach and multivariate granger causality," Int. J. Econ. Financ., vol. 5, no. 7, pp. 26–31, 2013.

[34] I. Al-Turaiki, M. Aloumi, N. Aloumi, and K. Alghamdi, "Modeling traffic accidents in Saudi Arabia using classification techniques," in 2016 4th Saudi International Conference on Information Technology (Big Data Analysis), KACSTIT 2016, 2016, pp. 1–5.

[35] M. A. Alharbe, "Awareness ability and influences on raising of traffic accidents through the content of social media in the internet of things: a practical empirical study by the internet of things and multimedia on university students in western Saudi Arabia," in Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018, 2019, pp. 48–51.

[36] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in 2018 15th Learning and Technology Conference, L and T 2018, 2018, pp. 40–45.

[37] A. S. Al-Ghamdi, "Time series forecasts for traffic accidents, injuries, and fatalities in Saudi Arabia," J. King Saud Univ.—Eng. Sci., vol. 7, no. 2, pp. 199–217, 2018.

[38] C. Olah, "Understanding LSTM networks," Blog, 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[39] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," vol. v1, pp. 1–14, December 2016.

Optimization of Multi-Product Aggregate Production Planning using Hybrid Simulated Annealing and Adaptive Genetic Algorithm

Gusti Eka Yuliastuti[1], Agung Mustika Rizki[2]
Wayan Firdaus Mahmudy[3]
Faculty of Computer Science
Brawijaya University, Malang, Indonesia

Ishardita Pambudi Tama[4]
Faculty of Industrial Engineering
Brawijaya University, Malang
Indonesia

***Abstract*—In the planning of aggregate production, company stakeholders need a long time due to the many production variables that must be considered so that the production value can meet consumer demand with minimal production costs. The case study is the company that produces more than a type of product so there are several variables must be considered and computational time is required. Genetic Algorithm is applied as they have the advantage of searching in a solution space but are often trapped in locally optimal solutions. In this study, the authors proposed a new mathematical model in the form of a fitness function aimed at assessing the quality of the solution. To overcome this local optimum problem, the authors refined it by combining the Genetic Algorithm and Simulated Annealing so called hybrid approach. The function of Simulated Annealing is to improve every solution produced by Genetic Algorithm. The proposed hybrid method is proven to produce better solutions.**

Keywords*—Agreggate; genetic algorithm; hybrid; production planning; simulated annealing

I. Introduction

The industrial world in Indonesia has experienced tremendous rapid development, especially in manufacturing. This certainly can have an impact on increasing fierce competition between companies. Each company has almost the same goal of wanting to always meet consumer demand. However, the company cannot get rid of one of the problems that must be faced, namely stochastic consumer demand or can be said to be always changing. Every company must be able to overcome these problems so that consumers do not switch to other competing companies. From these basic problems, indirectly every company is required to increase its product production effectively and efficiently to meet consumer demand. Several strategies are usually implemented by the company, among others level strategy, time flexibility strategy and mixed strategy. Each strategy has its advantages and disadvantages so it will be less than optimal in handling stochastic consumer demand problems.

In addition to problems related to consumer demand, some problems are quite crucial for the company that is related to production costs. The problem faced by the company was known after the author interviewed the owner of the company. Every company must be able to optimize all company resources to produce a product that is following consumer demand but with minimal cost. Because the amount of production costs is very influential to the profits of the company, so the calculation of production costs needs to be done optimally. To achieve the goals of the company, of course, there is a need for optimal production planning. By the optimal production planning, it will be able to minimize production costs but still be able to meet consumer demand so that the profits obtained by the company will be maximized. One way to do production planning in the medium term is aggregate planning. Aggregate production planning is a procedure to translate the expected sales demand and also the capacity of product production availability for production plans in the coming period [1].

Based on this background, the author conducts research related to production planning to address the problem of stochastic consumer demand and minimize production costs incurred. This problem is important to solve because it is relate to something crucial in the company. The method that will be applied by the author is the Genetic Algorithm, but solving the problem using the Genetic Algorithm is often a new problem that is the occurrence of early convergence or the solution achieved in several iterations there is no significant increase. Therefore, the author will give a little modification to the Genetic Algorithm and then combine it with the Simulated Annealing method to get more optimal results. The Simulated Annealing method works to improve the solution generated by the Genetic Algorithm after the solution is fixed by Simulated Annealing, the best solution will be returned to the Genetic Algorithm to continue the process until the stop condition is reached. The selection of the Simulated Annealing method by the author is based on previous research by Zhang *et al.*, [2]. Their hybrid is done by entering the process flow of the Simulated Annealing method into the Genetic Algorithm process flow in general. The fundamental difference in this study compared to before is the modification of the process flow of the Genetic Algorithm. By this modification, the aim is that the resulting solution will be more optimal and also the Genetic Algorithm will avoid early convergence.

II. Literature Review

Several studies have previously been carried out related to production planning which includes aggregate production, scheduling, distribution and so forth.

Fahimnia *et al.,* [3] optimize the integration of aggregate production and distribution planning using Genetic Algorithms. This Genetic Algorithm is used to determine the combination of the production of various types of products in several manufacturing plants and also the distribution of finished products from the factory to many users either through direct or indirect transportation routes.

The combination of two methods namely Genetic Algorithm and Taboo Search was carried out by Ramezanian *et al.,* [4] to solve the problem of aggregate production planning in a two-phase production system. The problem in this research is that the system is multi-period, multi-product and multi-machine. They developed the Mixed Integer Linear Programming (MILP) model for aggregate production planning in two general phases. Because these problems are quite complex, they apply Genetic Algorithms and taboo search. Computational results show that by applying both methods, a solution with better quality and more efficient solutions for large scale problems such as this is found.

Genetic Algorithms are also applied by Chistyakova *et al.,* [5] to the decision support system for optimal planning. Genetic Algorithms are considered algorithms that make it possible to use in solving planning problems that can be reconfigured. After testing the system, the results have proven to be more effective with Genetic Algorithms.

Erfanian and Pirayesh [6] also use Mixed Integer Linear Programming (MILP) to integrate the aggregate production planning and maintenance planning. The purpose of aggregate planning is to determine the level of production and inventory to minimize the total cost of production with limited labor and equipment resources. Also, the purpose of maintenance planning is to improve equipment efficiency to meet production requirements. In the proposed model, integrated aggregate production planning and maintenance are developed to determine the optimal production plan and preventive maintenance for each period.

The issue of integration of production and maintenance planning is also a topic of issues raised by Ettaye *et al.,* [7]. They also apply the Genetic Algorithm to solve these problems. Because according to them, planning maintenance activities can hamper manufacturing operations in terms of cost, quality and time. Besides, it is necessary to ensure the availability of production equipment to meet consumer demand. By using Genetic Algorithms, integrated planning is more optimal.

There is another study that takes a situation of the case of production planning at an oil company conducted by Siregar *et al.,* [8]. Production planning is less effective and efficient because it only relies on demand data from the past period. This has an impact on high-cost losses. Siregar *et al.,* are used the Goal Programming approach to optimize the company's production planning. This method is used to maximize or minimize the resources of several holding companies with an objective function. Goal Programming is a mathematical model that is suitable for solving multi-purpose problems because it uses variable deviation.

Based on several studies that have been done before, it is proven that Genetic Algorithms can overcome optimization problems. So the method that will also be applied by the author is the Genetic Algorithm. But solving the problem using the Genetic Algorithm apart often begin new problems namely the occurrence of early convergence or solutions achieved in several iterations there is no significant increase. Therefore, the author will give a slight modification to the Genetic Algorithm and then combine it with the Simulated Annealing method so that the aim is to get more optimal results. Modifications made by the author are changing the reproductive process both in crosssover and mutation and also genetic parameters that will be made adaptive to adjust the conditions at that time.

III. Aggregate Production Planning

Production planning is planning, organizing beforehand about people, materials, machinery and other equipment and capital needed to produce goods or services in a certain period in the future by what is predicted. Aggregates are demand forecasts for various goods combined into homogeneous units. Aggregate production planning relates to the production capacity used to respond to predicted consumer demand [9].

The case study that will be completed in this research is a home textile company that manufactures school uniforms of 18 types of products including superiors and subordinates. In each type of product, there are eight different sizes. To determine the level of production in the coming period is quite difficult because there is more than one type of product and also each product is done by different workers. Besides that, consumer demand for each product is also different and always changing. This is quite difficult for the company to do production on all types of products. They need to consider various parameters and adjust the rules that apply to the company in achieving optimal solutions. The optimal solution the company wants is that consumer demand is fulfilled with minimal production costs. In this study, several production parameters are used as shown in Table I.

TABLE. I. Parameters in Production Process

Para-meter	Description
n	Number of workers
rt	Average production amount per worker in a day (regular time)
ot	Average production amount per worker in a day (overtime)
sc	Average production amount per worker in a day (subcontract)
jrt	Working hours in a day (regular time)
jot	Maximum hours of overtime
msc	Maximum number of subcontract worker
brt	Production cost (regular time)
bot	Production cost (overtime)
bsc	Production cost (subcontract)
rp	Number of worker recruited (hiring)
prp	Average amount of production per new worker in a day
brp	Cost of recruiting worker
mp	Number of laid-off worker
bmp	Cost of laying off worker (firing)
kp	Number of production shortages
i	Number of inventory product
bi	Cost of each inventory product

IV. Modified Genetic Algorithm

Genetic Algorithm is a method adapted from the natural evolutionary process [10]. The terms in the Genetic Algorithm also use terms that are commonly used in the evolutionary process including genes, chromosomes, individuals, populations, generations, and reproduction. Genes are the smallest parts that represent solution units. Chromosomes are a collection of genes if assembled will be a solution. Such chromosome lines are also referred to as individuals. A collection of individuals is called a population. Whereas generation is a time when the population is still alive and the generation will continue to move forward with a population that will always change. Reproduction is the process of producing new individuals.

A. Chromosome Representation

The type of chromosome encoding used to represent the solution is real-coded. If there is a fractional value, then the value will be rounded when calculating the fitness value. Rounding the value in a way if the value is more than half it will be rounded up and vice versa. In a series of chromosomes, there are 18 segments that indicate the number of products. In each segment, there are 12 genes indicating the number of periods counted (for 12 months). The total length of the chromosome is 216 genes.

B. Reproduction Operator

Reproduction operators are needed to produce new individuals. There are two reproduction operators, crossover and mutation [11]. The difference between these operators is in the number of parent chromosomes used. In the crossover process, requires 2 parent chromosome, whereas in mutation only need 1 parent chromosome. The two operators are used at the same time but the rate should be determined at the beginning. For example, the crossover rate (*cr*) is set at 0,6 while the mutation rate (*mr*) is set at 0,4. Based on those rates, we can conclude that the reproduction process focuses more on the crossover.

1) Crossover process: The crossover method used one-cut-point-crossover in each segment to aim that the individual produced is more varied. Then generate a random value in each segment between intervals [1] because there are 12 genes in each segment indicates 12 months. The random value will be used as a reference gene that becomes the point of intersection. For example, on the parent chromosome, there are 12 genes with 3 segments as shown in Fig. 1.

In segment 1, we generate a random value and get a value that is 2. It means the second gene is cut as shown in Fig. 2.

In segment 2, we generate a random value and get a value that is 3. It means the third gene is cut as shown in Fig. 3.

In segment 3, we generate a random value and get a value that is 1. It means the first gene is cut as shown in Fig. 4.

Based on the crossover process, it produces the following child chromosome as shown in Fig. 5.

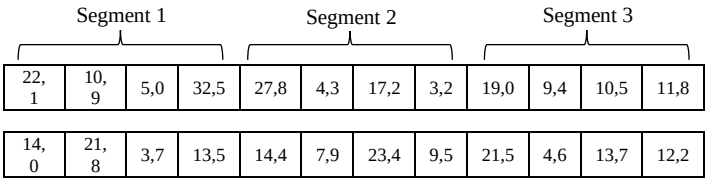


Fig. 1. Illustration of Chromosome Representation for Crossover.

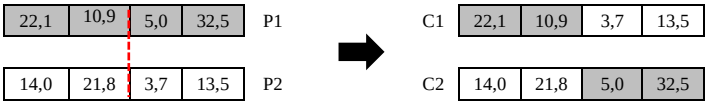


Fig. 2. Illustration of Crossover in Segment 1.

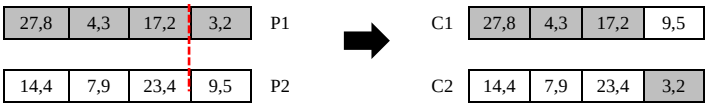


Fig. 3. Illustration of Crossover in Segment 2.

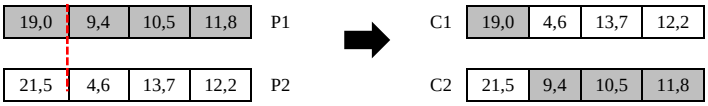


Fig. 4. Illustration of Crossover in Segment 3.

Fig. 5. Illustration of Child Chromosome Result.

2) Mutation process: The mutation method used position-based-mutation in each segment to minimize the occurrence of early convergence. Then generate a random value in each segment between intervals [1] because there are 12 genes in each segment indicates 12 months. The random value will be used to determine the position of the point whose value will be changed. The value at that position will be changed by generating a random value with a range of ± 10 from the initial value. For Example, on the parent chromosome, there are 12 genes with 3 segments as shown in Fig. 6.

In segment 1, we generate a random value and get a value that is 1. It means the first gene has to be changed. The new value is obtained by another random value and get a value that is 27,4 as shown in Fig. 7.

In segment 2, we generate a random value and get a value that is 4. It means the fourth gene has to be changed. The new value is obtained by another random value and get a value that is 12,5 as shown in Fig. 8.

In segment 3, we generate a random value and get a value that is 3. It means the third gene has to be changed. The new value is obtained by another random value and get a value that is 17,3 as shown in Fig. 9.

Based on the mutation process, it produces the following child chromosome as shown in Fig. 10.

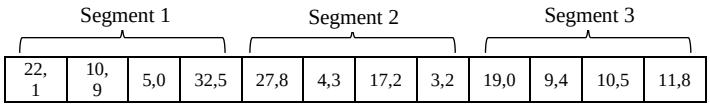


Fig. 6. Illustration of Chromosome Representation for Mutation

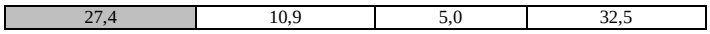


Fig. 7. Illustration of Mutation in Segment 1.

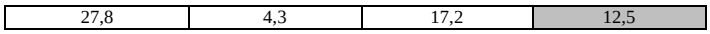


Fig. 8. Illustration of Mutation in Segment 2.

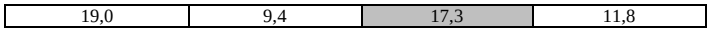


Fig. 9. Illustration of Mutation in Segment 3.

22, 1	10, 9	3,7	13,5	27,8	4,3	17,2	9,5	19,0	4,6	13,7	12,2
14, 0	21, 8	5,0	32,5	14,4	7,9	23,4	3,2	21,5	9,4	10,5	11,8

Fig. 10. Illustration of Child Chromosome Result.

C. Selection Operator

The type of selection operator used is elitism for reasons that are not complex so it does not take a long time at the time of execution. In conducting individual selection, it is necessary to have a unit of value to compare or hereinafter referred to as fitness value. This fitness value does not have a patent formula because it will never be the same because the formula is structured to solve the problem at hand. The fitness value formulation for this problem is shown as (2).

$$Total\ of\ Production\ Cost = (\sum rt.brt) + (\sum ot.bot) + (\sum sc.bsc) + (\sum rp.brp) + (\sum mp.bmp) + (\sum i.bi) \quad (1)$$

$$Fitness\ Value = \frac{100000000}{Total\ \ Production\ Cost + Penalty} \quad (2)$$

A penalty is some conditions that can reduce the quality of a solution such as inventory value minus. It means the company cannot complete consumer demand. If some rules are not suitable for the company related to the number of workers, the time of workers and so on, it can count as a penalty.

D. Adaptive Parameter

The adaptive parameter in the Genetic Algorithm referred to the value of the crossover and mutation rates which can adapt depending on the situation and conditions that occur at that time. The value will change and be randomized if in 10 iterations there are no better results.

E. Hybrid with Simulated Annealing

The Simulated Annealing (SA) and Genetic Algorithm (GA) certainly have their respective weaknesses. This research will combine the two algorithms with the hope that they will get more optimal results. The hybrid process of Simulated Annealing Algorithm and Genetic Algorithm (HSAGA) consists of combining the Genetic Algorithm with Simulated Annealing Algorithm iteratively [2]. The difference between this study and previous research is the modification of AG using adaptive parameters which are then proposed to be the method of Hybrid Simulated Annealing and Adaptive Genetic Algorithm (HSAAGA). At HSAAGA, SA is considered as an operator in AGA that functions to manage the population optimally [12]. During the initial AGA process, genetic operators are used as usual to produce new individuals and then SA operators are used to identifying these individuals.

V. Result and Analysis

The testing results of the proposed method namely Hybrid of Simulated Annealing - Adaptive Genetic Algorithm (HSAAGA) as shown in Table II. In the previous trial, all the best parameters were obtained from each method. All the best parameters will be used by the author to test the HSAAGA method. The best parameters in the Adaptive Genetic Algorithm are population size = 70, number of generations = 1000, combination of cr = 0.5 and mr = 0.5. As for the best parameters on Simulated Annealing namely the number of new solutions generated as many as 9, temperature = 175, reduction factor = 0.5 and maximum iteration = 100.

From 10 trials, an average fitness value of 0.4480 was obtained. The best solution with the highest fitness value is 0.4713 with execution time during 19701866 ms.

Knowing that this proposed method is the best method in solving the problem of aggregate production planning with optimal results, it needs to be proven by comparing it with other methods as shown in Table III. Each method will be run 5 times with the same time limit i.e. during 120 minutes so that the resulting solution is fairer in the comparative test of this method.

TABLE. II. The Testing Results of HSAAGA

	Trials										**Average**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Fitness Value	0,4713	0,4490	0,4439	0,4320	0,4414	0,4483	0,4468	0,4486	0,4400	0,4477	0,4480
Execution Time (ms)	19701866	19965587	19712745	19820371	19854968	20131036	19846575	19827304	19826300	19795550	19848230

TABLE. III. Methods Comparation

Methods	**Fitness Value**					
	Trials					**Average**
	1	**2**	**3**	**4**	**5**	
SA	0,22845	0,25940	0,23383	0,21125	0,22366	0,231321
GA	0,29984	0,30854	0,28576	0,31006	0,27509	0,295862
AGA	0,32141	0,32881	0,32942	0,34106	0,33106	0,330356
HSA-AGA	0,47132	0,44907	0,44395	0,43202	0,44145	0,447567

The proposed method of HSAAGA produces the highest average fitness value. Adaptive Genetic Algorithm that improved using Simulated Annealing was quite effective as evidenced by the increase in fitness value. The existence of adaptive parameters value of *cr* and *mr* in the Genetic Algorithm is very important in increasing yield. In addition, modifications during the reproduction process also play an important role in expanding the search space for solutions so that it is likely to achieve optimal solutions.

The actual aggregate production planning for the first product carried out by the company requires a production cost of Rp. 645,703,000. Different results are obtained by applying the HSAAGA method to carry out aggregate production planning for the first product.

The total production costs to be incurred by the company if applying the HSAAGA method is Rp. 612,731,500. There is a difference in the production costs of Rp. 32,971,500, - in the production of the first product which is quite influential for the company because the products produced by the company are quite a lot, namely 18 products. By applying this proposed method HSAAGA, the company can carry out aggregate production planning to get optimal results effectively and efficiently.

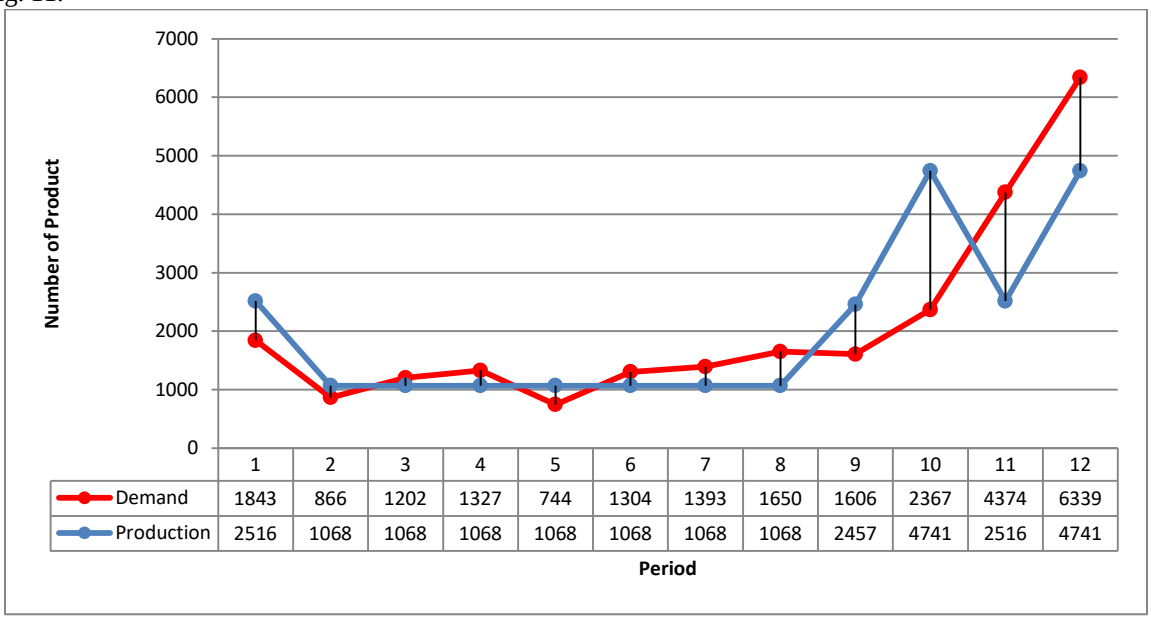
After obtaining the final solution in the form of the amount of production from the application of the method then the writer compares with consumer demand to see the difference as shown in Fig. 11.

VI. Conclusion

Aggregate production planning using the proposed method can produce an optimal solution because in determining the quality of a solution considered the total production costs incurred, the total number of workers, the total cost for workers and also some penalties that are not in accordance with company rules. So that, the final solution achieved represents the total minimum production costs and minimum penalties by hiring workers as needed for the production process, this can certainly have a good impact on the profits of the company.

The hybrid carried out between Simulated Annealing and the Adaptive Genetic Algorithm produced the most optimal solution. The purpose of this hybrid, the authors want to further increase the exploitation of the quality of the solutions that have been produced. Simulated Annealing functions to improve the solutions produced by Adaptive Genetic Algorithms. Every improvement can certainly improve the quality of the resulting solution so that the solution is considered more optimal. The most optimal solution resulting from the hybrid of Simulated Annealing and Adaptive Genetic Algorithm to carry out multi-product aggregate production planning is the fitness value of 0.4713223 with a production cost of Rp. 2,176,631,500.

For future work, this hybrid method can be used for another case study. On the other hand, another heuristic method can be added to increase the solution from this hybrid method.



	1	2	3	4	5	6	7	8	9	10	11	12
Demand	1843	866	1202	1327	744	1304	1393	1650	1606	2367	4374	6339
Production	2516	1068	1068	1068	1068	1068	1068	1068	2457	4741	2516	4741

Fig. 11. Comparison of Demand and Production.

REFERENCES

[1] M. Kumar and A. N. Haq, "Hybrid Genetic — Ant Colony Algorithms for Solving Aggregate Production Plan," vol. 4, no. 1, pp. 103–111, 2005.

[2] Y. Zhang, Q. Huang, F. Gao, and X. Sun, "Optimal Reservoir Operation Using a Hybrid Simulated Annealing Algorithm-Genetic Algorithm," Bio-Inspired Comput. Theor. Appl., 2010.

[3] B. Fahimnia, L. Luong, and R. Marian, "Genetic Algorithm Optimisation of An Integrated Aggregate Production – Distribution Plan in Supply Chains," Int. J. Prod. Res., no. October 2013, pp. 37–41, 2012.

[4] R. Ramezanian, D. Rahmani, and F. Barzinpour, "An Aggregate Production Planning Model for Two Phase Production Systems: Solving with Genetic Algorithm and Tabu Search," Expert Syst. Appl., vol. 39, pp. 1256–1263, 2012.

[5] T. B. Chistyakova, A. S. Razygrayev, and R. V Makaruk, "Decision Support System for Optimal Production Planning Polymeric Materials Using Genetic Algorithms," in Soft Computing and Measurements (SCM), 2016, pp. 257–259.

[6] M. Erfanian and M. Pirayesh, "Integration Aggregate Production Planning and Maintenance Using Mixed Integer Linear Programming," IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2016–Decem, pp. 927–930, 2016.

[7] G. Ettaye, A. E. L. Barkany, and A. E. L. Khalfi, "Applying Genetic Algorithm for Integrated Planning of Production and Maintenance," in Logistics and Supply Chain Management (LOGISTIQUA), 2017, pp. 166–170.

[8] I. Siregar, U. Tarigan, D. A. Siregar, R. M. Sari, M. M. Tambunan, and T. H. Nasution, "Determination of Optimal Production Planning in Palm Oil Company (Case Study)," in International Conference on Industrial Engineering and Applications, 2017.

[9] J. Heizer and B. Render, Production and Operations Management Strategies and Tactics. New Jersey: Prentice Hall, 1993.

[10] W. F. Mahmudy, R. M. Mariana, and L. H. S. Luong, "Hybrid Genetic Algorithms for Multi-period Part Type Selection and Machine Loading Problems in Flexible Manufacturing System," in IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM), 2013.

[11] V. N. Wijayaningrum and W. F. Mahmudy, "Optimization of Ship's Route Scheduling Using Genetic Algorithm," Indones. J. Electr. Eng. Comput. Sci., vol. 2, no. April, pp. 180–186, 2016.

[12] G. E. Yuliastuti, A. M. Rizki, W. F. Mahmudy, and I. P. Tama, "Determining Optimum Production Quantity on Multi- Product Home Textile Industry by Simulated Annealing," J. Inf. Technol. Comput. Sci., vol. 1, no. 1, 2018.

Hybrid Topic Cluster Models for Social Healthcare Data

K Rajendra Prasad[1]
Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India

Moulana Mohammed[2]
Dept. of CSE
Koneru Lakshmaiah University
Guntur, India

R M Noorullah[3]
Dept. of CSE
Koneru Lakshmaiah University
Guntur, India

***Abstract*—Social media and in particular, microblogs are becoming an important data source for disease surveillance, behavioral medicine, and public healthcare. Topic Models are widely used in microblog analytics for analyzing and integrating the textual data within a corpus. This paper uses health tweets as microblogs and attempts the health data clustering by topic models. The traditional topic models, such as Latent Semantic Indexing (LSI), Probabilistic Latent Schematic Indexing (PLSI), Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and integer Joint NMF(intJNMF) methods are used for health data clustering; however, they are intractable to assess the number of health topic clusters. Proper visualizations are essential to extract the information from and identifying trends of data, as they may include thousands of documents and millions of words. For visualization of topic clouds and health tendency in the document collection, we present hybrid topic models by integrating traditional topic models with VAT. Proposed hybrid topic models viz., Visual Non-negative Matrix Factorization (VNMF), Visual Latent Dirichlet Allocation (VLDA), Visual Probabilistic Latent Schematic Indexing (VPLSI) and Visual Latent Schematic Indexing (VLSI) are promising methods for accessing the health tendency and visualization of topic clusters from benchmarked and Twitter datasets. Evaluation and comparison of hybrid topic models are presented in the experimental section for demonstrating the efficiency with different distance measures, include, Euclidean distance, cosine distance, and multi-viewpoint cosine similarity.**

Keywords*—Multi-viewpoint based metric; traditional topic models; hybrid topic models; topic visualization; health tendency

I. INTRODUCTION

Twitter, Facebook, and microblogs [21], [22], [23] reveals the opinions of public and assessment of this social data [13], [14] is an emerging need in the applications like topics detection [1], [4], product promotion in business [4], political predictions [6], and health recommendations [2], [6]. Rapid urbanization is posing the number of public health-related problems, including accidents and injuries, healthcare disparities, and increasing disease burdens due to changes in lifestyle and nutrition, as well as increased environmental pollution. Twitter data contain sufficient health-related information, which dynamically updates millions of user posts for various health topics with all polarities of information [18]. The rate of flow of information increased because of tweets and retweets. In this paper, relationships between tweets as the interactions based on users (mentions), actions (reply and retweet), and contentment similarity are defined. A massive amount of data is generated, which makes it more difficult to extract topics [3] from these text documents manually. Topic modeling is an approach, which extracts implicit topics from document sets. There are several independent topics with different topic probabilities, and topics are consisting of words with different word probability. People frequently search for health-related topics in social media for envisaging solutions towards healthcare. Document clustering has become an increasingly important technique for unsupervised document organization [5], automatic topic extraction, and fast information retrieval [35], [38], or filtering. Clustering techniques [8], [17], used based on topic classification, namely, LSI [27], PLSI [29], LDA [26], supervised LDA [66], multi-class sLDA [67], medLDA [68], NMF [19], [36] and intJNMF [37], [40], sNMF [7] are specifically used for text mining. Thes methods are most suitable for post clustering and cannot able to determine the pre-cluster tendency in such cases; there is possible to get poor health clustering results. Bezdek et al. proposed [9], [20], [15] visual access tendency (VAT) for assessment of health data, which imposes good health topics (or clusters) visually. VAT use the Prim's logic for re-ordering the dissimilarity features for the set of data objects and saves this data in the matrix of reordered dissimilarity matrix (RDM) [15] and shows the visual clusters while displaying the image of RDM. The final image of RDM is known as VAT Image that gives the clarity of visualized square-shaped dark blocks along the diagonal for representing the clusters. We aim to investigate which public health issues are discussed in social media and in particular Twitter, and we use both VAT and traditional topic models in the proposed hybrid framework to overcome the problem of health cluster tendency, these hybrid topic models are VNMF, VLDA, VLSI, and VPLSI. Euclidean, cosine based, and multi-viewpoint cosine based metrics are used for finding similarity features between tweets documents in proposed models. Text documents similarity features are very sparse and high dimensional. Cosine based metric computes the distance between objects concerning magnitude and direction of document vectors. Hence, it strongly supports topics clustering than Euclidean. In traditional similarity/dissimilarity uses only a single viewpoint, which is the origin. In a multi-viewpoint cosine similarity based metric, we used many different viewpoints; objects assumed not to be in the same cluster. Using this more accurate assessment of how close or distant a pair of points if we look at them from many different viewpoints and average of similarities

measured relatively from the views of all other documents outside that cluster. The overall similarity is determined by taking an average over all the viewpoints not belonging to the cluster. Multi-viewpoints cosine similarity offers a more informative assessment of similarity than the single-origin point-based similarity measure. Our work is useful in the development of smart healthcare applications for recognition of health problems [31], [32] based on symptoms, and produces social recommend solutions for each specific health problem. The key contribution of our work is summarized as follows:

1) Pre-processing of tweets is initiated for removing of unwanted symbols, URLs, and stemming on words for maintaining uniformity in text analysis.

2) Estimation of health topics tendency is done.

3) Visual clusters are developed for these health topics.

4) Hybrid topic cluster models are proposed for topics tendency and health cluster analytics.

5) Proposed techniques give a solution for health classification and recommended solutions for health problems.

The remaining part of the paper is organized as follows: Section 2 presents the related work of topic models; Section 3 describes the proposed hybrid topic models; Section 4 discusses the experimental study; Visual cluster analysis and discussion are presented in Section 5. Finally, Section 6 presents the conclusion and future scope of the work.

II. Related Work of Topic Models

With minimal human intervention, topic models can discover prominent topics in the text without specific knowledge of the dataset and used to derive clusters. Topic models have a wide range of applications especially in the fields of text mining and information retrieval. Probabilistic models describe the topics distributions of terms and subjects, whereas non-probabilistic models used for describing the topics from the importance of terms.

Latent semantic indexing (LSI) [27] useful for defining of term-document matrix with latent semantic factors that use the singular value decomposition (SVD) [28], [29]. Post cluster tendency methods [16], [17] are usually considered as lazy topic learners. To overcome this problem, pre-cluster tendency methods preferred in cluster tendency assessment. Normalized cut (Ncut) described as Ncut-NMF [24], Xiaohui Yan *et al.* employed alternating non-negative least squares (ANLS) [25] algorithm for Ncut-NMF. Latent Dirichlet Allocation (LDA) [26], a Bayesian three-level hierarchical statistical model of PLSA implements posterior distribution using Bayesian inferences. Topics derived from interactions between tweets and re-tweets in the intJNMF method [49]. For visual assessment of clusters VAT and cosine VAT (cVAT) [18] extracts the dissimilarity features of documents concerning topic features and stored in dissimilarity matrix DM. The crisp partition matrix [17] is a broad approach of visual techniques, and it discovers document clustering results to topics, proposed as visual topic models in this paper. In this paper, we used a broad approach of VNMF, VLDA, VLSI, and VPLSI for the automation of tweets clustering of health tweets with prior knowledge of cluster health tendency and is presented in the proposed framework.

A. Similarity and Clustering Documents

Similarity metrics play an essential role in the success or failure of a clustering method. The effectiveness of the clustering algorithm depends on the appropriateness of the similarity measure to the data available. Documents comparison concerning topics features is performed with either distance or similarity metrics. Smaller the distance between two objects, the more similar they are to each of the feature vectors. Euclidean distance is the most common metric for computing similarity features between documents, used in most of the traditional topic modeling algorithms computed as in "(1)".

$$\text{D(Twn, Twm)}=\sqrt{(x1-y1)^2 \pm -- \pm (xz-yz)^2} \quad (1)$$

Cosine similarity is the best suitable when the data is in high-dimensional and sparse. It is a big similarity score in text mining and information retrieval. Cosine measure is used in spherical k-means algorithm [45], min-max cut graph-based spectral method [46], average weight [47], normalized cut [48] and document clustering using pairwise similarity score [49], [50]. In [51], Strehl et al. compared four measures Euclidean, cosine, Pearson correlation, and extended Jaccard, and concluded cosine and Jaccard are the best ones on web documents. Lenco et al. [52] introduced a similar context-based distance learning method for categorical data. Lakkaraju et al. [53] employed a conceptual tree-similarity measure to identify similar documents. Chim and Deng [54] proposed a phrase-based document similarity. The cosine metric uses both the magnitude and direction of the vectors. Similarity features between two tweet documents computed using "(2)".

$$\text{Cosine (Twn, Twm)} = \frac{Twn.Twm}{\|Twn\|\|Twm\|} \quad (2)$$

Multi-viewpoint based cosine similarity: We have a more accurate assessment of how close or distant a pair of points is if we look at them from multi-viewpoints computed and this similarity measure is called as multi-viewpoint based cosine similarity, and it is shown in "(3)".

$$\text{Sim}\left(d_i, d_j\right)_{d_i, d_j \epsilon S_r} = \frac{1}{n-n_r} A \quad (3)$$

Where $A = \sum_{d_h \epsilon S \setminus S_r} Sim(d_i - d_h, A = \pi r^2 - d_h)$

The compatible visual hybrid topic models with Euclidean, cosine based, and multi-viewpoints cosine based metrics for assessment of twitter health topics is described as follows:

Step1: The term-document matrix is transformed into a topic-document matrix with a size of n x m, in which n and m describe the number of topics and documents, respectively.

Step2: Dissimilarity matrix (DM) is computed using distance metrics, and it is shown in the following matrix.

$$\text{DM} = \begin{pmatrix} \text{T11} & \dots & \text{T1m} \\ \dots & \dots & \dots \\ \text{Tm1} & \dots & \text{Tmm} \end{pmatrix}$$

Step3: Reorder the dissimilarity features of tweet documents using the procedure of VAT [9], which resulting matrix is RDM.

Step4: The RDM matrix is normalized within the scale of 0 to 1, then display the RDM grayscale image for visualizing topic clusters, in which the topic clusters are shown as square-shaped dark colored blocks.

The visual images were displayed with Euclidean, cosine based distance, and multi-viewpoint based distance RDMs, more clarity of visual topics are shown in multi-viewpoints cosine based RDM. The proposed framework uses the topic derivation techniques along with visual models for deriving clusters estimations health tweets and explore the health topics based classification results. Respective proposed models are illustrated in the following section.

III. Proposed Hybrid Topic Models

For clusters estimations, visual models show an impressive result for unlabeled datasets. Traditional models include PLSI, NMF, LDA, and intJNMF are popular in the determination of topics clustering results. It is required to determine the prior classification of social health data clusters information; which informative assessment produces the quality of topics clusters. In such cases, users guessing is not required for topics clustering models; thus, there is a broad scope to improving the efficiency of topics cluster models. A combination of visual model (i.e., VAT) and topic models are proposed, by nature, these are known as hybrid models, which are Visual LSI (VLSI), Visual PLSI (VPLSI), Visual NMF (VNMF), and Visual LDA (VLDA). Proposed hybrid models compute the low-rank matrices for active learning of topics of tweets documents. The low-rank matrices are less sparse that can be used for representing the related topics and tweets documents rather than the relationship between terms and tweet documents. The proposed VNMF, VLDA, VPLSI, and VLSI exploit the relationships between documents and ignores retweet interactions. The VLSI uses the singular value decomposition (SVD) for finding the document scores for topics. The dissimilarity features between documents are computed based on document scores.

Algorithm 1: VLSI

Input: X, term-document matrix in w-d form
Output: Number of Topics (or Clusters)
Method:

1. Compute the document scores ‘V’ for given X using SVD [ ] approach.
2. Compute the Dissimilarity Matrix (DM) of V
3. Find the Reordered Dissimilarity Matrix (RDM) of V using [15] and display image(RDM).
4. Assess a good number of topics from the image of RDM

The term-document matrix 'X' is constructed for the set of tweet documents, and it is taken as input in algorithm 2 (VPLSI). Step 1 computes the prior probability P(d, w) with the probabilities of P(d) and P(w|d), whereas P(d) and P(w|d) describes the initial probability, and conditional independence probability respectively. The values of P(w|z), P(z), P(d|z) are re-calculated by EM procedure [30], which support for extraction of certain topics distributions from the set of tweet documents. Probabilities of topic-document are stored in the variable of V. Re-estimations of probabilities are described in Step 3. The topic-document probabilities are stored into matrix V. Step 4 computes the dissimilarity matrix 'DM' using distance metrics (Either Euclidean or Cosine). Step 4 uses the Prim's logic for dissimilarities of data objects and finds the result in another matrix, known as re-ordered dissimilarity matrix (RDM). Step 6 finds the image of RDM, which shows the clusters in the shape of square dark colored blocks. Estimations of clusters are performed by the assessment of square-shaped dark colored blocks from the image of RDM. All these steps are described in the proposed VPLSI, which shown as follows.

Algorithm 2: VPLSI

Input: X, term-document matrix in w-d form
Output: Number of Topics (or Clusters)
Method:

1. Compute the prior probability $P(d, w) = P(d)P(w|d)$ with
2. Conditional probability independence $P(w|d) = \sum_{z \epsilon Z} P(w|z)P(z|d)$
   Re-estimate the parameters using EM algorithm, and
   Update the value of P(d,w). where$P(d, w) = \sum_{z \epsilon Z} P(w|z)P(z)P(d|z)$
3. Store the values of P(d|z) into V
4. Compute Dissimilarity Matrix (DM) of V
5. Reordered Dissimilarity Matrix (RDM) of DM.
6. Assess a good number of topics from the image of RDM

The proposed VPLSI uses the term-documents [10] and related parameters of EM for determining the topic-document matrix. The terms-correlation features are not considered in the estimation of topic parameters. These features are essential in the determination of topics of tweet documents. For this reason, another visual topic model VNMF is proposed; it initially finds the topic-document matrix with consideration of the term-correlation matrix 'S.' It derives the useful conceptual (or hidden) topics before deriving the term-topic matrix ‘U’. The steps of VNMF are shown in Algorithm 3.

Algorithm 3: VNMF

Input: X, Term-Document Matrix
S, Term-Correlation Matrix
Output: Number of Topics (or Clusters)
Method:

1. Term-topic matrix U minimizes the function L(U), which satisfies L(U)= $\|S - UU^T\| >= 0$
2. Apply convergence for finding $U = SU(U^TU)^{-1}$
3. Apply convergence for finding topic-document matrix V using $V = (U^TU)^{-1}U^TX$
4. Compute Dissimilarity Matrix (DM) of V
5. Compute Reordered Dissimilarity Matrix (RDM) of V using [15]
6. Assess the suitable number of topics from the image of RDM.

Term-document X shows the association between tweet documents and their respective terms. Correlations between the terms are also taken as input (S) of VNMF. Step 1 shows the term-topic matrix, which satisfies the objective function L(U). Step 2 and 3 show the optimal term-topic matrix 'U' derivations with convergence and derives the optimal topic-document matrix V. In V; the documents are denoted as document vectors. Step 4 computes the distances between document vectors using distance metrics and stored in dissimilarity matrix (DM). In Step5, the values of DM are reordered according to Prim's logic, and final re-ordered values are stored in RDM and procedure given in [11]. In step 6, the RDM image is displayed that shows the visual clusters through dark-colored square-shaped blocks. Each square-shaped dark colored block shows an individual crisp partition for a similar set of objects. With the crisp partitions of square-shaped blocks, the cluster labels of objects are predicted which gives the complete topics clustering results.

Algorithm 4: VLDA

Input: X, Term-Document Matrix
Output: V, Topic- Document Matrix, Number of Topics (or Clusters)
Method :

1. Generate multiple topics P(W) with the likelihood
2. Estimate word distribution and topic convergence distribution
   using Dirichlet distribution by enforcing the sparse conditions.
3. Find an optimal solution using the EM algorithm
   3.1 E-Step: Calculate Posterior probability using Bayesian inferences
   3.2 M-Step: Re-estimate probability of document covering topics and update the values into 'V'
4. Compute Dissimilarity Matrix (DM) of V.
5. Reordered Dissimilarity Matrix (RDM) of V using [15].
6. Assess the good number of topics from the image of RDM.

Another topic model, say, LDA finds the Dirichlet coefficients [39] for various topics while performing topics clusters for the set of tweet documents. It uses EM concept to update the probability topics of documents and saves the values into V. Dissimilarity values of V and RDM are computed for further assessment of a number of clusters and clustering objects for the topics of tweet documents. All these steps are described in Algorithm 4.

IV. Experimental Study

Baseline topic models [12] are enhanced as hybrid topic cluster models to separate topics from tweet documents visually. Proposed models derive topics or clusters with pre-clustering techniques and visually represented. Here we combine traditional topic models with VAT. We come up with hybrid topic models with Euclidean, cosine based and multi-viewpoint cosine similarity measures for extracting and visualizing topic clouds and health tendency of different health topics from the benchmark and Twitter datasets.

A. The Architecture of Hybrid Topic Models

Fig. 1 shows the architecture of proposed hybrid topic modeling. It supports the entire knowledge discovery procedure, including analysis, inference, evaluation, and applications for health data clustering. In the data layer, read the health-related data of 2-topics to 20-topics, TREC2014, and TREC2015 health-related keyword phrases are considered in the extraction of health data from Twitter. In the processing layer, it provides an implementation of VNMF, VLDA, VintJNMF, VPLSA, and VPLA topic models. Besides that, this architecture presents three aspects of how to evaluate the performance of proposed algorithms by using internal validity indexes, external validity indexes, and computational complexity. We used internal indexes DB, CHI, SI, XI, PC, PEI, and SM to measure given clustering structures, and external validity indexes CA, NMI, Precision, Recall, and F-Score to measure the fitness of data and expected structure. The total number of iterations for convergence, time for execution (in seconds), and allocated memory (in Kbs) are taken as measures in the evaluation of topic models.

B. Datasets Description

The experiments are executed with i7 processor, 16 GB RAM, under MATLAB 2019. Implementation details are as follows: Tweets documents and benchmark datasets are extracted with keyword phrases, preprocessed with the NLP tool. VAT used for assessment of the number of topics from tweet documents. Derived topic-document matrix is treated as input to partition matrix and derived VAT images for topics clustering of social data.

Subsets of labeled social health-related data sets from 2-topics to 20-topics are extracted from Twitter. Collected documents consist of tweets, and retweets related to healthcare are preprocessed by standard NLP tools for removing irrelevant data in tweet documents. Datasets used in our study are described in [61]. The standard bench-marked health key phrases of TREC2015 [33] and TREC2014 [34] are used for the extraction of relevant health tweets phrases are mentioned in [61].

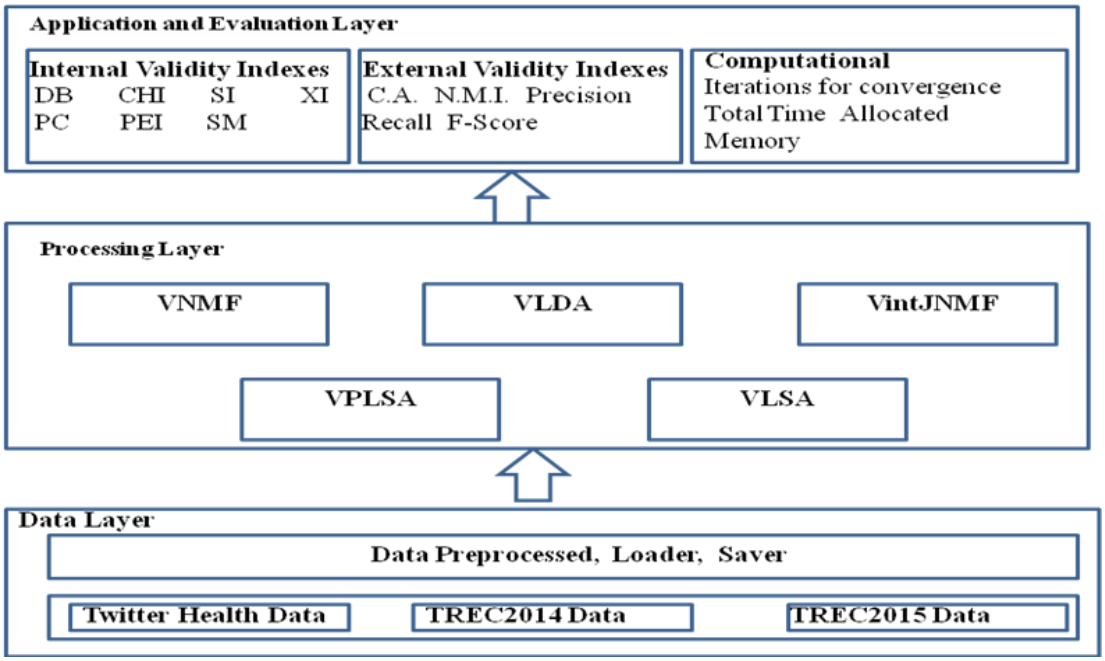


Fig. 1. Architecture of Hybrid Topic Models.

C. Features of Hybrid Topic Algorithms Comparison

Hybrid topic cluster modeling algorithms are compared for different datasets with three factors i.e., data set size, the number of clusters, and data set type. Datasets of [61] shows the details of these features of datasets used in our experiments and the respective result analysis shown in the following sections. For each feature, four tests are made, one for each algorithm. Results are represented in the form of tables and graphs in the following sections.

V. Visual Cluster Analysis And Discussion

A. Topics Clouds Description

Word cloud is a tool to describe results and trending keywords with the visual representation of word content commonly used to represent user-generated content in the respective health-related topic. It depicts the most frequently utilized words, the most important word is the biggest in size, and likewise, the size of the word decreases with each word's frequency correlated with font size. Fig. 2 shows the sample experimental word clouds formation of health data sets from Twitter, TREC2014, and TREC2015 keyword phrases data sets with following generalized steps.

1) Text is extracted from collected tweet documents.

2) Bag of words or bag of topics is extracted from preprocessed tweets.

3) Infrequent words are removed from these bag of words or bag of topics.

4) By using hybrid topic models, relevant words are found and represented as a word cloud based on the number of health-related topics.

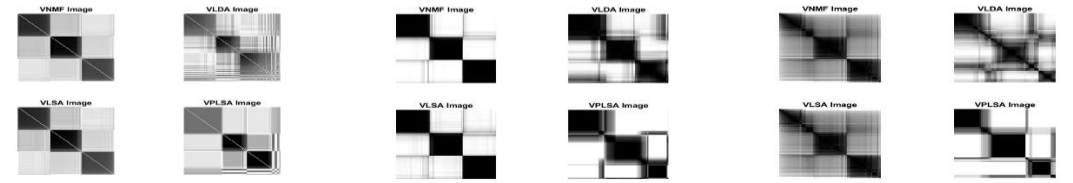
5) Words, along with the size depicting intensity, are colorfully-plotted.

B. Assessment of Health Tendency

The visual evidence of topics clusters i.e., VAT images for the proposed models is shown in Fig. 3 for health tweets of TREC2015-3 keyword, TREC2015-4 keywords, 5-topics, 15-topics, and 20-topics respectively. The VAT images show the visual assessment of topics clusters for VNMF, VLDA, VLSA, and VPLSA using Euclidean, cosine based and multi-viewpoint cosine similarity measures, in which, every individual topic represented as dark square-shaped blocks along the diagonal. The quality of clusters is recognized with more clarity of square-shaped dark-colored in VAT Images. From the visual evidence of Fig. 3, it is observed that more clarity of visual topics is found in most of the health-related datasets using a multi-viewpoint cosine metric than Euclidean and cosine based metrics. Several topics increase in some cases; more clarity is observed under cosine based metrics. Significance of these visual results stated that VNMF, VLSA, and VPLSA efficiently performed for detection of health topics cluster tendency in healthcare applications and observed that VLDA shows the less clarity of visual results when compared to other models.

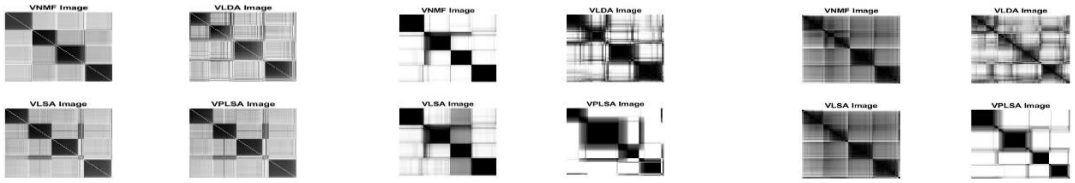
Word cloud of 2-topics data Word cloud of 4 keyword phrases

Fig. 2. Sample Word Clouds of Health Data Sets.



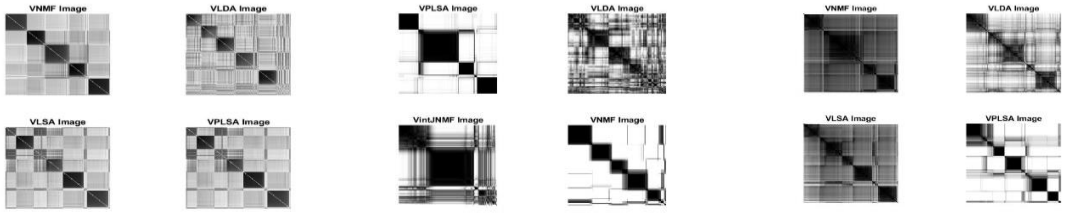
Multi-viewpoint Cosine Metric Cosine Metric Euclidean Metric

TREC2015 Three Keyword Phrases



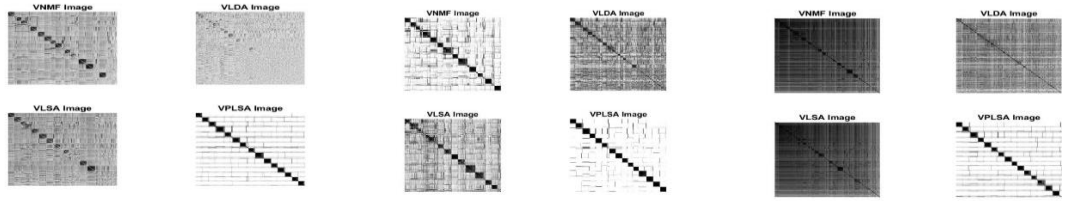
Multi-viewpoint Cosine Metric Cosine Metric Euclidean Metric

TREC2015 Four Keyword Phrases



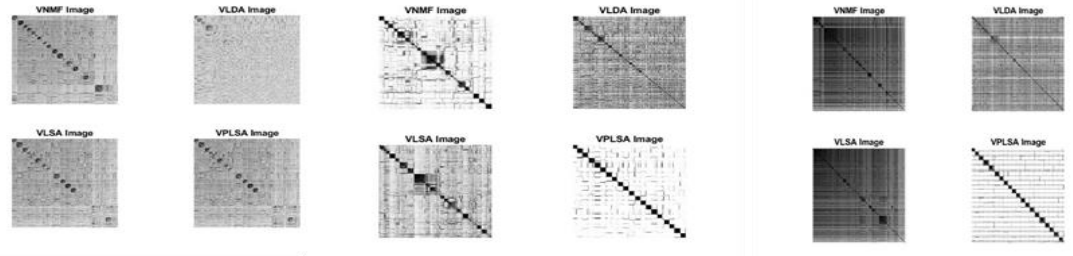
Multi-viewpoint Cosine Metric Cosine Metric Euclidean Metric

5-topics health tweets datasets



Multi-viewpoint Cosine Metric Cosine Metric Euclidean Metric

15-topics health tweets datasets



Multi-viewpoint Cosine Metric Cosine Metric Euclidean Metric

20-topics health tweets datasets

Fig. 3. Visual Clusters of Health Tweets Datasets.

C. Performance Measures Evaluation

Our proposed mechanisms use both traditional topic models and visual technique i.e., VAT to find cluster tendency and to represent visually. A combination of VAT and topic models are used in our proposed framework (hence, they called as VNMF, VLDA, VLSI, and VPLSI) and experimented with Euclidean, cosine based and multi-viewpoint cosine based metrics. Evaluation of proposed techniques are measured with five external validity indexes, namely, clustering accuracy (CA) [41], normalized mutual information (NMI) [42], precision (P), recall (R) and F-Score (F) [43], [44] and seven internal validity indexes viz., Davies-Bouldin index (DB) [55], [56], [60], Calinski-Harabasz Index (CHI) [55], [56], Silhouette Index (SI) [55], [56], Xie-Beni Index (XI) [57], Partition Coefficient (PC) [59], Partition Entropy Index (PEI) [57], [58], and Separation Measure (SM) [60]. Health tweets are assigned to topic clusters which are maintained the highest similarity with the topic clusters to improve the value of CA. Proposed visual topic models properly compute the similarity computations. NMI [42] computes the cluster accuracy by computation of mutual information I(W;C) divided by the clusters W and classes C. Here mutual information I(W;C) statistically computed by Eq. (4).

$$I(W;C) = \sum_k \sum_j P\left(w_k \cap c_j\right) log \frac{P(w_k \cap c_j)}{P(w_k)P(c_j)} \quad (4)$$

Variables k and j denote the number of clusters in W and C. w_k represents the cluster at index k from the topic clusters W and CJ represent the specific topic cluster with index k measured in evaluated clusters by proposed topic models. Precision (P) computes the fraction of relevant topics objects among extracted topics objects; Recall (R) computes the fraction of relevant topics objects among total topics objects. Equations of precision and recall are shown in"(5)" and "(6)" respectively.

$$P = \frac{TP}{TP+FP} \quad (5)$$

$$R = \frac{TP}{TP+FN} \quad (6)$$

Whereas TP refers to the true positives, TN refers to the true negative, FP refers to the false positive, and FN refers to the false negative. F-measure (F) computes both precisions and recall harmonic means, and it is shown in "(6)"

$$F = 2\, x \frac{PxR}{P+R} \quad (7)$$

External validity indexes, i.e., CA, NMI, P, R and F-Score of 5-topics, 10-topics, 15-topics, and 20-topics health data tweets of proposed visual topic models are shown in Table I. VLSA and VPLSA under multi-viewpoint cosine distance similarity measure shows better performance than other hybrid models in case of CA and NMI external validity indexes. VNMF under multi-viewpoints performs better in precision (P), recall (R), and F-Score (F). As number topics are increased, external validity index values are decreased; however, except VLDA other three models under multi-viewpoint cosine metric maintains good clustering accuracy values.

Internal validity indexes, i.e., DB, CHI, SI, XI, PC, PEI, and SM of 5-topics, 10-topics, 15-topics, and 20-topics of proposed visual topic models are shown in Table I. It is observed that in case of DB internal validity index VLSI, VPLSI under multi-viewpoint cosine shows better performance, as the number of topics increases VNMF under multi-viewpoint shows better results. CHI and SI internal validity indexes perform well under multi-viewpoints in VLSA and VPLSA models, whereas XI, PEI, and SM under Euclidean distance perform well for VNMF and PC results are scattered under Euclidean distance. It is proved that the multi-viewpoint cosine metric is greatly succeeded in text clustering as per the overall observations of both external and internal validity indexes performance measures.

External validity indexes, i.e., CA, NMI, P, R and F-Score and internal validity indexes DB, CHI, SI, XI, PC, PEI and SM of TRC2014 and TRC2015 Health keyword phrases datasets of proposed visual topic models are shown in Table II and Table III respectively. It is observed that for both.

TREC2014 and TREC2015 datasets all external indexes perform well under multi-viewpoint cosine similarity for VNMF and VLSA models and all internal indexes except XI and CHI perform well under cosine metric similarity for VNMF.

Fig. 4 shows the comparative results with a bar graph of external validity indexes for 4-topics, 8-topics, 12-topics, and 20-topics health tweets, TREC2014 4-keyword, 3-keyword, and TREC2015 4-keyword phrases, 5-keyword phrases. Fig. 5 shows comparative results of internal indexes with line graphs of 18-topics, 20-topics, and TREC2014 and TREC2015 3-keyword and 4-keyword phrases. It shows the performance of hybrid visual topic models with Euclidean, cosine based and multi-viewpoint cosine based comparative metric analysis of VNMF, VLDA, VLSA, and VPLSA models. From this analysis, it is confirmed that VLSA, VPLSA and VNMF is capable of clustering health tweets with a better accuracy rate under multi-viewpoint cosine based similarity metrics. Comparative analysis shows that better performance values are obtained in VLSA, VPLSA, and VNMF visual topic models are more suitable topic models for accessing topics and for discovering complete clustering results of health datasets.

TABLE. I. EXTERNAL AND INTERNAL INDEXES OF HEALTH TWEETS DATASETS

Tweets Dataset C.A.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5700	0.3550	**0.9760**	**0.9760**	0.8650	0.4100	0.9000	0.3100	0.5700	0.3550	0.9750	0.9750
10-Topics	0.6600	0.2825	**0.6700**	**0.6700**	0.6000	0.3075	0.6225	0.2550	0.6600	0.2825	0.6600	0.6600
15-Topics	**0.5250**	0.2000	0.4917	0.2100	0.5217	0.2233	0.5150	0.1917	**0.5250**	0.2000	0.4917	0.2100
20-Topics	0.4325	0.1725	**0.4800**	**0.4800**	0.4775	0.2475	0.5550	0.2013	0.4325	0.1725	**0.4800**	**0.4800**
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5052	0.1080	**0.9318**	**0.9318**	0.7836	0.2077	0.9000	0.0949	0.5560	0.2009	0.8154	0.0941
10-Topics	0.5834	0.1501	**0.6098**	**0.6098**	0.5534	0.2321	0.6004	0.1630	0.4949	0.1891	0.4722	0.1602
15-Topics	**0.4977**	0.1487	0.4813	0.1552	0.4871	0.1641	0.4846	0.1443	0.3824	0.1743	0.3294	0.1457
20-Topics	0.4907	0.1545	**0.5211**	**0.5211**	0.5181	0.2452	0.6063	0.1976	0.3813	0.2401	0.3548	0.2107
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8825	0.4679	**0.9750**	0.3180	0.8825	0.4679	1.0000	0.3259	0.8825	0.4679	0.9200	0.3259
10-Topics	**0.7655**	0.3827	0.6700	0.2410	0.6798	0.5083	0.7225	0.2842	**0.7655**	0.3827	0.5075	0.2564
15-Topics	**0.6459**	0.3156	0.4917	0.2183	**0.6459**	0.6359	0.5550	0.2078	**0.6459**	0.6359	0.3567	0.2078
20-Topics	**0.5850**	0.5650	0.4800	0.2381	0.6235	0.3117	0.5550	0.2089	0.5489	0.3591	0.3012	0.2170
Recall (R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8400	0.4850	0.9750	0.3200	0.8400	0.4850	1.0000	0.3250	0.8400	0.4850	0.9200	0.3250
10-Topics	**0.7700**	0.4000	0.6700	0.2450	0.7250	0.5000	0.7225	0.2850	**0.7700**	0,4000	0.5075	0.2625
15-Topics	**0.6733**	0.3100	0.4917	0.2233	0.6633	0.6533	0.5550	0.2100	**0.6733**	0.6233	0.3567	0.2100
20-Topics	**0.6125**	0.6025	0.4800	0.2463	0.6088	0.3387	0.5550	0.2162	**0.6487**	0.3725	0.3012	0.2225
F-Score (F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8297	0.4673	0.9750	0.3173	0.8297	0.4673	**1.0000**	0.3246	0.8297	0.4673	0.9200	0.3246
10-Topics	**0.7609**	0.3833	0.6700	0.2414	0.6867	0.4908	0.7225	0.2837	**0.7609**	0.3833	0.5075	0.2564
15-Topics	**0.6534**	0.3023	0.4917	0.2201	**0.6534**	**0.6534**	0.5550	0.2078	**0.6534**	**0.6534**	0.3567	0.2078
20-Topics	**0.6642**	0.4654	0.4800	0.2411	0.6542	0.3181	0.5550	0.2115	**0.6422**	0.3540	0.3012	0.2189
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	10.7540	64.4572	**3.4784**	**3.4784**	3.4804	11.0546	2.3716	14.4076	9.2761	21.0979	4.3503	14.3609
10-Topics	18,0747	48.0752	**14.8103**	**14.8103**	20.3867	27.7623	18.6712	17.8868	17.6975	37.4516	18.8138	23.0892
15-Topics	**26.4144**	83.4127	31.2064	31.6627	26.5696	31.1415	26.5796	27.8605	32.7241	55.3354	39.5790	35.3903
20-Topics	**33.8310**	82.9828	37.5860	37.5860	19.8077	36.7545	41.6406	34.7402	35.1299	63.0640	41.5584	39.1396
C.H.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	7.6755	0.5287	**189.750**	**189.750**	48.6197	2.8234	550.3889	2.4142	6.1710	2.8001	30.8947	3.0535
10-Topics	13.5893	1.6944	**20.5715**	**20.5715**	15.9857	2.0776	20.3201	3.4162	14.1817	1.5947	20.5714	3.0896
15-Topics	**8.5933**	1.4371	6.6995	2.1976	7.6664	1.4882	7.7097	2.1805	3,5418	1.3641	2.9697	1.9680
20-Topics	9.2787	1.0614	**10.5519**	**10.5519**	9.9539	2.0615	2.9526	2.6605	7.0669	1.1551	7.7888	2.0700
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	-0.0766	-0.0732	**0.7560**	0.4328	0.2717	-0.0940	-0.0912	-0.0872	-0.0883	-0.0852	-0.4355	0.3462
10-Topics	**0.0207**	-0.0966	0.0804	0.0804	-0.0090	-0.0928	-0.1038	-0.0954	-0.2759	-0.1170	-0.1012	-0.2012
15-Topics	**-0.0790**	-0.0959	-0.0979	-0.8642	-0.0859	-0.0986	-0.0997	-0.0874	-0.3685	-0.1811	-0.0929	-0.2432
20-Topics	-0.1246	-0.1932	**-0.1140**	-0.2243	-0.1976	-0.1162	-0.1174	-0.1246	-0.3277	-0.1814	-0.1898	-0.1124
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	17.7659	472.6124	274.7771	274.8100	21.0856	24.0764	32.1530	21.2387	0.0399	0.0426	**0.0373**	1.1370

10-Topics	5.3797	0.0006	**0.0002**	0.0006	0.0462	0.7154	0.0007	0.1338	0.0334	0.0957	0.0088	2.3679
15-Topics	0.5134	0.9549	0.3745	5.1716	0.0211	0.5424	0.5649	0.5300	**0.0018**	0.0394	0.0022	1.2323
20-Topics	1.2912	0.2041	1.5867	0.5031	0.1529	1.1463	0.0203	0.2374	**0.0027**	0.0148	0.0057	3.6486
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.7349	0.4414	0.7511	0.7511	0.7688	0.3464	0.8345	0.9024	0.5802	0.3133	0.5133	**0.9084**
10-Topics	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	**0.2895**	0.1033	0.2624	0.1000
15-Topics	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.1688	0.0846	**0.1713**	0.0667
20-Topics	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	**0.1622**	0.0592	0.1682	0.0500
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5935	1.1376	**0.5757**	**0.5757**	0.4937	1.3136	0.3898	0.6504	0.8667	1.3735	1.0116	0.6023
10-Topics	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	**1.6654**	2.2841	1.7323	2.3026
15-Topics	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	**2.1304**	2.5716	2.1212	2.7081
20-Topics	2.9957	2.9957	2.9957	2.8857	2.9957	2.9957	2.9957	2.9957	2.3842	2,8926	**2.3037**	2.9957
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.3370	0.5471	0.0748	0.0748	0.1064	0.3627	0.0320	0.0196	0.8494	1.3215	0.4773	**0.0213**
10-Topics	0.6133	0.0443	0.5459	2.2830	0.0008	0.0000	0.0012	0.0014	0.0076	0.0230	**0.0007**	5.7261
15-Topics	0.6538	0.0619	0.0377	2.8245	0.0033	0.0002	0.0034	0.0044	**0.0001**	0.0010	0.0034	1.8791
20-Topics	1.2911	0.0850	3.0754	0.3444	0.0357	1.8886	0.5004	0.2595	**0.0002**	0.0054	0.0065	3.9345

TABLE. II. EXTERNAL AND INTERNAL INDEXES OF TREC2014 HEALTH KEYWORD PHRASES DATASETS

TREC2014 C.A.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.6250	**1.0000**	0.9750	**1.0000**	0.7500	**1.0000**	0.9750	0.9750	0.7000
3Keyword	**1.0000**	0.8750	**1.0000**	0.4083	**1.0000**	0.9083	**1.0000**	0.4833	0.9833	0.8917	0.9833	0.4833
4Keyword	0.9812	0.4500	0.9750	0.4063	**1.0000**	0.7250	**1.0000**	0.4500	0.8500	0.8250	0.9688	0.4437
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.0456	**1.0000**	0.8313	**1.0000**	0.1887	**1.0000**	0.8313	0.8313	0.1187
3Keyword	**1.0000**	0.6285	**1.0000**	0.0408	**1.0000**	0.7169	**1.0000**	0.0906	0.9291	0.6875	0.9291	0.0763
4Keyword	0.9368	0.2184	0.9157	0.1452	**1.0000**	0.4395	**1.0000**	0.1530	0.6362	0.5839	0.9011	0.1616
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5833	**1.0000**	**1.0000**	**1.0000**	0.7941	**1.0000**	**1.0000**	**1.0000**	0.8148
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4047	**1.0000**	**1.0000**	**1.0000**	0.4602	**1.0000**	**1.0000**	0.9833	0.4602
4Keyword	0.9939	0.9939	0.9750	0.3982	0.9939	0.9939	**1.0000**	0.4413	0.6709	0.6709	0.9688	0.4865
Recall(R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5250	**1.0000**	**1.0000**	0.9750	0.6750	**1.0000**	**1.0000**	0.8750	0.5500
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4083	**1.0000**	**1.0000**	**1.0000**	0.4583	**1.0000**	**1.0000**	0.9833	0.4583
4Keyword	0.9939	0.9939	0.9750	0.4000	0.9938	0.9938	**1.0000**	0.4437	0.7063	0.7063	0.9688	0.5000
F-Score(F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5526	**1.0000**	**1.0000**	0.9873	0.7297	**1.0000**	**1.0000**	0.9333	0.6567
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4054	**1.0000**	**1.0000**	**1.0000**	0.4581	**1.0000**	**1.0000**	0.9833	0.4581
4Keyword	0.9937	0.9937	0.9750	0.3984	0.9937	0.9937	**1.0000**	0.4404	0.6561	0.6561	0.9688	0.4899
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.8600	0.8573	0.8712	3.3995	**0.6901**	0.7650	0.6906	1.2297	0.9290	2.8748	0.9324	2.0255
3Keyword	1.6057	2.8889	1.7053	24.8802	1.3063	1.5675	**1.3167**	4.1082	1.8455	2.1106	1.8785	5.9762
4Keyword	2.3388	19.3556	2.6151	7.9024	**1.8554**	3.8763	1.8750	6.1840	3.5706	3.9035	2.8485	5.4655

C.H.I.	Multi-viewpoint Cosine				Cosine				Euclidean			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	3.7917	1.4453	3.6289	0.0042	3.9685	0.0265	2.2770	0.0023	404.5580	378.7549	**405.7344**	8.5830
3Keyword	1.3269	0.0474	1.1660	0.0012	3.2002	0.0958	2.2716	0.0079	**195.2073**	72.2194	190.1012	6.4386
4Keyword	**346.6194**	3.4159	256.675	5.0847	1.7212	0.0184	1.2061	0.0060	34.3966	34.3527	104.3518	7.1884
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.9895	0.9661	0.0684	0.6542	**0.9989**	0.8004	0.8692	0.6542	0.8949	0.8591	0.0901	0.4326
3Keyword	0.9652	0.3941	-0.0605	0.7642	**0.9832**	0.5578	0.1658	0.1456	0.7648	0.4706	0.1531	0.5242
4Keyword	0.8520	-0.0693	-0.0743	0.6532	**0.9626**	0.1030	0.0658	0.0426	0.1637	0.2522	-0.0421	0.2436
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.3861	0.7182	0.3878	0.6867	**0.0386**	1.8074	0.0657	1.6163	1.9705	1.2353	3.6027	1.6387
3Keyword	115.3133	64.260	**2.1006**	5.3383	14.3863	20.4709	25.9446	17.6322	94.034	30.1596	40.0134	29.9649
4Keyword	0.0994	1.0878	0.0232	**0.0334**	0.5471	0.1518	1.3669	4.6510	481.1688	33.1056	155.6778	210.9156
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**0.9899**	0.9697	0.9894	0.9608	**0.9989**	0.9291	0.9982	0.9387	0.9221	0.9470	0.9449	0.9272
3Keyword	0.9452	0.8571	0.9383	0.9082	**0.9781**	0.8620	0.9689	0.9683	0.8514	0.8304	0.8476	0.9589
4Keyword	0.8843	0.6080	0.8709	0.8585	**0.9534**	0.7126	0.9340	0.9059	0.7382	0.6849	0.7703	0.8727
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.0314	0.0683	0.0326	0.0763	**0.0036**	0.1304	0.0062	0.0998	0.1401	0.1015	0.1068	0.1169
3Keyword	0.1385	0.2964	0.1532	0.1985	0.0488	0.2774	0.0731	**0.0647**	0.2869	0.3375	0.2949	0.0824
4Keyword	0.2808	0.7751	0.3087	0.3249	**0.1070**	0.5851	0.1546	0.1995	0.5342	0.6416	0.4751	0.2563
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.0378	0.0413	0.0393	0.0431	**0.0258**	0.0383	0.0269	0.0326	0.0840	0.0395	0.0899	0.0360
3Keyword	0.0453	0.0993	0.0475	0.0412	**0.0224**	0.0469	0.0259	0.0255	0.0936	0.0569	0.0956	0.0263
4Keyword	0.0527	0.1942	0.0629	0.0524	**0.0227**	0.0541	0.0269	0.0266	0.5563	0.0764	0.1530	0.0289

TABLE. III. EXTERNAL AND INTERNAL INDEXES OF TREC2015 HEALTH KEYWORD PHRASES DATASETS

TREC2015 C.A.	Multi-viewpoint Cosine				Cosine				Euclidean			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8750	**1.0000**	**1.0000**	**1.0000**	0.8500	**1.0000**	0.9500	**1.0000**	0.7750
3keyword	**1.0000**	0.7750	**1.0000**	0.7583	**1.0000**	0.7048	**1.0000**	0.4996	**1.0000**	0.8083	**1.0000**	0.5333
4keyword	0.9625	0.6625	0.8313	0.8313	0.9375	0.5563	0.9500	0.4813	**0.9688**	0.8063	**0.9688**	0.4313
5keyword	0.6700	0.5400	0.6800	0.6800	**0.9100**	0.6700	0.8700	0.3750	0.8250	0.6000	0.7650	0.4000
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.4564	**1.0000**	**1.0000**	**1.0000**	0.3902	**1.0000**	0.7136	**1.0000**	0.2308
3keyword	**1.0000**	0.4479	**1.0000**	0.4530	**1.0000**	**1.0000**	**1.0000**	0.4996	**1.0000**	0.4942	**1.0000**	0.1901
4keyword	0.8865	0.3864	0.6339	0.6339	0.8614	0.3127	0.8828	0.1813	0.9011	0.5739	**0.9011**	0.1552
5keyword	**0.5879**	0.3942	0.5645	0.5645	0.8119	0.4590	0.7591	0.1636	0.6456	0.4027	0.5858	0.2490
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8974	**1.0000**	**1.0000**	**1.0000**	0.8462	**1.0000**	**1.0000**	**1.0000**	0.7750
3keyword	**1.0000**	**1.0000**	**1.0000**	0.7283	**1.0000**	**1.0000**	**1.0000**	0.7608	**1.0000**	**1.0000**	**1.0000**	0.4943
4keyword	0.9661	0.9661	0.8312	0.4558	0.9637	0.9637	0.9500	0.4913	**0.9939**	**0.9939**	0.9688	0.4620
5keyword	0.6830	0.6830	0.6800	0.5407	0.9321	0.9321	0.8700	0.3949	**0.9366**	**0.9366**	0.7650	0.3993
Recall (R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8750	**1.0000**	**1.0000**	**1.0000**	0.8250	**1.0000**	**1.0000**	**1.0000**	0.7750

3keyword	**1.0000**	**1.0000**	**1.0000**	0.7333	**1.0000**	**1.0000**	**1.0000**	0.7583	**1.0000**	**1.0000**	**1.0000**	0.5250
4keyword	0.9625	0.9625	0.8312	0.4625	0.8625	0.9625	0.9500	0.4937	**0.9938**	**0.9938**	0.9688	0.4750
5keyword	0.7700	0.7700	0.6800	0.5500	0.9200	0.9200	0.8700	0.3950	**0.9250**	**0.9250**	0.7650	0.4150
F-Score (F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8861	**1.0000**	**1.0000**	**1.0000**	0.8354	**1.0000**	**1.0000**	**1.0000**	0.7750
3keyword	**1.0000**	**1.0000**	**1.0000**	0.7303	**1.0000**	**1.0000**	**1.0000**	0.7562	**1.0000**	**1.0000**	**1.0000**	0.4931
4keyword	0.9619	0.9619	0.8313	0.4562	0.9621	0.9621	0.9500	0.4883	**0.9937**	**0.9937**	0.9688	0.4649
5keyword	**0.9562**	0.9356	0.6800	0.5347	0.9201	0.9201	0.8700	0.3942	0.9258	0.9258	0.7650	0.4006
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.8268	0.8784	0.8784	1.0548	**0.6930**	0.7328	0.6939	0.9486	0.8741	0.8871	0.8791	1.5546
3keyword	1.7844	5.7403	1.8031	3.1419	**1.3141**	1.6395	1.3205	1.9608	1.9231	2.7790	1.9400	8.7497
4keyword	2.5904	8.0642	4.5376	4.5376	**2.2253**	6.1540	2.3094	10.2185	2.7870	3.9111	2.8485	7.6919
5keyword	12.2397	16.9242	14.3625	14.3625	**3.2036**	7.6354	3.6686	7.6100	4.7661	12.7339	6.1253	8.8636
C.H.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	3.1217	0.9943	0.8843	0.1132	8.2324	0.6959	0.4773	0.0604	**609.161**	197.676	598.464	20.0786
3keyword	1.1547	0.0125	1.1464	0.0354	3.1818	0.0884	2.1479	0.0329	**183.429**	35.5853	181.120	4.3831
4keyword	215.57	10.9328	44.850	44.850	166.944	7.7988	**220.252**	6.3097	107.242	30.7358	104.351	9.0711
5keyword	24.652	16.2431	18.384	18.384	**122.026**	30.8903	74.6247	5.7980	31.9652	8.3416	25.0822	8.3665
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.9868	0.9358	0.6325	0.7642	**0.9944**	0.9228	0.5117	0.4562	0.9357	0.7539	0.2693	0.4328
3keyword	0.9589	0.0844	0.3051	0.8642	**0.9841**	0.5253	0.2400	0.2326	0.7898	0.2884	-0.0279	0.3452
4keyword	**0.7600**	-0.004	0.3390	0.4562	0.6952	-0.0412	-0.0729	-0.0652	0.6528	0.2322	-0.0395	0.2346
5keyword	0.0526	0.0432	-0.0467	0.0456	**0.5608**	0.1192	0.3533	0.2542	0.1978	-0.0553	-0.0642	0.0568
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.3658	1.0215	1.0215	1.3430	**0.1731**	1.3944	0.2150	2.1933	1.4691	1.3686	1.4751	2.5907
3keyword	**1.8426**	30.1387	1.9038	8.7804	29.7477	102.25	46.6774	301.5592	61.8303	125.707	117.7518	29.9772
4keyword	13.8912	121.3485	**13.1335**	**13.1335**	354.4803	77.7253	168.7514	102.0204	147.0701	77.3134	155.6775	110.7712
5keyword	1.4976	0.0041	**0.0007**	**0.0007**	0.3305	0.8615	0.0404	2.0835	0.4388	0.0599	0.1613	5.8658
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.9871	0.9504	0.9504	0.9311	**0.9944**	0.9465	0.9930	0.9150	0.9493	0.9400	0.9486	0.8841
3keyword	0.9374	0.8412	0.9367	0.8887	**0.9770**	0.8611	0.9667	0.8774	0.8092	0.7871	0.8081	0.8872
4keyword	0.8821	0.7395	0.8588	0.8588	**0.9417**	0.6818	0.9255	0.8929	0.7761	0.6586	0.7703	0.8848
5keyword	0.7606	0.6287	0.7778	0.7778	**0.9105**	0.6208	0.8683	0.8800	0.6637	0.4187	0.6734	0.8253
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.0373	0.0967	0.0967	0.1245	**0.0136**	0.0943	0.0173	0.1402	0.1018	0.1086	0.1032	0.1942
3keyword	0.1531	0.3230	0.1559	0.2328	**0.0530**	0.2812	0.0752	0.2283	0.3645	0.4129	0.3663	0.2132
4keyword	0.2854	0.5458	0.3283	0.3283	**0.1276**	0.6388	0.1687	0.2215	0.4628	0.6871	0.4751	0.2316
5keyword	0.4794	0.7980	0.5098	0.5098	**0.2102**	0.8050	0.2982	0.2558	0.7189	1.1633	0.6974	0.3793
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.0360	0.0429	0.0429	0.0444	**0.0267**	0.0332	0.0281	0.0352	0.0723	0.0391	0.0737	0.0451
3keyword	0.0522	0.0906	0.0514	0.0606	**0.0212**	0.0379	0.0231	0.0287	0.1296	0.0575	0.1316	0.0495
4keyword	0.0609	0.1215	0.0737	0.0737	**0.0236**	0.1017	0.0490	0.0303	0.1479	0.0749	0.1530	0.0271
5keyword	5.2845	0.8762	0.2876	0,0458	0.0281	0.0855	0.1334	**0.0214**	0.2848	0.9381	0.3779	0.0311

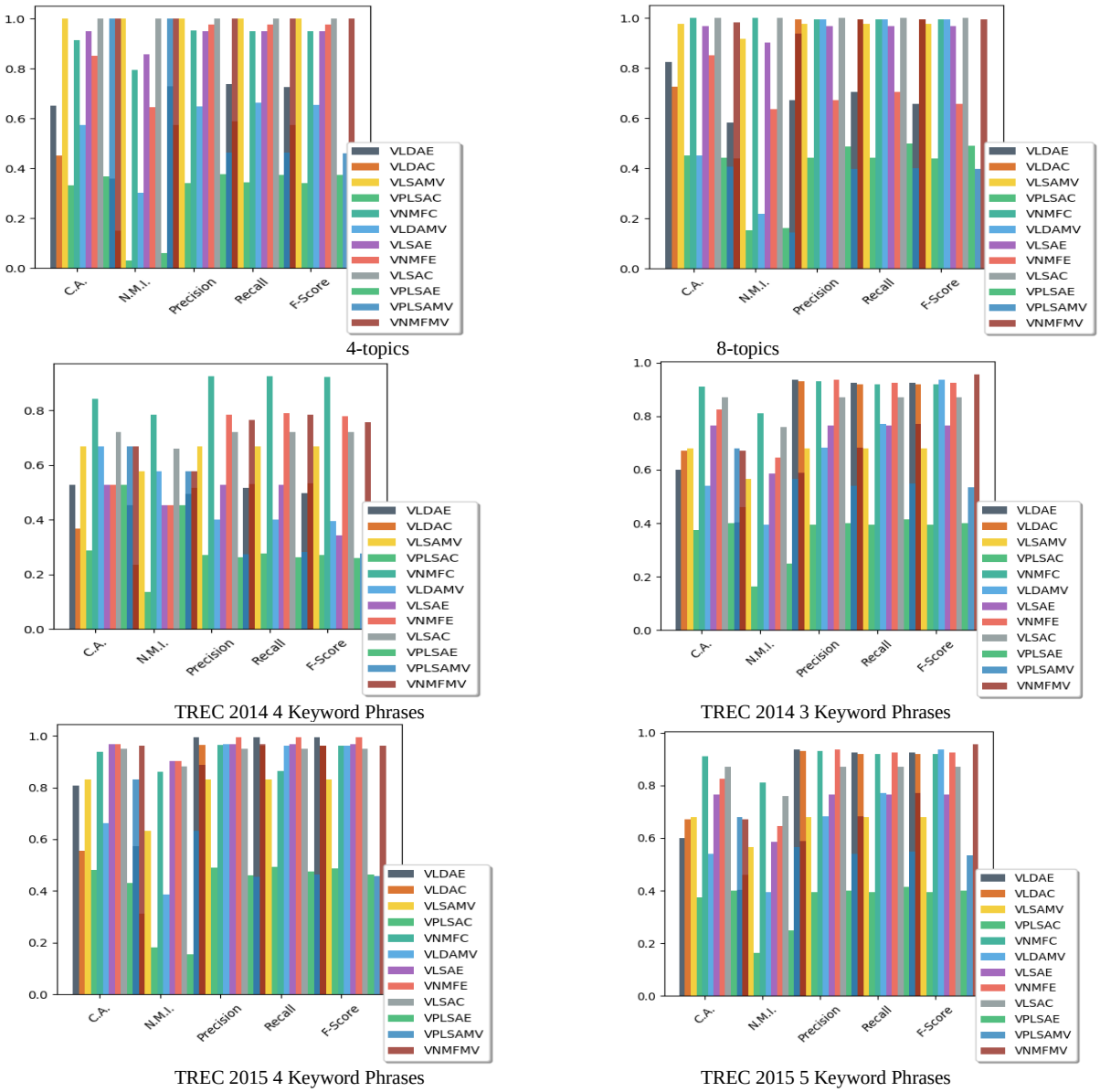


Fig. 4. Bar Graphs of External Validity Indexes of 4-Topics, 8-Topics, 12-Topics, 20-Topics, TREC2014-4, 3Keyword Phrases, and TREC2015-4 and 5 Keyword Phrases Datasets.

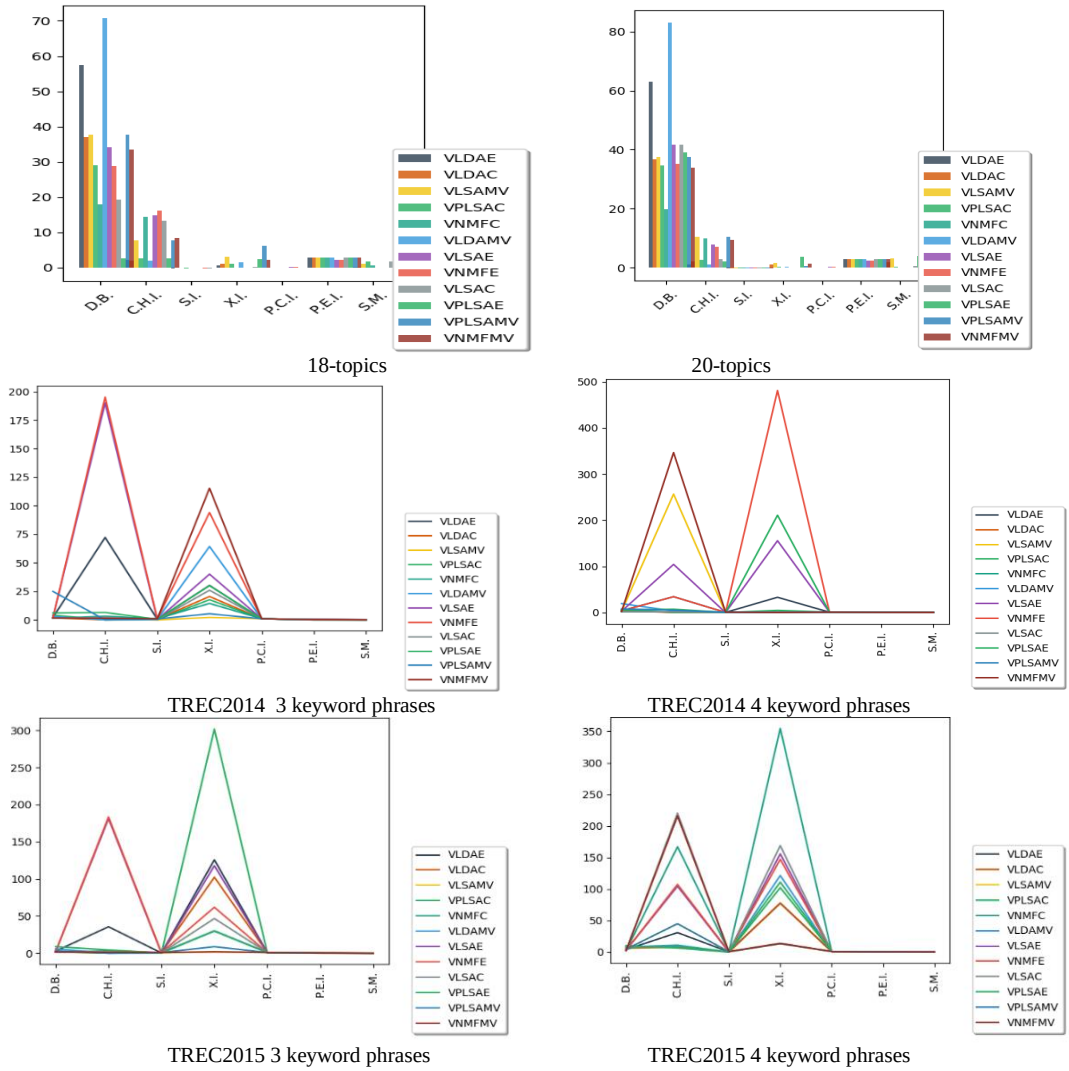


Fig. 5. Line Graphs of Internal Indexes of TREC2014 4, 3 Keyword Phrases and TREC2015 4, 3Keyword Phrases.

D. Convergence Study

Table IV shows the total number of iterations required for the execution of hybrid visual topic models with an error tolerance of 0.000001 for convergence of 2-topics to 20-topics of twitter health datasets and TREC2014 and TREC2015 2, 3, 4 and 5-keyword phrases of data sets under multi-viewpoints, cosine and Euclidean metrics.

Bar graphs are represented in Fig. 6 for 2-topics to 20-topics health datasets and TREC2014 and TREC2015 keyword phrases. On overall observations in all types of datasets, the total number of iterations required within error tolerance is less under multi-viewpoints cosine distance metric than Euclidean and cosine based metric. There are only a few cases where less number of iterations required under Euclidean distance such as 13-topics, 15-topics, 16-topics, and 20-topics.

E. Computational Complexity Analysis

Computational time in seconds and allocated memory in Kb are measured for evaluating the computational complexity of visual topic models. Table V presents the total computation time of visual topic models for various subsets, include 5-topics, 10-topics, 15-topics, and 20-topics of datasets, TREC2014, and TRC2015 all keyword phrases datasets. It presents the computation time using

Euclidean, cosine based and multi-viewpoint cosine based distance metrics for comparative analysis purpose. It is noted that VNMF under multi-viewpoint is taking less amount of time when the number of topics is less than five. As the number of topics increasing VLSA under cosine based metric taking less time than other models. Hence, VLSA under cosine based metric is the time-efficient model for topics clustering.

Fig. 7 shows the computational time analysis using bar graphs under Euclidean, cosine based and multi-viewpoints cosine based metrics. In the cosine metric, it is observed that VPLSA has taken less computational time than other models for 5-topics to 20-topics health data and TREC2014 and TREC2015 datasets.

TABLE. IV. TOTAL NUMBER OF ITERATIONS FOR CONVERGENCE

Under Multi-viewpoints

2topics	3topics	4topics	5topics	6topics	7topics	8topics	9topics	10topics	11topics	12topics	13topics	14topics	15topics	16topics	17topics	18topics	19topics	20topics
30	39	**37**	**59**	57	**63**	83	72	**71**	81	**90**	75	**86**	100	123	125	**110**	**111**	134

Under Cosine

2topics	3topics	4topics	5topics	6topics	7topics	8topics	9topics	10topics	11topics	12topics	13topics	14topics	15topics	16topics	17topics	18topics	19topics	20topics
29	**30**	41	68	**50**	69	**75**	**63**	92	**73**	106	80	111	90	107	113	118	112	116

Under Euclidean

2topics	3topics	4topics	5topics	6topics	7topics	8topics	9topics	10topics	11topics	12topics	13topics	14topics	15topics	16topics	17topics	18topics	19topics	20topics
32	44	46	68	63	69	86	75	82	89	92	**74**	94	**90**	**101**	**96**	154	116	**114**

TREC2014 Keyword Phrases									TREC2015 Keyword Phrases									
Multi-viewpoint			Cosine			Euclidean			Multi-viewpoint				Cosine				Euclidean			
2key	**3key**	**4key**	**2key**	**3key**	**4key**	**2key**	**3key**	**4key**	**2key**	**3key**	**4key**	**5key**	**2key**	**3key**	**4key**	**5key**	**2key**	**3key**	**4key**	**5key**
26	**50**	**56**	59	56	80	76	57	70	**30**	**48**	**50**	**56**	35	75	54	70	81	62	80	69

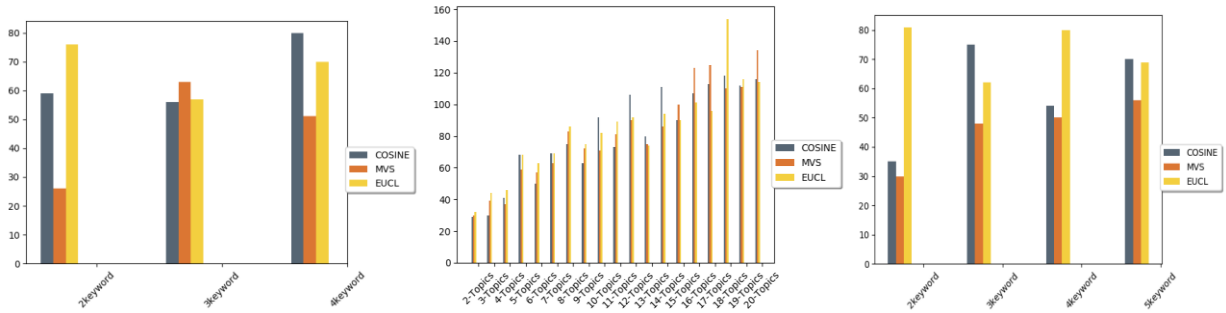


Fig. 6. Total Number of Iterations for Convergence of 2-Topics to 20-Topics, TREC2014, and TREC2015 Keyword Phrases.

TABLE. V. TOTAL TIME (SEC) TAKEN OF HEALTH TWEETS DATASETS FROM TWITTER

Tweets Dataset	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.080	0.118	0.054	0.860	0.336	0.313	**0.045**	1.042	0.394	0.289	0.153	1.094
10-Topics	0.331	0.767	0.147	5.860	0.703	0.475	**0.061**	7.722	0.716	0.589	0.137	6.786
15-Topics	1.716	1.621	0.265	24.031	2.689	1.056	**0.206**	22.158	2.683	0.986	**0.177**	22.291
20-Topics	1.704	0.856	0.130	37.928	5.492	0.845	**0.125**	39.860	4.358	1.154	0.124	38.964
TREC2014	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**0.032**	0.087	0.048	0.041	0.052	0.084	0.051	0.076	0.148	0.261	0.103	0.102
3Keyword	**0.117**	0.304	0.122	0.198	0.156	0.135	0.137	0.192	0.159	0.261	0.136	0.146
4Keyword	**0.129**	0.252	0.130	0.453	0.214	0.257	0.194	0.842	0.215	0.159	0.161	0.726
Trec2015	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**0.305**	0.365	0.357	0.356	0.376	0.498	0.388	0.367	0.326	0.312	0.355	0.321
3keyword	**0.052**	0.174	0.060	0.271	0.134	0.286	0.084	0.279	0.115	0.160	0.094	0.277
4keyword	**0.071**	0.445	0.059	0.503	0.193	0.132	0.073	0.430	0.144	0.184	0.061	0.648
5keyword	0.267	0.796	0.177	0.932	0.269	0.296	**0.110**	1.204	0.423	0.564	0.140	0.453

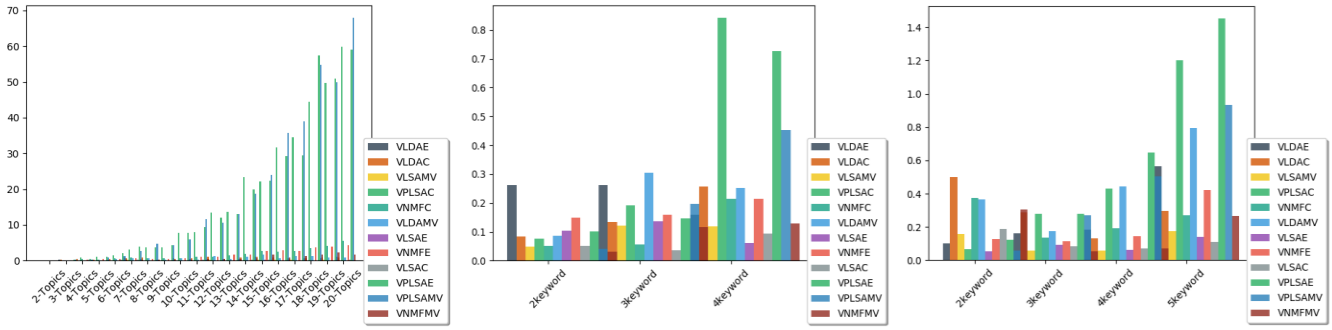


Fig. 7. Total Time is Taken (sec) for the Execution of 2-Topics to 20-Topics of Twitter Datasets and TREC2014 Keyword Phrases under Multi-Viewpoint Cosine based, Cosine based and Euclidean Distances.

A memory requirement is another important criterion for calculating computational complexity. In the experimental analysis, the memory allocated for successful running visual topic models are measured in terms of Kb for 2-topics to 20-topics, TREC2014 and TREC2015 datasets under three different similarity measures, i.e., Euclidean, cosine based, and multi-viewpoint cosine distance values are tabulated in Table VI. It is observed that VLSA under cosine based is more memory efficient than other visual hybrid models.

Fig. 8 shows the memory requirement comparison for visual topic models, and it is evaluated for 2-topics to 20-topics, TREC2014, and TREC2015 using Euclidean, cosine based, and multi-viewpoints cosine based distance that of other visual topic models, though VLSA and VNMF visual model performs well in other aspects i.e., results visually, which are more suitable for our further work in developing smart healthcare applications. Our proposed visual topic models outperform with other traditional topic models in two aspects; VLSI under cosine has taken less amount of space than other visual topic models in most of the health topics. VLSI performs well concerning space complexity than other models.

TABLE. VI. ALLOCATED MEMORY (KB) OF HEALTH TWEETS DATASETS FROM TWITTER

Tweets Dataset	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	14988	72	**0**	648948	37776	36736	6988	750960	29440	14312	1024	752208
10-Topics	43976	8084	2460	5268368	24944	52	**0**	6808840	62504	34640	2040	6090784
15-Topics	52248	17980	**4420**	23677472	88388	22912	4424	21312380	87696	22244	5284	21319836
20-Topics	26376	**156**	3812	68096596	80436	284	3816	58952964	79044	164	3812	57937436
TREC2014	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	692	**0**	**0**	**0**	**0**	**0**	68	**0**	13312	5656	1024	3336
3Keyword	14468	43880	6892	146116	**0**	**0**	**0**	123588	14708	26916	11516	85612
4Keyword	24104	24544	1024	266248	6036	12	1140	397804	8276	684	**0**	348036
TREC2015	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	13196	4388	2052	2528	12652	17308	1024	2132	**0**	**0**	68	64
3keyword	2180	68	**0**	133840	100	1524	1056	143400	2920	300	4	102472
4keyword	1140	1504	**0**	227408	3392	**0**	**0**	208868	3720	**0**	**0**	397396
5keyword	12788	64648	2488	501600	5304	**0**	**0**	620012	5340	**0**	**0**	611308

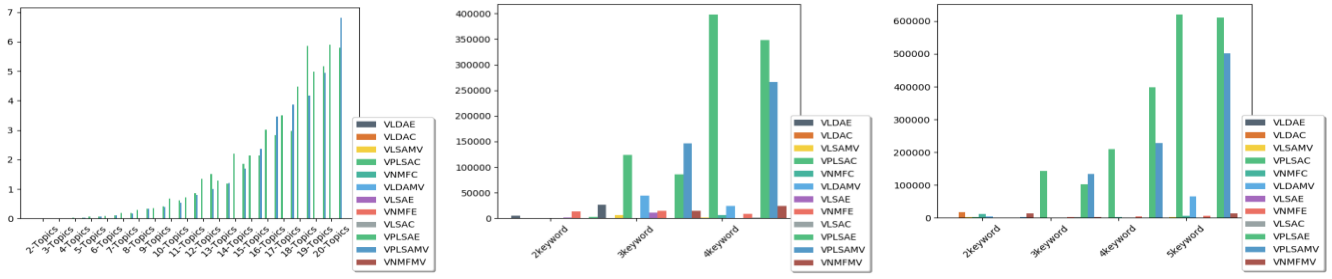


Fig. 8. Total Time is Taken (sec) for the Execution of 2-Topics to 20-Topics of Twitter Datasets and TREC2014 Keyword Phrases under Multi-Viewpoint Cosine based, Cosine based and Euclidean Distances.

First, cluster tendency or prior knowledge about social data is unknown in existing topic models, whereas our models assess the prior information of social data clusters visually without external interference. Secondly, our visual topic models effectively deal with a large amount of unlabeled social data in determining the number of clusters (or topics) visually. In our work, cluster validity is assessed by both internal and external cluster index measures with Euclidean, cosine based, and multi-viewpoints cosine based similarity measures.

VI. Conclusion and Future Work

Topic models are capable of finding hidden conceptual topics from such a vast number of terms of tweet documents. The topic models, LSI, PLSI, NMF, LDA, and intJNMF determines the topic clusters without knowledge of cluster tendency. Hybrid topic models overcome the problem of health cluster tendency and improve the performance of topic clustering. The empirical analysis of proposed hybrid topic model techniques is performed based on the parameters, such as convergence speed, time, and space computational complexities. In our experimental study, cosine based hybrid topics models much succeeded for the detection of hidden concepts or topics from tweets documents and observed numerical improvement with an increased rate of 40% to 45% over 2-topics to 10-topics and 35% to 30% over 11-topics to 20-topics. Among proposed visual topic models, VLSA and VNMF under multi-viewpoints cosine show the best performance in finding both numbers of topics for unlabeled twitter data and tweets clustering internal and external validity indexes, total numbers of iterations for convergence and less time is taken for execution. Hence, VNMF and VLSA show better performance in finding both the number of topics as well as clustering results for unlabeled twitter data and tweets. Based on the observation of space complexity in experimental, it needs to be improved as scalable visual topic models in future work for performing effective big data health clustering.

Acknowledgment

This work is supported by the Science & Engineering Research Board (SERB), Department of Science and Technology, Government of India for the Research Grant of DST project number ECR/2016/001556.

References

[1] J. Vosecky, D. Jiang, K. W.-T. Leung, K. Xing, and W. Ng, "Integrating social and auxiliary semantics for multifaceted topic modelling in twitter, " ACM Transactions on Internet Technology (TOIT), vol. 14, no. 4, pp. 27, 2014.

[2] Chen Z, and Liu B, "Mining topics in documents: standing on the shoulders of big data, " in Proc. 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1116-1125, Aug. 2014.

[3] Amrutha Benny and Mintu Phili, " Keyword Based Tweet Extraction and Detection of Related Topics, " ICICT2014, vol. 46, pp. 364-371, 2015.

[4] McCallum, A. Corrada-Emmanuel, and X. Wang, “Topic and role discovery in social networks,” Computer Science Department Faculty Publication Series, pp. 3, 2005.

[5] Vandana Singh, and Sanjay Kumar Dubey, " Opinion mining and analysis: A literature review, " 2014 IEEE conference, pp. 25-26, Sep. 2014.

[6] Menna Allah Hassan, " A comparative study of classification algorithm in e-health environment, " 2016 IEEE conference (ICDIPC), April 2016.

[7] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, “Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering,” IEEE Transactions on Neural Networks, vol. 22, no. 12, pp. 2117–2131, 2011.

[8] Kolini, Farzan, Janczewski, and Lech, "Clustering and Topic Modelling: A New Approach for Analysis of National Cybersecurity Strategies," PACIS 2017 Proceedings no. 126, pp.1-12, 2017.

[9] James C Bezdek " VAT: A tool for visual assessment of cluster tendency, " IJCNN2002, Feb 2002, DOI: 10.1109/IJCNN.2002 .1007487.

[10] Wuhan, "TF-IDF based Feature Words Extraction and Topic Modeling for Short Text," ICMSS2018, Jan. 2108, DOI: 10.1145/3180374.3181354.

[11] Alexandra Schofield " Pulling out the stop: Rethinking stop words removal for topic models, " 15th conference of ACM, pp:432-36, April 2107.

[12] Amritanshu Agrawal " What is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE), " IEEE transaction on software engineering, Aug 2016, DOI: 10.1016/j.infsof.2018.02.005.

[13] Dredze M "How Social Media Will Change Public Health," IEEE Intelligent Systems, Vol.27, no.4, pp.81–84, Aug. 2012.

[14] Dr Choudhury M, Gamon M, Counts S, and Horvitz E, "Predicting Depression via Social Media," in Proc. 7th International Conference on Weblogs and Social Media, 2013.

[15] Kumar, D., Bezdek, J. C., Palaniswami, M., Rajasegarar, S., Leckie, C., & Havens, T. C. (2016). A Hybrid Approach to Clustering in Big Data. IEEE Transactions on Cybernetics, 46(10), 2372–2385. DOI:10.1109/tcyb.2015.2477416.

[16] K. Rajendra Prasad and M. Suleman Basha, "Improving the performance of speech clustering method," 10th International Conference on Intelligent Systems and Control (ISCO), 2016, DOI: 10.1109/ISCO.2016.7726878.

[17] Eswara Reddy B, and K. Rajendra Prasad, "Improving the performance of visualized clustering method," International Journal of Systems Assurance Engineering and Management, Vol. 7, pp.102-111, Dec.2016.

[18] Y. Hu, A. John, F. Wang, and S. Kambhampati, "Et-LDA: Joint topic modelling for aligning events and their twitter feedback, " in AAAI Conference on Artificial Intelligence (AAAI 2012), vol. 12, Toronto, Ontario, Canada, pp. 59–65, July 2012.

[19] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modelling based on interactive nonnegative matrix factorization," IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 1992–2001, 2013.

[20] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, “Learning topics in short texts by non-negative matrix factorization on term correlation matrix,” in Proceedings of the SIAM International Conference on Data Mining (SIAM 2013, San Diego, California, USA: SDM, July 2013.

[21] M. Albakour, C. Macdonald, and I. Ounis, "On scarcity and drift for effective real-time filtering in microblogs," in Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM2013), pp.419–428, October 2013.

[22] J. Li, Z. Tai, R. Zhang, W. Yu, and L. Liu, “Online busty event detection from microblog,” in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference, pp. 865–870, Dec. 2014.

[23] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models, ” The International AAAI Conference on Web and Social Media (ICWSM), vol. 10, pp. 130–137, May 2010.

[24] Yan, X, and Guo, J, " Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization, " in Proc. CIKM 2012, Miami, HI, USA, pp. 2259–2262, 2012.

[25] Yan, X, and Guo, J, "Learning Topics in short text Using Ncut-weighted non-negative matrix Factorization on term correlation matrix," http://xiaohuiyan.com/papers/TNMF-SDM-13.pdf

[26] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation, " The Journal of Machine Learning Research, Vol.3, pp.993–1022, 2003.

[27] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, " Indexing by latent semantic analysis, " Journal of American Society for Information Sciences, Vol.41, no.6, pp:391-407, 1990.

[28] Singular Value Decomposition [Online] Available: web.mit .edu/be.400/www/SVD/Singular_Value_Decomposition.htm.

[29] Thomas Hofmann, "Probabilistic latent semantic indexing," SIGIR, ACM, pp:50-57, 1999.

[30] Dempster A, Laird N, and Rubin D, "Maximum likelihood from incomplete data via the EM algorithm," J. royal statistics society B 39, pp: 1-38, 1977.

[31] Madhuri Avula, Narasimha Prasad Lakkakula, and Murali Prasad Raja" Bone cancer detection from mri scan imagery using mean pixel intensity" IEEE 8th Asia Modelling Symposium, pp:141-146, 2014.

[32] PR Anisha, C Kishor Kumar Reddy, and LV Narasimha Prasad" A pragmatic approach for detecting liver cancer using image processing and data mining techniques" IEEE International Conference on Signal Processing and Communication Engineering Systems, pp: 352-357, 2015.

[33] TREC2015 https://trec.nist.gov/pubs/trec24/trec2015.html

[34] TREC2014 https://trec.nist.gov/pubs/trec23/trec2014.html.

[35] C. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval ", online edition, vol. 1, Cambridge, 2008, https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[36] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems 13, NIPS 2000, Denver, CO, USA, pp. 556–562, 2000.

[37] R. Nugroho, J. Yang, Y. Zhong, C. Paris, and S. Nepal, "Deriving topics in twitter by exploiting tweet interactions," Proceedings of the 4th IEEE International Congress on Big Data, New York, USA: IEEE Services Comp. Community, July 2015.

[38] R. Pochampally and V. Varma, "User context as a source of topic retrieval in twitter," Workshop on Enriching Information Retrieval (with ACM SIGIR). Beijing, China: ACM, pp. 1–3, July 2011.

[39] Paul M, and Girju R, "A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics," Proc.24th AAAI-10 Conference on Artificial Intelligence, Atlanta, USA, 2010.

[40] Robertus Nugroho, Jian Yang, Weiliang Zhao, Cecile Paris, and Surya Nepal, "What and With Whom? Identifying Topics in Twitter Through Both Interactions and Text ", Journal of Latex Class Files, Vol.14, No.8, August 2015.

[41] Pattanodom, M., I am-On, N., and Boongoen, T. " Clustering data with the presence of missing values by ensemble approach," 2016 Second Asian Conference on Defense Technology (ACDT). doi:10.1109/acdt.2016.7437660.

[42] Alessia Amelio and Clara Pizzuti, " Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods? , " IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[43] Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., & Yao, H. (2019). Research on Topic Detection and Tracking for Online News Texts. IEEE Access, 7, 58407–58418. DOI:10.1109/access.2019.2914097.

[44] Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on-the topic model. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).doi:10.1109/icis.2016.7550929.

[45] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data USING clustering," Machine Learning, vol. 42, nos. ½, pp. 143-175, Jan. 2001.

[46] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering," Proc. IEEEInt'lConf.DataMining(ICDM),pp.107-114,2001.

[47] H.Zha, X.He, C.Ding, H.Simon, and.Gu, "SpectralRelaxation for K-Means Clustering," Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064,2001.

[48] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug.2000.

[49] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274,2001.

[50] Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag,2007.

[51] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf.Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pp. 58-64, July 2000.

[52] D. Ienco, R.G. Pensa, and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Proc. Eighth Int'l Symp. IntelligentDataAnalysis(IDA),pp.83-94,2009.

[53] P. Lakkaraju, S. Gauch, and M. Speretta, "Document Similarity Based on Concept Tree Distance," Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132,2008.

[54] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept.2008.

[55] Ujjwal Maulik and Sanghamitra Bandyopadhyay, " Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, December 2002.

[56] Tapana Mekaroonkamon and Sarwan Wongsa, "A Comparative Investigation of the Robustness of Unsupervised Clustering Techniques for Rotating Machine Fault Diagnosis with Poorly-Separated Data," 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016.

[57] Dan A. Simovici, Member, IEEE, and Szymon Jaroszewicz, "An Axiomatization of Partition Entropy," IEEE Transactions on Information Theory, vol. 48, no. 7, July 2002.

[58] Kelemen, O., Tezel, O., Ozkul, E., Tiryaki, B. K., and Agayev, E, "A comparison of validity indices on fuzzy C-means clustering algorithm for directional data," 2017 25th Signal Processing and Communications Applications Conference (SIU).doi:10.1109/siu.2017.7960557.

[59] Kuo-Lung Wu, "An analysis of robustness of partition coefficient index. 2008 IEEE International Conference on Fuzzy Systems," IEEE World Congress on Computational Intelligence.doi:10.1109/fuzzy .2008.4630393.

[60] Vergani, A. A., & Binaghi, E, "A Soft Davies-Bouldin Separation Measure," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). DOI:10.1109/fuzz-IEEE.2018.8491581.

[61] K. Rajendra Prasad, Moulana Mohammed, R M Noorullah,, " Visual Topic Models for Healthcare Data Clustering", Evolutionary Intelligence, Oct, 2019, pp:1-17, https://doi.org/10.1007/s12065-019-00300-y.

Lizard Cipher for IoT Security on Constrained Devices

Ari Kusyanti[1], Rakhmadhany Primananda[2], Kalbuadi Joyo Saputro[3]
Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

***Abstract*—Over the past decades, security become the most challenging task in Internet of Things. Therefore, a convenient hardware cryptographic module is required to provide accelerated cryptographic operations such as encryption. This study investigates the implementation of Lizard cipher on three Arduino platforms to determine its performance. This study is successful in implementing Lizard cipher on Arduino platform as a constrained devices and resulting 0.98 MB of memory utilization. The execution time of Lizard cipher is compared among Arduino variants, i.e Arduino Mega, Arduino Nano and Arduino Uno with ANOVA test. Tukey's HSD post-hoc test reveales that the execution time is significantly slower in Arduino Mega compared to Arduino Nano and Arduino Uno. This result will help IoT security engineers in selecting a lightweight cipher that is suitable for constraints of the target device.**

***Keywords*—Lizard cipher; IoT security; Arduino; ANOVA**

I. Introduction

In recent years, Internet of Things (IoT) gets its popularity due to its integrated architecture that connects 'things' to share information between them. Microcontroller is one of the 'things' that is widely adopted for IoT considering its various functionalities. Since the increasing numbers of devices that connect to IoT, the possibility of security violation is also increasing [1] [2]. One of the solutions is by encrypting the data that can reduce the security risk. There is solution for the problem based on hardware which is called Hardware Security Modules (HSM). However HSM are expensive to be implemented in IoT.

For this purpose, three widely used target microcontrollers in the Internet of Things context, i.e. Arduino Uno, Nano and Mega are selected. Arduino is an easy to use and low cost device and equipped with open source platform. Arduino utilizes a hardware known as the Arduino development board and software for programming the code known as the Arduino IDE (Integrated Development Environment).

Apart from the targeted devices, the cipher itself has to be taken into consideration when implemented in Internet of Things environment. Lizard [3] is created with low-cost scenarios in mind. The design of Lizard is inspired from Grain v1 [4], but with the smaller inner state which is 121 bits and the larger key of 120 bits. It outperforms Grain v1 in memory area usage and power consumption; hence Lizard is suitable for power-constrained devices.

The objective of this study is to investigate the implementation of Lizard cipher on three Arduino platforms to determine its performance i, i.e. Arduino Mega, Arduino Uno and Arduino Nano. To evaluate the performance, the memory utilization and execution time of Lizard is observed. To analyze the data, ANOVA test is used to examine the difference implementation of Lizard on three Arduino.

II. Related Study

Based on previous studies, a number of cryptographic algorithms had been implemented on microcontrollers, such as [5] which evaluates the performance of public key cryptography such as RSA and ECC with various libraries on Arduino. The work of [6] had evaluated 19 lightweight ciphers on three microcontroller platforms: 8-bit AVR, 16-bit MSP430, and 32-bit ARM. The ciphers are all block cipher namely AES [7], Chaskey [8], Fantomas [9], HIGHT [10], LBlock [11], LEA [12], LED [13], Piccolo [14], PRESENT [15], PRIDE [16], PRINCE [17], RC5 [18], RECTANGLE [19], RoadRunneR [20], Robin [21], Simon [22], SPARX [23], Speck [24], and TWINE [25]. According to [26] stream ciphers tend to perform faster and more efficient than block ciphers. Therefore in this study, Lizard cipher as one of stream cipher is evaluated so that it can help IoT security engineers when selecting a lightweight ciphers that suits the requirements of the constraints of the target device.

III. Lizard Architecture and Arduino

This section will disscuss the overview of Lizard cipher [3] and the target devices i.e. Arduino Mega, Arduino Nano and Arduino Uno, also the test to analyze the data observed namely ANOVA test.

A. Lizard Cipher

In this section, the details of Lizard cipher are discussed. First, the components of the cipher in detail is described. Then, initialization state is specified and finally keystream generation process is described.

Components of Lizard cipher:

Lizard cipher has three main blocks, i.e NFSR (Non Linear Feedback Shift Register) i.e. NFSR1 and NFSR2, and output function a as depicted in Fig. 1. Lizard requires inputs of 120 bits key and 64 bits IV (Initiation Vector) and produces one bit keystream at a time.

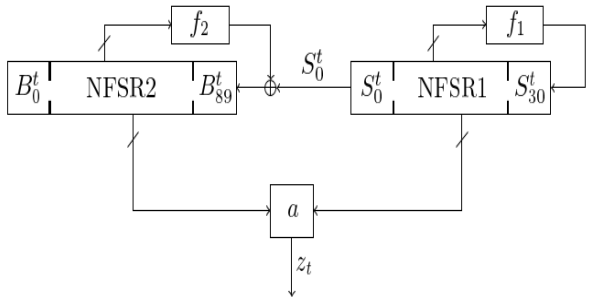


Fig. 1. Lizard Architecture.

NFSR (Non Linear Feedback Shift Register)

The first component on Lizard is NFSR which consist of two blocks, i.e NFSR1 and NFSR2. The difference between these two is in the feedback function. NFSR1's content is denoted as $s_i, s_{i+1}, , , s_{i+79}$, while NFSR2's is $b_i, b_{i+1}, , , b_{i+63}$. As for NFSR1, the update relation of NFSR1 is defined as

$$\begin{aligned} S_{30}^{t+1} = {} & S_0^t \oplus S_2^t \oplus S_5^t \oplus S_6^t \oplus S_{15}^t \oplus S_{17}^t \oplus S_{18}^t \oplus S_{20}^t \\ & \oplus S_{25}^t \oplus S_8^t S_{18}^t \oplus S_8^t S_{20}^t \oplus S_{12}^t S_{21}^t \\ & \oplus S_{14}^t S_{19}^t \oplus S_{17}^t S_{21}^t \oplus S_{20}^t S_{22}^t \\ & \oplus S_4^t S_{12}^t S_{22}^t \oplus S_4^t S_{19}^t S_{22}^t \oplus S_7^t S_{20}^t S_{21}^t \\ & \oplus S_8^t S_{18}^t S_{22}^t \oplus S_8^t S_{20}^t S_{22}^t \oplus S_{12}^t S_{19}^t S_{22}^t \\ & \oplus S_{20}^t S_{21}^t S_{22}^t \oplus S_4^t S_7^t S_{19}^t S_{21}^t \\ & \oplus S_4^t S_{12}^t S_{21}^t S_{22}^t \oplus S_4^t S_{19}^t S_{21}^t S_{22}^t \\ & \oplus S_7^t S_8^t S_{18}^t S_{21}^t \oplus S_7^t S_8^t S_{20}^t S_{21}^t \\ & \oplus S_7^t S_{12}^t S_{19}^t S_{21}^t \oplus S_8^t S_{18}^t S_{21}^t S_{22}^t \\ & \oplus S_8^t S_{20}^t S_{21}^t S_{22}^t \oplus S_{12}^t S_{19}^t S_{21}^t S_{22}^t \end{aligned}$$

Whilst update relation of NFSR2 is defined as

$$\begin{aligned} B_{89}^{t+1} = {} & S_0^t \oplus B_0^t \oplus B_{24}^t \oplus B_{49}^t \oplus B_{79}^t \oplus B_{84}^t \oplus B_3^t B_{59}^t \oplus B_{10}^t B_{12}^t \\ & \oplus B_{15}^t B_{16}^t \oplus B_{25}^t B_{53}^t \oplus B_{35}^t B_{42}^t \oplus B_{55}^t B_{58}^t \\ & \oplus B_{60}^t B_{74}^t \oplus B_{20}^t B_{22}^t B_{23}^t \oplus B_{62}^t B_{68}^t B_{72}^t \\ & \oplus B_{77}^t B_{80}^t B_{81}^t B_{83}^t \end{aligned}$$

Output Function

The final block of output function is defined as

$$z_t \coloneqq \mathcal{L}_t \oplus Q_t \oplus T_t \oplus T'_t$$

where

$$\mathcal{L}_t \coloneqq B_7^t \oplus B_{11}^t \oplus B_{30}^t \oplus B_{40}^t \oplus B_{45}^t \oplus B_{54}^t \oplus B_{71}^t$$

$$Q_t \coloneqq B_4^t B_{21}^t \oplus B_9^t B_{52}^t \oplus B_{18}^t B_{37}^t \oplus B_{44}^t B_{76}^t$$

$$\begin{aligned} T_t \coloneqq {} & B_5^t \oplus B_8^t B_{82}^t \oplus B_{34}^t B_{67}^t B_{73}^t \oplus B_2^t B_{28}^t B_{41}^t B_{65}^t \\ & \oplus B_{13}^t B_{29}^t B_{50}^t B_{64}^t B_{75}^t \oplus B_6^t B_{14}^t B_{26}^t B_{32}^t B_{47}^t B_{61}^t \\ & \oplus B_1^t B_{19}^t B_{27}^t B_{43}^t B_{57}^t B_{66}^t B_{78}^t \end{aligned}$$

$$T'_t \coloneqq S_{23}^t \oplus S_3^t S_{16}^t \oplus S_9^t S_{13}^t B_{48}^t \oplus S_1^t S_{24}^t B_{38}^t B_{63}^t$$

Initiaization state

The state initialization process can be divided into 4 phases, namely

a) Phase 1: Key and IV Loading
b) Phase 2: Grain-like Mixing
c) Phase 3: Second Key Addition
d) Phase 4: Final Diffusion

Prior to generating any keystream, there are four phases that has to be completed. Firstly, the cipher has to be initialized with the key and the IV. Let the bits of the key, k, denoted as k_i, $0 \le i \le 119$ and the bits of the IV denoted as IV_i, $0 \le i \le 63$. The initialization process is performed as follows.

$$B_j^0 = \begin{cases} K_j \oplus IV_j, & for\ j \in \{0, \dots, 63\}, \\ K_j, & for\ j \in \{64, \dots, 89\}, \end{cases}$$

$$S_i^0 \coloneqq \begin{cases} K_{i+90}, & for\ i \in \{0, \dots, 28\}, \\ K_j \oplus 1, & for\ i = 29, \\ 1, & for\ i = 30. \end{cases}$$

Secondly, the Grain-mixing like process, the cipher is clocked 128 times without producing any running key. In this process, the output function is fed back and xored with the input, both to the NFSR1 and to the NFSR2 as shown in Fig. 2. Similiarly for the phase 3, second key addition process is performed. After this initialization process is complete, keystream is generated in phase 4.

Key Generation state:

After completing all four stages, the first keystream bit that is used for plaintext encryption is z_{257}. The length of the keystream depends on the length of plaintext, since it will be encrypted bit by bit to produce ciphertext.

B. Arduino

Arduino is a microcontroller which can be easily used, low-cost and has capability of mini-computer [27]. Arduino builds upon a hardware known as the Arduino development board and software for developing the code known as the Arduino IDE (Integrated Development Environment). There are a variety of Arduino models with various features and enhancement in their latest versions that add more components. In this study, three types of Arduino are used, i.e. Arduino Mega, Arduino Uno, Arduino Nano.

The main difference among these Arduino is the concern of this study which lies on the processor. Arduino Mega 2560 features an ATmega2560 at its heart, while Arduino Nano is equipped with ATmega168 and Arduino Uno is based on ATmega328 MCU.

Table I presents the main characteristics of the target devices used, while in the next paragraphs provide a brief description of each Arduino type.

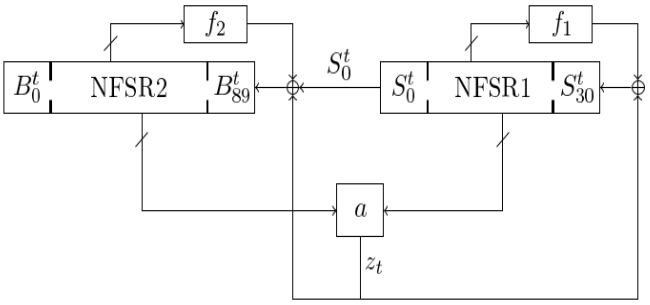


Fig. 2. Lizard State Iniziation.

TABLE. I. ARDUINO TYPE

Parameter	Nano	Uno	Mega
Processor	ATmega168	ATmega 328P	ATmega2560
Clock Speed	16 MHz	16 MHz	16 MHz
Flash Memory	16 kB	32 kB	256 kB
EEPROM	512 bytes	1 kB	4 kB
SRAM	1 kB	2 kB	8 kB
Voltage Level	5 V	5 V	5 V

The first target, Arduino Nano with ATmega168 as the processor has a maximum frequency of 16 MHz and with 1kB SRAM and 16 KB flash memory. The second target is Arduino Uno with ATMega328P as its heart. The processor has 2kB SRAM and 32kB flash memory. Lastly, Arduino Mega with ATmega 2560 processor is equipped with 8 KB SRAM and 256 KB flash memory. All of these three operate at V = 5 V.

C. A One-Way Analysis of Variance (ANOVA)

In the process of examining the relationship between variables, ANOVA is used to compare the means of two or more groups on the dependent variable. ANOVA a hypothesis-testing technique compares the variance between samples to variation within each sample. When the value of between variation is bigger than the within variation, hence the means of different samples is not equal. On the other hand, when the between and within variations are the same, hence there is no significant difference between sample means.

Before analyzing the data with ANOVA, the population has to follow a normal distribution, hence it is categorize as a parametric tests. On the contrary, an ANOVA test cannot be used to test the equality of the sample means. In this case, a non-parametric test is used since it does not rely on distributional assumptions. There are several assumptions that have to be met:

a) Assumption of independence: The samples have to be selected randomly independent to one another.

b) Assumption of normality: The population has to follow a normal distribution.

c) Assumption of homogeneity of variance: The variance of population has to be equal.

In ANOVA there are two hypotheses which are null hypothesis (H0) and alternative hypothesis (Ha). The null hypothesis for an ANOVA assumes the population means are equal, while alternative hypothesis assumes that at least one mean is not statistically equal.

In summary, an ANOVA test is used to determine at least one mean is different. However, an additional test must be conducted to determine which mean(s) is/are different which is called post-hoc test. There are several techniques for testing the differences between means, Tukey's HSD is used as the post-hoc test that helps to describe further data analysis technique for this study.

IV. IMPLEMENTATION

Arduino is an open source microcontroller board [27] that can be used to program by using free development software. The Arduino uses a simplified version of C/C++ programming language. Integrated Development Environment (IDE) is used to program and configure the Arduino which can be downloaded from the official website of Arduino freely.

The implementation of Lizard cipher can be an invaluable part of an overall security solution in IoT which provides accelerated cryptographic operations. Lizard cipher is successfully implemented in various Arduino by using Arduino IDE. It is able to generate keystream and is tested against test vectors. The experiment run 800 times for each Arduino with various combinations of Key and IV as inputs which generates keystream.

Descriptive statistics is conducted to observe the distribution of the data. Table II displays the summary of the descriptive statistics.

The average of execution time is 55813.4850 microsecond for Arduino Mega, 55596.5700 microsecond for Arduino Uno, and 55596.5700 microsecond for Arduino Nano as can be seen in Fig. 3.

Meanwhile, memory utilization is 0.98 MB for each Arduino. With less than 1 MB of memory utilization indicated that with fewer resources for cryptography leaves more resources available to applications. The collected data then analyzed by using ANOVA to determine whether there is statistically significant difference of execution time among Arduino Mega, Arduino Uno and Arduino Nano.

The null hypothesis for an ANOVA assumes the population means are equal. Hence, the null hypothesis as:

Ho: The mean of execution time of Lizard cipher is statistically equal across the three types of Arduino.

TABLE. II. DESCRIPTIVE STATISTIC

Execution Time	N	Mean	Std. Deviation
Arduino Mega	800	55813.4850	151.26459
Arduino Nano	800	55596.5700	151.29735
Arduino Uno	800	55596.5700	151.29735
Total	2400	55668.8750	182.56201

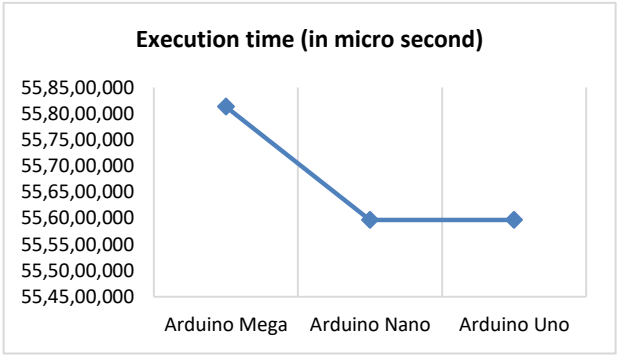


Fig. 3. Execution Time of Lizard.

The alternative hypothesis is:

Ha: At least one mean of execution time of Lizard cipher is not statistically equal.

There are several assumptions that has to be met before conducting ANOVA, i.e. assumption of independence, assumption of normality, and assumption of homogeneity of variance. As for assumption of independence, the sample has to be random and the value of one observation is not related to any other observation. This assumption is met since the data are taken from each Arduino individually without interfering one another. The second assumption is assumption of normality which indicates that the data are normally distributed. Based on the result of Kolmogorov-Smirnov test, the data is normal, so it can be proceed further. The last assumption is assumption of homogeneity of variance. By using Levene's F test the variances of the distributions in the data is equal. The result of ANOVA test is presented in Tables III to V. An Alpha level of .05 is used for all analyses. Based on Table III, the test for homogeneity of variance is not significant [LeveneF(2, 2397) = 0.00, p > .05] indicating that this assumption underlying the application of ANOVA is met.

ANOVA is used to examine whether the execution time of Lizard cipher is a function of the three Arduino variant. The independent variable represented the three different types of Arduino: 1) Arduino Mega; 2) Arduino Uno; and 3) Arduino Nano. The dependent variable is the execution time of Lizard cipher.

This study is successful in attaining main goal in determining whether Lizard cipher can be implemented in three variant of Arduino platform. It presents promising results that Lizard cipher occupies less than 0.98 MB of memory utilization for cryptography in Arduino which leaves enough memory available for other functionalities. The execution time of Lizard cipher is observed and compared among Arduino variants, i.e. Arduino Mega, Arduino Nano and Arduino Uno. There is a significant difference in execution time of Lizard cipher between Arduino Mega and Arduino Uno (p = 0.00), as well as between Arduino Mega and Arduino Nano (p = 0.00). However, there are no differences between Arduino Nano and Arduino Uno (p = 1.00).

TABLE. III. TEST OF HOMOGENEITY OF VARIANCES

Levene Statistic	df1	df2	Sig.
.000	2	2397	1.000

TABLE. IV. ANOVA EXECUTION TIME

Source	Sum of Square	Df	Mean Square	F	Sig.
Between Groups	25094462 .520	2	12547231 .260	548 .211	.000
Within Groups	54861539 .980	2397	22887 .584		
Total	79956002 .500	2399			

TABLE. V. POST-HOC TEST TUKEY'S HSD

(I) Arduino	(J) Arduino	Mean Difference (I-J)	Std. Error	Sig.
Mega	Nano	216.91500*	7.56432	.000
	Uno	216.91500*	7.56432	.000
Nano	Mega	-216.91500*	7.56432	.000
	Uno	.00000	7.56432	1.000
Uno	Mega	-216.91500*	7.56432	.000
	Nano	.00000	7.56432	1.000

Tukey's HSD procedures are used to determine which pairs of the three group means differed. Tukey's HSD is used since the assumption of homogeneity is met. According to Table V, Tukey's HSD post hoc test reveales that the execution time is significantly slower in Arduino Mega (55813.4850 ± 151.26459, p = .000) and Arduino Nano (55596.5700 ± 151.29735, p = .000) compared to Arduino Uno (55596.5700 ± 151.29735). There is no significant difference in execution time of Lizard cipher between Arduino Uno and Arduino Nano (p = 1.00).

V. CONCLUSION

This study is successful in attaining main goal in determining whether Lizard cipher can be implemented in three variant of Arduino platform to perform a number of important security-related functions. The result presents promising results that Lizard occupy less than 1 MB of memory utilization for cryptography in Arduino which will be enough memory available for other functionalities. The execution time of Lizard cipher are observed and compared among Arduino variants, i.e. Arduino Mega, ArduinoNao and Arduino Uno. There is a significant difference in execution time between Arduino Mega and Arduino Uno (p = 0.00), as well as between Arduino Mega and Arduino Nano (p = 0.00). However, there are no differences between Arduino Nano and Arduino Uno (p = 1.00). This result will assist IoT security engineers in choosing a lightweight cipher that suitable for constraint device.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Computer networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] S. A. Kumar, T. Vealey, and H. Srivastava, "Security in internet of things: Challenges, solutions and future directions," in 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016, pp. 5772–5781.

[3] Hamann, M., Krause, M., & Meier, W., 2017. LIZARD – A Lightweight Stream Cipher for Power-constrained Devices. Germany: University of Mannheim.

[4] Hell, M., Johansson, T., Meier, W.: Grain – A Stream Cipher for Constrained Environments. eSTREAM, ECRYPT Stream Cipher Project, Report 2005/010 (2005) http://www.ecrypt.eu.org/stream https://www.arduino.cc/en/main/boards [accessed on January 2019].

[5] A. D. Elbayoumy and Simon J. Shepherd. Stream or Block Cipher for Securing VoIP?. International Journal of Network Security, Vol.5, No.2, PP.128–133, Sept. 2007 128.

[6] Daniel, D.; Le Corre, Y.; Khovratovich, D.; Perrin, L.; Grobschadl, J.; Biryukov, A. Triathlon of Lightweight Block Ciphers for the Internet of Things. 2015. Available online: http://orbilu.uni.lu/bitstream/10993/25565/1/209.pdf.

[7] National Institute of Standards and Technology (NIST). Advanced Encryption Standard (AES). FIPS Publication 197, available for download at http://nvlpubs.nist.gov/nistpubs/fips/nist.fips.197.pdf, 2001.

[8] N. Mouha, B. Mennink, A. Van Herrewege, D. Watanabe, B. Preneel, and I. Verbauwhede. Chaskey: An efficient MAC algorithm for 32-bit microcontrollers. In A. Joux and A. M. Youssef, editors, Selected Areas in Cryptography — SAC 2014, volume 8781 of Lecture Notes in Computer Science, pages 306–323. Springer Verlag, 2014.

[9] V. Grosso, G. Leurent, F.-X. Standaert, and K. Varici. LS-designs: Bitslice encryption for efficient masked software implementations. In C. Cid and C. Rechberger, editors, Fast Software Encryption — FSE 2014, volume 8540 of Lecture Notes in Computer Science, pages 18–37. Springer Verlag, 2015.

[10] D. Hong, J. Sung, S. Hong, J. Lim, S. Lee, B. Koo, C. Lee, D. Chang, J. Lee, K. Jeong, H. Kim, J. Kim, and S. Chee. HIGHT: A new block cipher suitable for low-resource device. In L. Goubin and M. Matsui, editors, Cryptographic Hardware and Embedded Systems — CHES 2006, volume 4249 of Lecture Notes in Computer Science, pages 46–59. Springer Verlag, 2006.

[11] W. Wu and L. Zhang. LBlock: A lightweight block cipher. In J. López and G. Tsudik, editors, Applied Cryptography and Network Security — ACNS 2011, volume 6715 of Lecture Notes in Computer Science, pages 327–344. Springer Verlag, 2011.

[12] D. Hong, J.-K. Lee, D.-C. Kim, D. Kwon, K. H. Ryu, and D. Lee. LEA: A 128-bit block cipher for fast encryption on common processors. In Y. Kim, H. Lee, and A. Perrig, editors, Information Security Applications — WISA 2013, volume 8267 of Lecture Notes in Computer Science, pages 3–27. Springer Verlag, 2013.

[13] . J. Guo, T. Peyrin, A. Poschmann, and M. J. Robshaw. The LED block cipher. In Cryptographic Hardware and Embedded Systems — CHES 2011, volume 6917 of Lecture Notes in Computer Science, pages 326–341. Springer Verlag, 2011.

[14] K. Shibutani, T. Isobe, H. Hiwatari, A. Mitsuda, T. Akishita, and T. Shirai. Piccolo: An ultra-lightweight blockcipher. In B. Preneel and T. Takagi, editors, Cryptographic Hardware and Embedded Systems — CHES 2011, volume 6917 of Lecture Notes in Computer Science, pages 342–357. Springer Verlag, 2011.

[15] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. Robshaw, Y. Seurin, and C. H. Vikkelsoe. PRESENT: An ultra-lightweight block cipher. In P. Paillier and I. Verbauwhede, editors, Cryptographic Hardware and Embedded Systems — CHES 2007, volume 4727 of Lecture Notes in Computer Science, pages 450–466. Springer Verlag, 2007.

[16] M. R. Albrecht, B. Driessen, E. B. Kavun, G. Leander, C. Paar, and T. Yalçin. Block ciphers – Focus on the linear layer (feat. PRIDE). In J. A. Garay and R. Gennaro, editors, Advances in Cryptology — CRYPTO 2014, volume 8616 of Lecture Notes in Computer Science, pages 57–76. Springer Verlag, 2014.

[17] J. Borghoff, A. Canteaut, T. Güneysu, E. B. Kavun, M. Knezevic, L. R. Knudsen, G. Leander, V. Nikov, C. Paar, C. Rechberger, P. Rombouts, S. S. Thomsen, and T. Yalçin. PRINCE – A low-latency block cipher for pervasive computing applications. In X. Wang and K. Sako, editors, Advances in Cryptology — ASIACRYPT 2012, volume 7658 of Lecture Notes in Computer Science, pages 208–225. Springer Verlag, 2012.

[18] R. L. Rivest. The RC5 encryption algorithm. In B. Preneel, editor, Fast Software Encryption — FSE '94, volume 1008 of Lecture Notes in Computer Science, pages 86–96. Springer Verlag, 1995.

[19] W. Zhang, Z. Bao, D. Lin, V. Rijmen, B. Yang, and I. Verbauwhede. RECTANGLE: A bit-slice lightweight block cipher suitable for multiple platforms. Science China Information Sciences, 58(12):1–15, Dec. 2015.

[20] A. Baysal and S. Sahin. RoadRunneR: A small and fast bitslice block cipher for low cost 8-bit processors. In T. Güneysu, G. Leander, and A. Moradi, editors, Lightweight Cryptography for Security and Privacy — LightSec 2015, volume 9542 of Lecture Notes in Computer Science, pages 58–76. Springer Verlag, 2016.

[21] V. Grosso, G. Leurent, F.-X. Standaert, and K. Varici. LS-designs: Bitslice encryption for efficient masked software implementations. In C. Cid and C. Rechberger, editors, Fast Software Encryption — FSE 2014, volume 8540 of Lecture Notes in Computer Science, pages 18–37. Springer Verlag, 2015.

[22] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers. The SIMON and SPECK families of lightweight block ciphers. Cryptology ePrint Archive, Report 2013/404, 2013.

[23] D. Dinu, L. Perrin, A. Udovenko, V. Velichkov, J. Großschädl, and A. Biryukov. Design strategies for ARX with provable bounds: Sparx and LAX. In J. H. Cheon and T. Takagi, editors, Advances in Cryptology — ASIACRYPT 2016, volume 10031 of Lecture Notes in Computer Science, pages 484–513. Springer Verlag, 2016.

[24] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers. The SIMON and SPECK families of lightweight block ciphers. Cryptology ePrint Archive, Report 2013/404, 2013.

[25] T. Suzaki, K. Minematsu, S. Morioka, and E. Kobayashi. TWINE: A lightweight, versatile block cipher. In G. Leander and F.-X. Standaert, editors, Proceedings of the 1st ECRYPT Workshop on Lightweight Cryptography (LC 2011), pages 146–169, 2011.

[26] Arduino website http://www.arduino.cc.

[27] Will G Hopkins. 2016. A New View of Statistics. http://sportsci.org/resource/stats/ [accessed on February 2019]

A Generic Approach for Weight Assignment to the Decision Making Parameters

Md. Zahid Hasan[1], Shakhawat Hossain[2], Mohammad Shorif Uddin[3], Mohammad Shahidul Islam[4]
Department of Computer Science and Engineering Jahangirnagar University, Dhaka, Bangladesh[1, 3]
Department of CSE, International Islamic University Chittagong, Chattogram, Bangladesh[2]
Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh[4]

***Abstract*—Weight assignment to the decision parameters is a crucial factor in the decision-making process. Any imprecision in weight assignment to the decision attributes may lead the whole decision-making process useless which ultimately mislead the decision-makers to find an optimal solution. Therefore, attributes' weight allocation process should be flawless and rational, and should not be just assigning some random values to the attributes without a proper analysis of the attributes' impact on the decision-making process. Unfortunately, there is no sophisticated mathematical framework for analyzing the attribute's impact on the decision-making process and thus the weight allocation task is accomplished based on some human sensing factors. To fill this gap, present paper proposes a weight assignment framework that analyzes the impact of an attribute on the decision-making process and based on that, each attribute is evaluated with a justified numerical value. The proposed framework analyzes historical data to assess the importance of an attribute and organizes the decision problems in a hierarchical structure and uses different mathematical formulas to explicit weights at different levels. Weights of mid and higher-level attributes are calculated based on the weights of root-level attributes. The proposed methodology has been validated with diverse data. In addition, the paper presents some potential applications of the proposed weight allocation scheme.**

Keywords*—Multiple attribute decision problem; average term frequency; cosine similarity; weight setup for multiple attributes; decision making

I. Introduction

In decision-making approaches, the decision-makers need to obtain the optimal alternative from a set of predefined alternatives based on some decision parameters say attributes. Attribute's weight states the relative importance of an attribute and is numerically described to address the impact of an attribute on the decision-making process. A precise decision-making process mostly depends upon its attributes' weights. Decision attributes in decision problems can be organized in a hierarchical layer. The root-layer attributes of a decision problem expose the basic decision parameters whereas intermediate-layer attributes have a significant dependency on the root-layer attributes.

However, all the attributes in a decision solution are not equally important. So, to identify the importance of an attribute relative to other attributes in decision problem solution, a weight is assigned to each attribute. For example, suppose, a music school determines to appoint a music teacher and for that, the school committee sets two basic quality measurement attributes- knowledge on a musical instrument (KMI) and knowledge on geography (KG). However, the case may happen that, a candidate scores 120marks (40 marks in KMI and 80 marks in KG) out of 200 where another candidate obtains 100 marks (70 in KMI and 30 marks in KG). If each of these two attributes is given similar importance then the first candidate will be selected which is by no means, an optimal decision. So, it is important to set a specific weight to each attribute in a decision-making process. But there is no mathematical approach that helps the decision-makers to allocate weights to the decision parameters. Decision-makers very often depend on domain experts to determine attributes' weights manually which produces some uncertainties in the decision-making process and consequently, leads to a non-optional decision. In real-life scenarios, a decision-maker himself sets an identical weight to each decision-making attribute rather than depending on domain experts which in consequence makes the decision-making framework find a troublesome decision solution.

However, this paper proposes a very straightforward formula to allocate weights to the decision attributes. The paper proposes the Term Frequency to allocate weights to the root level attributes and the Cosine Similarity to generate weights for the intermediate level attributes. Both approaches analyze some historical data to emerge the weights of the decision attributes.

The paper is organized as follows: Section II provides a detailed description of the existing weight allocation methodologies. Based on the study of related works, a new generic weight assignment methodology is proposed in Section III. In Section IV, a numerical experiment on the proposed methodology is provided. The system is validated in Section V through the representation of its results at different critical situations. The paper is concluded in Section VI.

II. Related Works

In multiple attribute decision-making problems, the relative importance of each attribute allocated by an expert has a great impact to evaluate every alternative [1]. The multiple attribute decision-making (MADM) method offers a practical and efficient way to obtain a ranking of all the alternatives based on non-relative and inconsistent attributes [2]. For that reason, it is especially imperative to find a logical and sensible weight allocation scheme. In real life, the imperfect, inexact information and the impact of particular and individual preference lead to expanding the indecision and trouble in

weight calculation and distribution in the decision-making problem [3].

The decision scientists have proposed various methodologies for obtaining attributes' weights of the decision-making problems. The present methods can be generally divided into three groups: subjective, objective and integrated approach [4]. In the subjective approach, the decision-makers set the attributes' weights by using their preference knowledge [5]. AHP [6] and Delphi method [7] are the standard methods for determining subjective weights based on the preference of decision-makers. Ranked and point allocation methods proposed by Doyle et al [8] and rank order distribution method provided by Roberts and Goodwin [9] are also some subjective weighting approaches.

In the objective approach, the decision-makers set the attribute-weights by using applicable facts, rational implications, and viewpoints. Entropy-based method [10], TOPSIS method [11] and mathematical programming based method [12] follow the objective approach to allocate weights to the decision parameters. Decision-makers' preferences and evidence-facts of the specific problem are considered simultaneously in the integrated approach for obtaining the attributes-weights. Cook and Kress proposed a preference-aggregation model [13] which is actually an integrated approach for allocating weights to the decision attributes. Fan et al. [14], Horsky and Rao [15], and Pekelman and Sen [16] also constructed some optimization-based models.

However, it is crucial to select suitable attributes' weights in decision-making conditions since the diverse values of attributes' weights may effect in unlike ranking order of alternatives. Nevertheless, in most MADM scenarios, the preference of the attributes over the alternatives distributed by decision-makers is typically not adequate for the crisp numerical data, because things are uncertain, fuzzy and possibly inclined by the subjectivity of the decision-makers, or the knowledge and data about the problem domain are insufficient during the decision-making the process.

Based on the above analysis, many factors affect the weight allocation, there should be considered that the importance of attribute is reflected by the objective data and the subjective preference of the decision-makers. But there is no better method to fuse the subjective and objective weights in existing MADM literature. With the motivation of establishing a weight allocation approach in a uniform as well as effective way, this paper proposes the current generic scheme.

III. Methodology

A. Problem Statement

Decision-makers use different methodologies to capture different types of decision problems. Different methodologies calculate alternatives' scores in different ways. However, almost all the available advanced decision-making approaches set some specific weights to distinguish the decision-making parameters based on their relative importance on the decision-making process. Based on the decision-making process and decision approach, the attributes' weights setup process can be categories into the following classes.

1) The rule-based decision-making processes set a random weight for each of its attributes. Some rule-based approaches like the Evidential Reasoning (ER) approach use those random values to calculate the final result. On the other hand, some advance decision-making approaches update the primarily set up random weights to enhance the decision-making accuracy. For example, the RIMER approach updates the activation weight to emerge an accurate decision result.

2) Decision-making approaches like deep learning methodologies use some random values as its primary activation weight. For example, ANN, RNN, CNN or GAN use some random values to activate their attributes' weights. Later, these approaches update the weights to reach the maximum accuracy in any type of decision-making process. So, in terms of decision attributes' weight generation, decision problems of the existing decision-making approaches can be listed as:

a) Rule-based decision-making approaches use some random values to update its decision attributes' weights which lead the decision-making approaches to some troublesome results.

b) Weight updating processes used in decision-making approaches are not enough matured and as a consequence decision-making approaches very often fail to reveal the decision results with enough confidence.

c) Deep learning approaches take huge time to update its activation weights as it requires multiple iterations to update attributes weights.

The above discussion makes it clear that decision approaches need to follow an appropriate weight setup process in order to generate an accurate decision result under any unfair circumstance like risk or uncertainties.

B. Problem Structuring

Based on types and nature, decision problems can be classified into two basic classes:

- Single-layer decision problems; and
- Multilayer decision problems.

In single-layer decision problems, decision solutions are made with the direct involvement of the decision attributes. Only the basic decision parameters that have a direct impact on decision problem solutions construct the single –layer decision architecture. This type of attribute can be represented using the architecture (see Fig. 1).

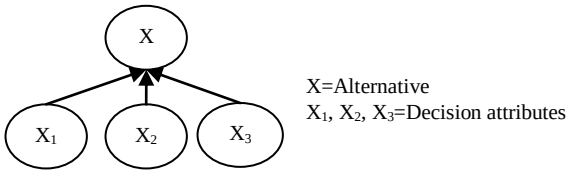


Fig. 1. Single Layer Decision Problem.

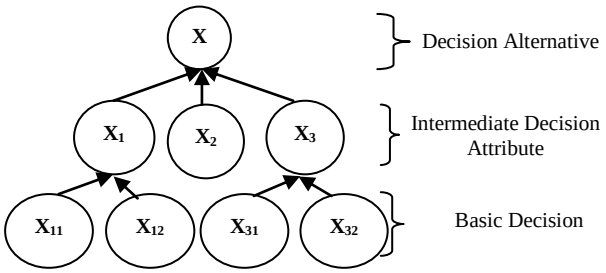


Fig. 2. Multi Layer Decision Problem.

On the other hand, multi-level decision problems depend on one or more intermediate levels to make the final decision rather than just depending on the basic attributes. These types of decision problems can be expressed by an Architectural Theory Diagram (ATD) or a hierarchical structure (see Fig. 2).

So, weight generation for decision attributes can necessarily be associated at root level or at both root and intermediate levels of the decision problem architecture.

C. Weight Generation for Root-layer Attributes

Weights for the basic attributes of a decision problem can be generated by analyzing the historical data. This paper proposes the Term Frequency (TF) as the data analyzing tool for weight assignment to the decision attributes. Term frequency is a very sophisticated technique for revealing the importance of a term for a specific purpose. Term frequency assigns weight to an attribute based on the number of times an attribute appears in a dataset. Term frequency of a term 't' can mathematically represented as

$$\text{Term Frequency, } TF = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

However, very often weight estimation of an attribute becomes troublesome because of different TF from different datasets. In that case, this paper proposes a normalized term frequency to expose the importance of an attribute. The normalization task is accomplished by calculating the average of the Term frequencies measured from different datasets. The average term frequency of an attribute can be represented as:

$$AvgTF_X = \frac{\sum_{i=1}^{N} TF_X}{L} \quad (2)$$

Here, $AvgTF_X$ refers to an average term frequency. The average TF of an attribute denotes the actual weight of an attribute.

$$Weight of attribute X = \frac{(TF_{x1} + TF_{x2} + TF_{x3} + \ldots + TF_{xn})}{n} \quad (3)$$

Here, $X = \{X_1, X_2, X_3, \ldots, X_n\}$ denotes the value of attributes and n is the total number of attributes. So, the average TF will be considered as the final weights of the root-level attributes.

D. Different Cases of Weight Generation using Term Frequency

Weight generation by analyzing the historical data can be categorized into two basic classes based on the data types.

1) Weight calculation from textual data: In case of textual datasets, the weight of an attribute can be assigned based on the times the attributes appears in each document. Thus, the weights of an attribute in different datasets are calculated. Finally, the average weight can be measured. For example, Table I shows the weight calculation process from two different documents.

The term frequency of "Dhaka" for each of these documents can be calculated as:

$$TF(\text{"Dhaka"}, d1) = \frac{3}{22} = 0.136 \ and$$

$$TF(\text{"Dhaka"}, d2) = \frac{5}{20} = 0.25$$

Where, *d1* and *d2* both are the two separate documents. So, the average *TF* or the actual weight of "Dhaka" is calculated by using *equation 3*.

$$AvgTF = 0.386$$

TABLE. I. CALCULATION OF TERM FREQUENCY FROM TEXT

Document 1 (*d1*)		**Document 2 (*d2*)**	
Term	**Term Count**	**Term**	**Term Count**
Dhaka	3	Dhaka	5
is	5	is	2
the	4	a	4
capital	1	big	1
of	6	population	1
Bangladesh	3	country	7

TABLE. II. WEIGHT CALCULATION FROM DIFFERENT TEXTUAL DATASETS

Diabetes Dataset 1 (Total 334 Patients)			**Diabetes Dataset 2** (Total 203Patients)			**Weight Calculation**
Attributes	**Term Count**	**Term Frequency**	**Attributes**	**Term Count**	**Term Frequency**	**Average TF**
Hunger	304	0.91	Hunger	165	0.81	0.86
Peeing more often	278	0.83	Peeing more often	170	0.83	0.83
Dry Mouth	311	0.93	Dry Mouth	173	0.85	0.89
Blurred Vision	90	0.26	Blurred Vision	70	0.34	0.3

2) Calculation of weight from numerical data: To generate the weight of an attribute from a dataset, the ration of the number of data samples containing that attribute and the total number of the sample data is calculated. Finally, the average weight of the attribute calculated from different datasets is measured. For example, we consider two Type-I diabetes patients' datasets where the first and second dataset both contain 334 and 203 diabetes-patients. There are four attributes (Hunger, Peeing, Dry mouth, and blurred vision) related to Type-I diabetes considering their corresponding affected patients as listed in Table II. Patients experiencing any of these four symptoms are represented and their corresponding term frequency is calculated by using *equation 1*. After calculating all the term frequency for both datasets, the average term frequency is calculated by using *equation 3*.

E. Weight Generation for Intermediate-layer Attributes

Very often decision-makers need to consider some intermediate-level attributes to calculate the efficiency of a particular attribute in order to rank the decision attributes. Attributes at the intermediate level are not always seemed to be present directly in the datasets. However, their derivatives mostly determine their significance in decision-making tasks. Weight generation for the attributes of the intermediate level can be accomplished by using Cosine Similarity. Cosine Similarity measures the similarity between two non-zero vectors in a vector space model [17]. Suppose, $\alpha$and $\beta$are two non-zero vectors in a vector space model. So, cosines similarity between $\alpha$and βcan be demonstrated as,

$$\text{Cosine Similarity} = \frac{\alpha \,.\, \beta}{\|\alpha\|\|\beta\|} \quad (4)$$

To generate weights for the intermediate-level attributes, one of the non-zero vectors, $\alpha$is constructed with the numeric values measured from the basic attributes of the decision problems which is always 1. The other vector, β is constructed with the Term Frequency of the attributes.

$$\text{Weight of Attribute } A_i = \frac{\sum_{i=1}^{N} \alpha_i \beta_i}{\sqrt{\sum_{i=1}^{N} {\alpha_i}^2}\sqrt{\sum_{i=1}^{N} {\beta_i}^2}} \quad (5)$$

For example, suppose, an intermediate-level attribute 'I' has two derivatives $'I_1'$ and $'I_2'$ and the TF of$'I_1'$and$'I_2'$ are calculated as 0.9 and 0.87 respectively. So, the vectors α andβ for A_i will be constructed as,

α= [1, 1] and

β= [0.9, 0.87]

Axiom-1: The value of the non-zero vectors used in the vector space model to generate the weight of intermediate-level attributes in the architectural theory diagram of decision process is always considered 1. For example in *equation 4*, each parameter of vector α is considered 1 which can be represented as,

$\bar{\alpha} = [1,1,1, \ldots.,1]$

In weight generation process explained in the earlier sections of this paper, one of the vector say α contains the frequency of the attributes. For example, to determine the weight of skin in the following decision problem, vector $\bar{\alpha}$ is constructed as $\bar{\alpha} = [1,1]$

Here, the weight of the skin is calculated based on the term frequency of purpura and edema in the historical datasets. To determine the importance of skin condition in the CKD diagnosis process, the skin condition is decomposed into its symptoms like edema and purpura (see Fig. 3). These two symptoms construct a vector. And the frequency of edema and purpura in the dataset constructs another vector. Here the value of the vector is always considered 1 to ensure the presence of edema and purpura in skin condition determination only once. If the value for edema would be considered 0, the skin condition determination process would be accomplished without considering edema. On the other hand, if the value in the vector for edema is considered 2, edema will actually be considered twice to determine skin condition (see Fig. 4).

From the above three scenarios, it becomes clear that the value for $\bar{\alpha}$ should be $\bar{\alpha} = [1, 1]$ todetermine the skin condition perfectly by avoiding the scenarios 2 and 3.

Axiom-2: The weight of an attribute in decision-making process is always $0 < w \leq 1$. Here w is the weight of an attribute and the value of w is neither 0 nor less than 0, $w <\neq 0$. Because if w=0, then the attribute has no impact on the decision-making process. Again the weight of an attribute can never be negative. If the weight of an attribute is negative then that attribute will have an adverse impact on decision process.

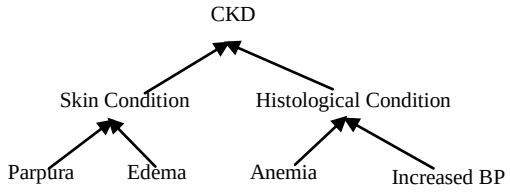


Fig. 3. Hierarchical Structure of CKD Diagnosis Problem.

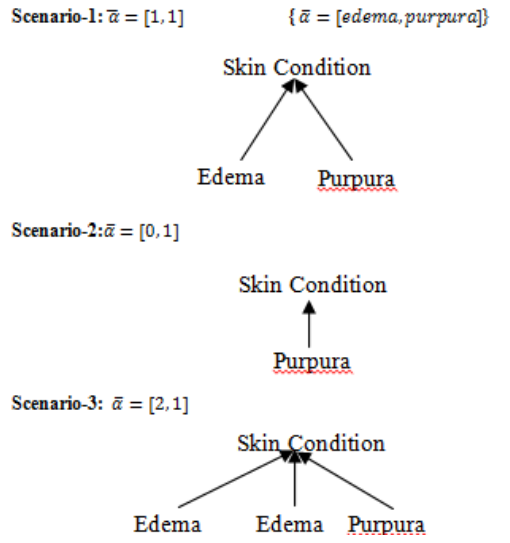


Fig. 4. Different Scenarios of Vector Construction Process.

IV. Numerical Experiment

To conduct a numerical experiment, multiple sets of data from four different hospitals in Bangladesh have been collected. A total amount from 4000 CKD (Chronic Kidney Disease) patients has been taken into consideration in this experiment.

To make the experiment transparent and adaptable, the inputs and outputs of the experiment are expressed as follows:

X: Chronic Kidney Disease

X_1: General Condition

X_2: Gastrointestinal Condition

X_3: Skin Condition

X_4: Hematologic Condition

X_{11}: Pain on the side or mid to lower back

X_{12}: Fatigue

X_{13}: Mental Depression

X_{14}: Headaches

X_{21}: Vomiting

X_{22}: Loss Body Weight

X_{23}: Change Taste

X_{31}: Edema

X_{32}: Purpura

X_{41}: Blood in Urine

X_{42}: High Blood Pressure (HBP)

X_{43}: Loss of Appetite

X_{44}: Protein in Urine

In order to set weights to the CKD diagnosis attributes say symptoms of CKD, the presence of symptoms in datasets is carefully analyzed. To make the analysis convenient, the CKD diagnosis problem is organized in a hierarchical structure. Decision attributes are structured in a two layered architecture based on attribute nature and their impact on diagnosis process (see Fig. 5).

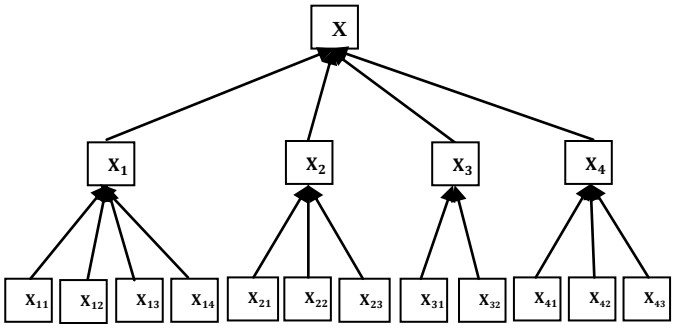


Fig. 5. Expression of Decision Problem in Hierarchical Structure.

A. Weight Calculation for Root -Layer Attributes by using Average Term Frequency

Term frequency calculates how many patients are experiencing a particular symptom over the total number of patients in a specific dataset. Based on how many times a symptom is present in a dataset, the weight for that symptom (attribute) is determined. The following algorithm provides a clear direction to the computation of the weights for the root-layer attributes.

Algorithm-1: TF calculation

1. Start
2. Define

$$d \leftarrow dataset$$
$$t \leftarrow rootlevelattributes$$
$$n \leftarrow numberofattributeappearsinadatset$$
$$m \leftarrow totalnumberofpatientinadatset$$
$$TF \leftarrow attribute's\ \ weight$$

3. Compute $TF \leftarrow \frac{n}{m}$
4. end

The following tables (Table. III-VI) show the weighs of different symptoms based on the TF calculation from different datasets. The TF of a symptom is calculated by using equation (1) and the weight is calculated by using equation (3).

Algorithm-2 in the next page provides an instruction for the implementation of equation (3) to generate weights for root-level attributes.

Algorithm-2: Weight Calculation for Root-layer attributes

Step 1: Start

Step 2: Define $i \leftarrow number\ of\ dataset$
$j \leftarrow number\ of\ attributes$
$TF \leftarrow Term\ Frequency$
$M \leftarrow total\ number\ of\ TF$
$N \leftarrow average\ TF$

Step 3: Set $TF \leftarrow 0$
$M \leftarrow 0$
$N \leftarrow 0$
$a \leftarrow 1$
$b \leftarrow 1$

Step 4: $M = M + TF$
Increase a by 1
Go to **Step 3** until $i = a\ \ and\ j = b$

Step 5: Compute $N = \frac{M}{i}$
Print N

Step 6: End

For example, for a dataset as shown in Table III, the value of TF for $\boldsymbol{X_{41}}$is calculated as 984/990=0.99 where 990 is the total number of patients and 984 is the number of patients experiencing the symptoms "Blood in Urine". Therefore, the TF of the attribute $\boldsymbol{X_{41}}$ is 0.99 in case of dataset 1. However; the average TF of X_{41} provides its actual weight which 0.97 as shown in Table VII.

TABLE. III. CKD DATASET FROM BANGLADESH MEDICAL COLLEGEB (990 PATIENTS)

	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
Attribute's Density	770	970	550	223	406	601	333	379	421	984	980	801	705
Term Frequency	0.77	0.97	0.55	0.22	0.41	0.6	0.33	0.38	0.42	0.99	0.98	0.8	0.71

TABLE. IV. CKD DATASET FROM SHAHEEDSUHRAWARDY MEDICAL COLLEGE (1013 PATIENTS)

	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
Attribute's Density	840	970	575	201	450	650	320	405	533	1001	998	905	695
Term Frequency	0.82	0.95	0.56	0.19	0.44	0.64	0.31	0.4	0.52	0.98	0.98	0.89	0.68

TABLE. V. CKD DATASET FROM BANGABANDHU MEDICAL COLLEGE (997 PATIENTS)

	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
Attribute's Density	699	965	507	245	423	629	400	405	467	981	987	790	695
Term Frequency	0.7	0.96	0.5	0.24	0.42	0.63	0.4	0.4	0.46	0.98	0.98	0.79	0.69

TABLE. VI. CKD DATASET FROM DHAKA MEDICAL COLLEGE (1104 PATIENTS)

	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
Attribute's Density	809	1065	701	307	447	667	397	508	498	1050	1077	850	745
Term Frequency	0.73	0.96	0.63	0.27	0.4	0.6	0.35	0.46	0.45	0.95	0.97	0.77	0.67

TABLE. VII. ROOT-LEVEL ATTRIBUTES' WEIGHT CALCULATION

Attributes	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
Attributes' weights	0.75	0.96	0.56	0.23	0.42	0.62	0.35	0.41	0.46	0.97	0.97	0.81	0.69

TABLE. VIII. WEIGHT CALCULATION OF INTERMEDIATE-LAYER ATTRIBUTES

Intermediate-level Attributes	X_1				X_2			X_3		X_4			
Root-Level Attributes	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{41}	X_{42}	X_{43}	X_{44}
α	1	1	1	1	1	1	1	1	1	1	1	1	1
β	0.75	0.96	0.56	0.23	0.42	0.62	0.35	0.41	0.46	0.97	0.97	0.81	0.69
Attribute's weight	0.92				0.97			0.35		0.99			

B. Weight Calculation for Intermediate -Layer Attributes by using Cosine Similarity

Attribute weights for intermediate layer are determined by the cosine similarity. Cosine similarity measures the cosine angle between two non-zero vectors. Intermediate layer attribute weight is calculated by the following algorithm shown in Table VIII. In Table VIII, the two non-zero vectors **α** and **β** are defined according to the Section 3.3.

Algorithm-3: Weight Calculation for Intermediate-layer attributes

Step 1: Start
Step 2: Define W ← weight
Step 3: Set $d \leftarrow 0$

$n \leftarrow 0$
$e \leftarrow 0$
$f \leftarrow 0$
$for\ all\ j = 1\ to\ n$
$for\ all\ i = 1\ to\ n$
Compute $d \leftarrow a_i \times b_i$
Increase i by 1
End for
Compute $n = n + d$
End for

Step 4: Set $m \leftarrow 0$

$for\ all\ i = 1\ to\ n$
$m \leftarrow a_i \times a_i$
$e \leftarrow sqrt\ (m)$
End for

Step 5: Set $m \leftarrow 0$

$for\ all\ i = 1\ to\ n$
$m \leftarrow b_i \times b_i$
$f \leftarrow sqrt\ (m)$
End for

Step 6: $W = \frac{n}{e \times f}$
Step 7: Print W
Step 8: End

V. Result and Discussion

Weight assignment process has been validated through some numerical experiments. For that, weights generated by the proposed system are compared to the benchmark results. The benchmark weights are calculated by analyzing a huge number of historical data by a group of domain experts. The accuracy of the proposed framework is found very close to the benchmark results.

To demonstrate that the proposed system performs better than any of the existing weight allocation approaches, some selected weight allocation methods have been experimented.

To analyze the performance of artificial neural network (ANN), the same input attributes and four datasets of the previous problem have been considered. In ANN, when input comes in the neural unit it is multiplied by a random weight of the corresponding node, and then the summation of the output of every node is performed. The final output comes out after transferring the current result into the sigmoid activation function for checking the current result against the threshold. To analyze the CKD problem, we consider a total of 12 attributes and feed the attribute values to the feedforward neural network with one hidden layer. The backpropagation algorithm sets the weights after 10 successful iterations.

On the other hand, some popular rule-based approaches like RIMER consider some random values as the attribute's initial weights which are later updated based on some thresholds. The domain experts manually set up the thresholds for making the weight updating process operational. To make a comparison among different weight allocation processes, the CKD diagnosis problem described earlier in this paper has been considered to generate the weighs by using the RIMER approach.

The other two prominent weight allocation methods- AHP and TOPSIS have also been utilized to generate weights for the attributed to a similar decision problem.

The Table IX provides a clear description of the performance level of different weighs assignment approaches. From the above table, it becomes very clear that TOPSIS and AHP methods result much differently than the other three methods. The ANN, RIMER and Proposed methodologies provide almost similar weights for almost all decision attributes. However, the optimal weight allocating framework can be selected by comparing the results of different methods against the benchmark results (see Fig. 6).

TABLE. IX. Visualization of Weight Generation Variations

Attributes	AHP	TOPSIS	ANN	RIMER	Proposed Approach
X_1	0.55	0.57	0.7	0.61	0.75
X_2	0.68	0.65	0.9	0.75	0.96
X_3	0.43	0.39	0.54	0.6	0.56
X_4	0.12	0.15	0.2	0.25	0.23
X_5	0.4	0.33	0.45	0.65	0.42
X_6	0.4	0.43	0.72	0.75	0.62
X_7	0.24	0.27	0.45	0.3	0.35
X_8	0.23	0.27	0.37	0.3	0.41
X_9	0.67	0.71	0.9	0.85	0.97
X_{10}	0.75	0.75	0.95	0.95	0.97
X_{11}	0.57	0.62	0.75	0.6	0.81
X_{12}	0.51	0.56	0.7	0.8	0.69

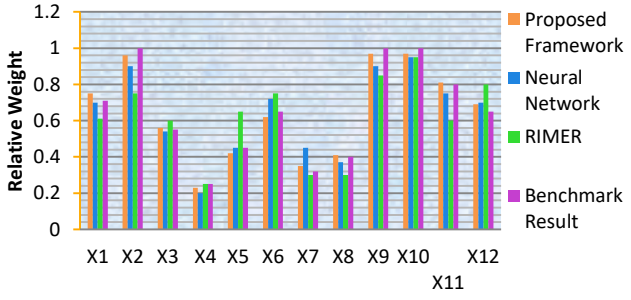


Fig. 6. Comparative Result of NN, RIMER, Benchmark Result with Proposed Framework.

The above graph visualizes that, the proposed weight allocation method is very close to the benchmark results. Therefore, it can be strongly claimed that the proposed weight assignment approaches provide the optimum weights to the decision parameters.

V. Conclusion

Weight assignment is an important task in decision-making approaches. The solution of any decision problem mostly depends upon its attributes' weights. Unfortunately, there is no sophisticated framework for assigning weights to the decision parameters. So, decision-makers need to depend on human sensing factors to assign weights to the attributes that leads the decision-makers to a non-optimal solution. However, this paper proposes a weight setup method for multiple attribute decision problems where the weights will be calculated from some historical data without any engagement of the domain experts. The decision problem in this paper is represented in a hierarchical structure where the attribute layers are divided into intermediate and root layers. Weights for root-layer attributes are first determined by the term frequency of an individual dataset and then average term frequency is computed from the total number of datasets to find out the final weights. Cosine similarity is applied to compute the weights for intermediate-layer attributes which actually measures the similarity between two non-zero vectors. This can efficiently evade the unreasoning assessment values due to the lack of knowledge or partial experience of the experts. Additionally, a numerical experiment provided in this paper validates the effectiveness of the proposed methodology which can proficiently help the decision-makers assign the accurate weights to the decision attributes.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W. Ho, X. Xu, and P. K. Dey, "Multi-criteria decision making approaches for supplier evaluation and selection: a literature review," European Journal of Operational Research, vol. 202, no. 1, pp. 16–24, 2010.

[2] Taho Yang, Yiyo Kuo, David Parker, and Kuan Hung Chen, "A Multiple Attribute Group Decision Making Approach for Solving Problems with the Assessment of Preference Relations," Mathematical Problems in Engineering, vol. 2015, 10 pages, 2015.

[3] Hongqiang Jiaoa and Shuangyou Wang, Multi-attribute decision making with dynamic weight allocation, Intelligent Decision Technologies 8 (2014) 225–230.

[4] JalilHeidaryDahooie, EdmundasKazimierasZavadskas, Amir SalarVanaki, Hamid Reza Firoozfar, MahnazLari&ZenonasTurskis (2019) A new evaluation model for corporate financial performance using integrated CCSD and FCM-ARAS approach, Economic Research-EkonomskaIstraživanja, 32:1, 1088-1113,2019.

[5] Quan Zhang and HongWeiXiu, "An Approach to Determining Attribute Weights Based on Integrating Preference Information on Attributes with Decision Matrix," Computational Intelligence and Neuroscience, vol. 2018, 8 pages, 2018.

[6] R. J. Ormerod and W. Ulrich, "Operational research and ethics: a literature review," European Journal of Operational Research, vol. 228, no. 2, pp. 291–307, 2013.

[7] C. L. Hwang and M. J. Lin, Group Decision Making Under Multiple Criteria: Methods and Applications, Springer, Berlin, Germany, 1987.

[8] J. R. Doyle, R. H. Green, and P. A. Bottomley, "Judging relative importance: Direct rating and point allocation are not equivalent," Org. Behav. Human Decis. Process., vol. 70, no. 1, pp. 65–72, Apr. 1997.

[9] R. Roberts and P. Goodwin, "Weight approximations in multi-attribute decision models," J. Multi Crit. Decis. Anal., vol. 11, no. 6, pp. 291–303, Nov. 2002.

[10] C. L. Hwang and K. Yoon, Multiple Attribute Decision Making: Methods and Applications. Berlin, Germany: Springer, 1981.

[11] S. H. Zanakis, A. Solomon, N. Wishart, and S. Dublish, "Multi-attribute decision making: A simulation comparison of select methods," Eur. J. Oper. Res., vol. 107, no. 3, pp. 507–529, Jun. 1998.

[12] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," Eur. J. Oper. Res., vol. 2, no. 6, pp. 429–444, Nov. 1978.

[13] W. D. Cook and M. Kress, "A data envelopment model for aggregatingpreference rankings,"Manag. Sci., vol. 36, no. 11, pp. 1302–1310, 1990.

[14] Z.-P. Fan, J. Ma, and Q. Zhang, "An approach to multiple attribute decision making based on fuzzy preference information on alternatives, "Fuzzy Sets. Syst., vol. 131, no. 1, pp. 101–106, Oct. 2002.

[15] D. Horsky and M. R. Rao, "Estimation of attribute weights frompreference comparisons,"Manag. Sci., vol. 30, no. 7, pp. 801–822,Jul. 1984.

[16] Shu, Peng & Ma, Jian & Zhang, Quan. (2002). An approach to multiple attribute decision making based on fuzzy preference information on alternatives. Fuzzy Sets and Systems. 131. 101-106. 10.1016/S0165-0114(01)00258-5.

[17] Md Zahid Hasan, Shakhawat Hossain, Md. Arif Rizvee and Md. Shohel Rana, "Content based Document Classification using Soft Cosine Measure" International Journal of Advanced Computer Science and Applications(IJACSA), 10(4), 2019.

A Comparative Study of the Most Influential Learning Styles used in Adaptive Educational Environments

Othmane ZINE[1], Aziz DEROUICH[2], Abdennebi TALBI[3]
Higher School of Technology of Fez
Sidi Mohamed Ben Abdellah University
Fez, Morocco

***Abstract*—E-learning has evolved from traditional content delivery approaches to a personalized, adaptive and learner-centered knowledge transfer. In the way of customizing the learning experience learning styles represent key features that cannot be neglected. Learning style designates any representative characteristic of an individual while learning, i.e. a particular way of dealing with a given learning task, the preferred media, or the learning strategies adopted in order to achieve a task. Despite the fact that the use of learning styles in adaptive educational environments has become controversial, but there is no empirical evidence of its usefulness. The main objective of our paper is to respond to the question "What learning style model is most appropriate for use in adaptive educational environments?"**

***Keywords*—Adaptive educational environments; learning style; Felder-Silverman; personalization; learner modeling**

I. INTRODUCTION

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers (all standard papers).

Along with the rapid progression of the ICTs for education and the increased efficiency of data acquisition and processing methods, the improvement of the learning quality of e-learning systems became appealing.

As demonstrated in the literature, the learner plays a central role in the complex learning process. And as noticed by teachers, learners aim at different goals, have diverse necessities, distinct backgrounds, skills and other significant characteristics [1]. They vary extremely in the speed and manner with which they collect new information and ideas, and in the confidence with which they process and use them. For example, some studies have highlighted that adult learners learn differently from younger ones (adult learners and young ones don't learn the same way) [2]. This makes each learner's requirements and preferences unique. The approach to instruction in which a single teaching scenario is used for all learners, better known in the literature as the 'one-size-fits-all' approach, is often unsuitable [3].

Student modeling is the process whereby an adaptive learning system creates and updates a student model by collecting data from several sources implicitly (observing user's behavior) or explicitly (requesting directly from the user). Traditionally, the majority of the student modeling systems has been limited to maintain assumptions related to student's knowledge (acquired during evaluation activities) without paying too much attention to student's preferences.

Learner's individual differences have remarkable potential that should be exploited to provide more accurate guidelines and learning support. This last lead to a better understanding of the subject, to the enhancement of learner's performance and motivation and consequently the optimization of the learning outcomes.

On the one hand, cognitivist and constructivist theories of learning revealed that several learning strategies should be integrated to accommodate individual differences and learning style [4]. On the other hand, researchers claimed that if a learner has a strong preference for a particular learning style, the strategies and even the learning resources should match that style to improve the learning experience [5].

As it is well known, the conventional face to face learning mode grants an accurate recognition of learners' behaviors from their facial expressions, questions and interactions, and so, allows the tutor to choose the right moment and tools to intervene constructively. Thus, it remains imperative to consider these interactions, even if they might be ambiguous in a distance-learning situation.

Traditional e-learning frameworks assume that each learner learns in the same manner through common learning material and instructional design. They are limited in terms of providing personalization and consider a "one fits for all" mindset without considering learner preferences and individual traits. Numerous learners encounter disappointment with such type of frameworks and get exhausted as they progress through their learning cycle since the material is static, and lacks sensitivity to the need of the learner [6]. Plausible remediation is the use of Adaptive Educational Systems (AES) that negate the learning management system the above-mentioned approach and adopt the "one fits one" one.

Learner diversity that exists in the classroom plays a role in influencing the teaching and learning process in the classroom. While all types of learners still need to be addressed, variety in instructional approaches can be used to address this diversity [7]. In addition, the various learning styles as a great deal of ongoing research indicate that learners have different strengths

and preferences in the way they absorb and process information [8]. Studies in psychology point out that people show noteworthy individual differences in problem-solving and decision-making activities. For instance, students with a solid inclination for a particular learning style may experience problems in learning if the teaching style does not coordinate with their learning style.

Even if researchers still argue on the usefulness of considering students' learning styles in adult education, the use of learning styles measures continues to be popular. And despite the absence of rigorous research findings to support this practice, there is no evidence for its ineffectiveness.

Some researchers claim that learning styles don't match the way the brain stores and reason about information. Others mention that there might be an optimal way to explain a particular subject, but the same style can't always be the best for a specific learner. And rather than focusing on one's best learning style, it is more interesting to on the worst, and try to improve learner's ability to learn in every style.

Even if the idea of learning styles hasn't reached maturity and hasn't proven a total success, there are still many strong reasons which show that employing learning styles can be beneficial to learners. One of the major reasons to use learning styles is because it encourages variety (i.e. as long as a learner feels at ease in the process of learning, no matter what his learning style is, he will learn better). Another reason and as involvement matters, a multimodal classroom is more engaging (i.e. if a learner prefers learning through activities, reading and listening to lectures will make him feel bored and discourage him). Even if learning styles turn out to be nothing more than a personal preference, it is still a creative and smart way to engage learners and enhance their motivation. Moreover, learning styles remind us that each learner is different, and while it is nearly impossible to satisfy all learners, success opportunities can be given to everyone by varying the teaching way at least.

This paper starts by critically examining the most influential learning style models according to the literature and presents a comparative summary of these learning style models (emphasizing their implication on teaching highlighting their strengths and weaknesses). Afterward, it shows how to measure the chosen learning style. Thereafter, it shed light on the impact of learning styles on learners preferred multimedia type. Next, it exposes the instructional design for learning path identification using the Felder-Silverman learning styles model. And finally, it presents a statistical study conducted to identify higher education learners default learning style.

II. Learning Style Models Classification

On the one hand, cognitivist and constructivist theories of learning revealed that several learning strategies should be integrated to accommodate individual differences and learning style [4]. On the other hand, researchers claimed that if a learner has a strong preference for a particular learning style, the strategies and even the learning resources should match this style to improve the learning experience [5].

The appellations 'learning style' and 'cognitive style' are commonly used interchangeably, even if cognitive style may denote a specific facet of learning style [4]. Moreover, learning styles are commonly associated with terms as "learning preferences", "learning skills", "learning strategies" and "learning approaches" [9]. This diversity of interpretations and terminologies led to the development of many learning style models.

Given the variety related to learning style, and the existence of a large number of learning models [9], a categorization of these models helps to identify their key features.

The model of Curry's onion can be used to group learning theories into three primary layers according to the degree of stability over time of the preferences represented by each one [10], [11] (see Fig. 1).

- Instructional preference styles (the outer layer of the onion): The least persistent over time, they deal with various modes of information delivery, may often change and therefore are less important in learning.
- Information processing styles (the middle layer): More stable over time than the instructional preference ones, they cope with the information processing way that influences the way learners memorize, infer and interpret information.
- Cognitive personality styles (the inner layer): The most unalterable over time, they are based on personality traits that have a more significative influence on learner's interaction with the learning environment.

Curry's onion model relies on theoretical assumptions and lacks experiential evidence to determine learning style stability [9]. And so, Coffield suggested "The families of learning styles" to classify learning style models with reference to several learning style overviews and on quantitative evidence. This spectrum was inspired from the onion model as well as analyses and overviews by key figures in the learning styles field [12]–[16].

The "families of learning styles" categorizes over seventy learning style models into (see Table I).

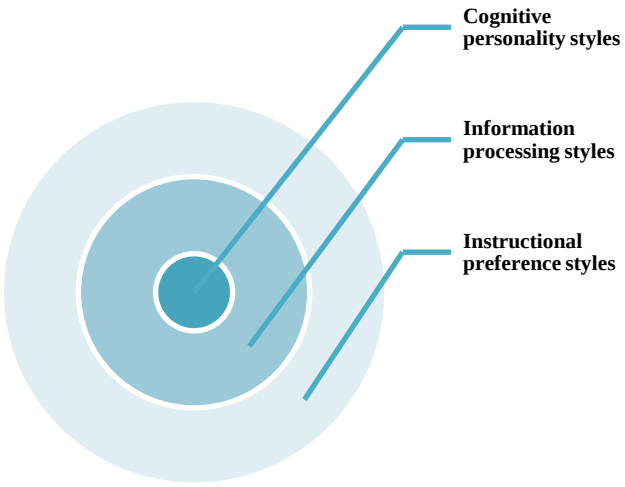


Fig. 1. Curry's Onion Learning Styles Model.

TABLE. I. COFFIELD'S FAMILIES OF LEARNING STYLES

Constitutionally-based learning styles and preferences	Dunn and Dunn; Gregorc; Bartlett; Betts; Gordon; Marks; Paivio; Richardson Sheehan; Torrance.
Cognitive structure	Riding; Broverman; Cooper; Gardner et al.; Guilford; Holzman and Klein Hudson; Hunt; Kagan; Kogan; Messick; Pettigrew; Witkin.
Stable personality types	Apter; Jackson; Myers-Briggs; Epstein and Meier; Harrison-Branson; Miller.
Flexibly stable learning preferences	Allinson and Hayes; Herrmann; Honey and Mumford; Kolb; Felder and Silverman; Hermanussen, Wierstra, de Jong and Thijssen; Kaufmann; Kirton; McCarthy.
Learning approaches and strategies	Entwistle; Sternberg; Vermunt; Biggs; Conti and Kolody; Grasha-Riechmann; Hill; Marton and Säljö; McKenney and Keen; Pask; Pintrich, Smith, Garcia, and McCeachie; Schmeck; Weinstein, Zimmerman and Palmer; Whetton and Cameron.

- Constitutionally based learning styles and preferences: Supposed to be fixed and very difficult to change, they are open to relatively easy environmental modification. These styles are mostly innate personality traits and represent the dominance of specific perceptual and sensory channels including the four sensory modalities: visual, auditory, kinesthetic and tactile.
- Cognitive structure: Presumed to be general habits of thought, they reflect intuitive and structural characteristics of the cognitive system and focus on the interactions of cognitive controls and cognitive processes [17].
- Stable personality types: Believed to be mostly stable but can change over time, they are viewed as embedded characteristics within the personality traits which are assumed to shape all aspects of an individual's interaction with the environment.
- Flexibly stable learning preferences: Assumed to have some long-term stability even if they can change slightly from one situation to another, they are viewed as crucial preferences rather than fixed characteristics. This family of learning styles classifies learners in accordance with a measure that mirrors the way they receive and process information.
- Learning approaches and strategies: Frequently changing depending on the situation, they came out from the drop of the learning styles for a holistic and active view of learning approaches and study strategies. They describe how learners prefer to tackle learning tasks generally according to their perceptions of a task and the adopted cognitive strategies [15].

III. LEARNING STYLE MODELS

In adaptive e-learning environments, various learning styles theories have been used. In this section, we focused on the most influential learning style models. Since they are the most suitable for the implementation of an adaptive e-learning environment, we are particularly interested in the flexibly-stable learning preferences. However, we decided to take one model from each other family into consideration for reviewing purposes.

A. Gregorc's Mind Styles Model (Constitutionally-based Learning Styles and Preferences Family)

Anthony Gregorc defines learning styles as stable, cognitive, affective, and physiological traits that serve as indicators of how learners perceive and deal with information and react during learning sessions. Furthermore, he argues the teaching strategies, the personality of individuals, and the media are highly correlated.

He claims that minds interact with any context through channels and that there are two dimensions of learners' innate abilities of perception and ordering, and distinguishes between four observable channels: abstract, concrete, random, and sequential tendencies (see Table II and Fig. 2). A combination of these tendencies is indicative of the individual style. And so, four learning styles are identified within Gregorc's model.

In order to determine learner's learning style, the Style Delineator has been developed, which is a 40-item self-report inventory involving the rank ordering of sets of words.

B. Riding Cognitive Style (Cognitive Structure Family)

Riding Cognitive Style model is mainly focused on how cognitive skills develop [16]. Its authors state that the cognitive style is the individual's way of thinking and at the same time the individual's favorite and habitual approaches of organizing and representing information. Furthermore, they define a learning strategy as the processes used by the learner to comply with a learning activity requirement. Besides, they claim that while strategies may be learned and developed over a period of time, styles are static and are relatively innate characteristics of any individual.

The authors claim that their model is oriented essentially to the cognitive skills developing approaches, and so, it influences the study orientation, the instructional inclination, the hands-on learning, the social attitude, and managerial skills.

Riding and Cheema reviewed the descriptions, correlations, methods of assessment, and effect on the behavior of over 30 models, and concluded that they could be grouped into two uncorrelated dimensions: one concerning to cognitive organization (holist-analytic); and one involving mental representation (verbal-imagery) (see Table III and Fig. 3). And Riding states that the first dimension originates from Witkin research on field dependence and field independence [18], while the second is based on the dual coding theory of Paivio [19].

TABLE. II. GREGORC'S MIND STYLES MODEL

Dimension	**Pole**	**Learning style**
Perception	Concrete (C)	1. AS 2. AR 3. CS 4. CR
	Abstract (A)	
Ordering	Sequential (S)	
	Random (R)	

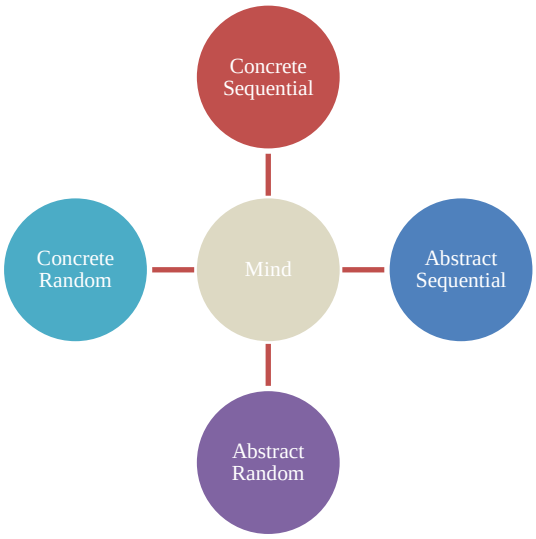


Fig. 2. Gregorc's Four Channels Model.

TABLE. III. RIDING COGNITIVE STYLE

Dimension	Pole	Learning style
Cognitive organization	Holist (H)	1. HV 2. HI 3. AV 4. AI
	Analytic (A)	
Mental representation	Verbal (V)	
	Imager (I)	

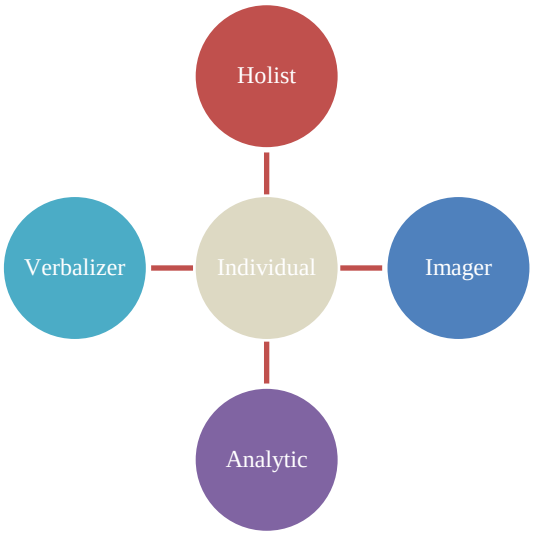


Fig. 3. Riding Cognitive Style uncorrelated Dimensions.

In order to retrieve learner's cognitive style, Riding has developed an assessment method named the Cognitive Styles Analysis (CSA) [20]. The cognitive representation dimension test items are entirely visual and the score is relying on a response speed comparison on a matching task and on embedded figures task analytic preference. And the items for the verbal-imagery dimension are all verbal and are in accordance with the relative speed of categorizing items as being similar through their conceptual similarity or color.

C. Myer-Briggs Type Indicator Theory (Stable Personality Types Family)

Developed by Isabel Briggs Myers and her mother, Katharine Cook Briggs, MBTI is a model based on the theory of psychological types of Carl Jung. It aims at helping each person to understand his unique personality. This model is based on the belief that the variances in behavior from one person to another can be expressed in terms of preferences between polarities and each person has a natural preference [21]. Accordingly, when someone uses his favorite pole, he/she generally succeeds better and feels more skilled.

Furthermore, MBTI's dimensions define the four main dichotomies of psychic life and represent humans' personality core functions [22]. Each dimension has two uncorrelated poles, and each person is predisposed to one pole in each dimension (see Table IV). So, this model allows the generating of sixteen unique personality types.

In order to detect learning styles, three forms of the Myers-Briggs Type Indicator instrument were developed (a standard 93-item version, an extended 126-item version, and an abbreviated 50-item version). And in all instances, scores are given to generate one of the sixteen unique personality types [23].

D. Felder-Silverman Learning Style Model (Flexibly Stable Learning Preferences Family)

The Felder-Silverman learning style model (FSLSM) [5] is, another model based on the work of Carl Gustav Jung, a widely used in adaptive educational systems focusing on learning styles. It describes the learning styles of engineering learners in a detailed way. It distinguishes between preferences on four measurements (dimensions) which are linked to the four dimensions of information (see Table V): information processing (How does the learner prefer to process information), information perception (What type of information does the learner prefer in order to perceive), information input (Through which sensory channel is external information most effectively perceived), and information understanding (How does the learner progress towards understanding) and therefore enables adaptive learning systems to provide a better-tailored learning material [1].

TABLE. IV. MBTI LEARNING STYLE

Dimension	Pole	Learning style
Attitude	Introvert (I)	1. ISTJ 2. ISFJ 3. INFJ 4. INTJ 5. ISTP 6. ISFP 7. INFP 8. INTP 9. ESTP 10. ESFP 11. ENFP 12. ENTP 13. ESTJ 14. ESFJ 15. ENFJ 16. ENTJ
	Extravert (E)	
Information processing	Sensing (S)	
	iNtuitive (N)	
Decisions making	Thinking (T)	
	Feeling (F)	
Environment evaluation	Judging (J)	
	Perceiving (P)	

TABLE. V. FELDER-SILVERMAN LEARNING STYLE MODEL

Dimension	Pole	Learning style
Perception	Intuitive (I)	1. IAVQ 2. IAVG 3. IAEQ 4. IAEG 5. IRVQ 6. IRVG 7. IREQ 8. IREG 9. SAVQ 10. SAVG 11. SAEQ 12. SAEG 13. SRVQ 14. SRVG 15. SREQ 16. SREG
	Sensing (S)	
Processing	Active (A)	
	Reflective (R)	
Input	Visual (V)	
	vErbal (E)	
Understanding	seQuential (Q)	
	Global (G)	

Moreover, FSLSM enables the learning style model to consider exceptional behavior which means that learners with a high preference for a certain behavior can act sometimes differently. This model rates the learner's learning style in a scale of four dimensions to define sixteen distinct learning styles.

In order to detect learning styles, Felder and Soloman elaborated a 44-item questionnaire, named the Index of Learning Styles [24], where 11 questions are asked for each dimension [25].

E. Kolb's Experiential Learning Theory (Flexibly Stable Learning Preferences Family)

Inspired by the works of John Dewey and Jean Piaget, the American psychologist Kolb developed a four-stages learning style model named "The Experiential Learning Theory" in the early 70s. According to Kolb, the experience is the key element of any learning process, and knowledge comes from the blending of grasping experience and transforming it [26].

He defined a learning model composed of these two orthogonal dimensions (see Table VI). The Grasping dimension poles are Concrete Experience (when the learner is confronted to a new situation or a remake of similar previous experience) and Abstract Conceptualization (when the learner's reflections initiate a new understanding or the expansion of the current knowledge). Likewise, the Transforming poles are Reflective Observation (when the learner observes the new experience and positions in accordance with his/her prior knowledge) and Active Experimentation (when the learner puts his or her newly acquired or expanded knowledge into practice).

Kolb proposed a four-stage hypothetical learning cycle and claimed that these four stages are interrelated, with each one leading to the following. He assumed that we can enter the learning cycle at any stage and learners will show a preference for some phases more than others. A learner can start with direct experience and makes it specific or of an abstract experience (AC/CE). Then these experiences (concrete or abstract) are transformed into knowledge when we reflect and think about them or when we experiment an active form of the received information (RO/AE).

For instance, student A goes through a concrete situation and accumulates experience (CE), this leads him to make some observations and reflections about the situation (RO), later on he will build abstract concepts and theories to explain these observations (AC), which he can actively experiment and validate to make decisions or resolve problems (AE). Once the circle is complete the learning outcome leads to the construction of new experiences which triggers the cycle of learning all over again (see Fig. 4).

In order to assess an individual's preferred modes of learning, Kolb elaborated a forced-choice ranking questionnaire that the subject has to complete, named the Kolb's Learning Style Inventory. It evaluates the individual's abilities throughout two spectrums: concrete experience to abstract conceptualization, and active experimentation to reflective observation. The first version of the LSI appeared in 1976, and it was revised several times (1985, 1999, and 2011) [27].

F. Honey and Mumford's Model (Flexibly Stable Learning Preferences Family)

Grounded in Kolb's theory, Honey and Mumford learning style model describes learning styles as behaviors and attitudes that determine individual learning preferences [28], [29]. They claim that people learn in the same way as experimental scientists conduct research and that learners' learning styles differ according to the phases of the learning process which they are best at. Despite the fact that their theory also relies on the steps of the experiment process, it doesn't assume the establishment of bipolar dimensions, as is the case with Kolb.

TABLE. VI. KOLB LEARNING STYLE MODEL

Dimension	Pole	Learning style
Grasping / Prehension	Concrete Experience (CE)	1. Diverging 2. (CE/RO) 3. Assimilating 4. (AC/RO) 5. Converging 6. (AC/AE) 7. Accommodating 8. (AC/AE)
	Abstract Conceptualization (AC)	
Transforming / Processing	Reflective Observation (RO)	
	Active Experimentation (AE)	

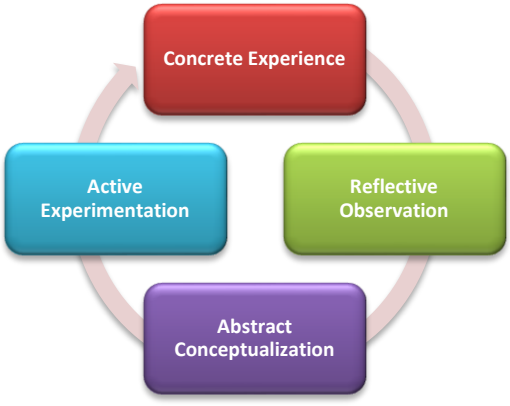


Fig. 4. The Experimental Learning Theory Learning Cycle

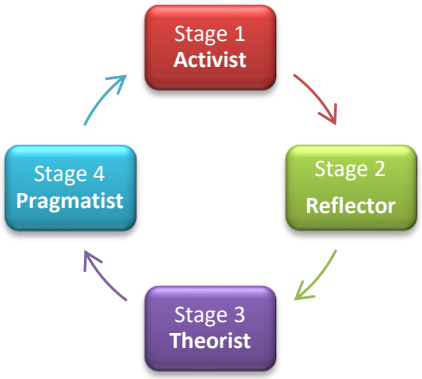


Fig. 5. H and M Learning Style Model.

The authors refer to the learning styles with the four stages (see Fig. 5): Activists (i.e. individuals who approve experiencing); Reflectors (i.e. individuals who rather reviewing experiences or pondering over facts); Theorists (i.e. individuals who favor reasoning); and pragmatists (i.e. individuals who prefer planning the next steps).

Honey and Mumford's intention is that learners should become proficient in all four stages of the learning cycle.

In order to assess an individual's learning style, Honey and Mumford Learning Style Questionnaire was developed. It consists of 80 items with true/false answers that probe preferences for four learning styles, with 20 items for each style.

G. *Entwistle's Approaches to Learning and Studying Inventory (Learning Approaches and Strategies Family)*

Noel Entwistle, a researcher in the field of educational psychology, and his colleagues developed a "teaching-learning process" experimental model in order to guide institutions to undertake a process of critical reflection on their adopted methods with the intention of reforming the whole learning environment to enhance the student learning quality (N. Entwistle 1990). This model intends to encompass the problematic influence structure that links motivation, academic performance and learning approaches with the indirect effects of teaching and assessment methods. Furthermore, it aims to identify the students' tendencies to adopt deep, surface and strategic approaches to learning and studying.

Entwistle defined a strategy as the manner a learner chooses to confront a specific learning assignment in accordance with its observed requirements and a style as an extensive description of a learner's favored approach of dealing with learning tasks generally [30]. Moreover, he distinguishes three separate learners' personality types in higher education courses in conformity with studies on the effects of personality on learning: non-committers, hustlers and plungers. While the first category of learners tends to be thoughtful, anxious and risk-averse, the second one is composed of competitive, dynamic but insensitive students, and the third gathers sensitive, thoughtless and self-reliant ones.

The Approaches to Learning and Studying Inventory (ALSI) have been developed to evaluate learners' approaches to learning and their perceptions about the course organization and the teaching impact (N. J. Entwistle, McCune, and Tait 1997). It was derived from evaluations of other measures, namely, the Approaches to Studying Inventory (ASI-1981); the Course Perception Questionnaire (CPQ-1981); the Revised Approaches to Studying Inventory (RASI-1995) (N. Entwistle and McCune 2004); and the Approaches and Study Skills Inventory for Students (ASSIST-1997).

The complete version of the ALSI questionnaire uses a Likert technique to determine students' attitudes by rating a series of related items that deal with the aspects of a specific construct (deep, surface and strategic) on a five-point scale. It incorporates three parts: The first one is concerned with students' perceptions of learning. The second relates to their study practices; The third involves students' preferences for different kinds of teaching (N. Entwistle and Tait 2013).

IV. Discussion

This section defines the comparative summary of the selected most influential learning style models. The following table (see Table VII) summarizes this discussion and shows the learning style's model name, family, instrument, strengths, and weaknesses.

The Gregorc's Mind Styles Model relies on individuals' instinctive abilities of 'perception' and 'ordering'. While Gregorc states that his model has high levels of internal consistency and test-retest reliability, no evidence for his theoretical claim is provided and significant uncertainties were expressed about its psychometric properties in the literature. Concerning validity, moderate correlations are reported for criterion-related validity, but there is no empirical evidence for construct validity. Some of the words used in the instrument are unclear or may be unfamiliar to the end-users which makes it irrelevant for the assessment of individuals.

Even if the Riding Cognitive Style Model is known for its simplicity, it lacks empirical evidence and suffers from unresolved conceptual problems and serious difficulties with its instrument. Within this model, only cognitive aspects of thinking and learning are dealt with without taking into account the affective and conative ones. Moreover, learning styles are assumed to be fixed, and metacognitive training which might lead to learning styles alteration is not considered. The author hasn't provided any evidence about its reliability, while other studies have revealed that internal consistency and test-retest reliability are very poor. And finally, its pedagogical impact is questionable.

The Myer-Briggs Type Indicator Model was specifically designed as a tool to classify an individual's personality type in general, and their approaches to relationships with others. Some researchers in the learning styles field choose to exclude the MBTI on the grounds that its scope as a personality measure outweighs cognitive regulations and behavior specifically related to learning. Furthermore, victim of its own massive commercial success of the MBTI as style measurement instrument, some of the critical and experiential examinations done with it are superficial and neglectful. That's why the research evidence to advocate it as an effective style evaluation and pedagogical support is still unconvincing. On the one hand, there has been considerable debate about the

construct validity of the MBTI and the irrelevant forced-choice format of the instrument. On the other hand, the stability of the MBTI types allocations is open to question in part because the middle scores are prone to misinterpretation due to small numerical differences. Finally, the practical application of MBTI in pedagogy is still ambiguous as there is no evident perception of how type dynamics impact on education.

The Felder-Silverman Learning Styles Model beneficiates from a considerable amount of available literature and has been frequently used by educators in different disciplines for providing adaptivity regarding learning styles in e-learning environments as it provides a detailed description of the different dimensions of the learner's style and taking into account inclinations on four measurements. Moreover, its instrument, the ILS, has undergone multiple studies according to the literature and has proved to be user-friendly and effective for instruction and assessment design in accordance with the learners' learning styles because the results are easy to interpret, and because the number of dimensions is controlled and can, in fact, be implemented. Even if there is no full consensus on studies results, they have shown satisfactory convergent and discriminant validity, scarce reliability, and satisfactory consistency.

The Kolb's Experiential Learning Theory is one of the first learning style models, based on an explicit theory, that engaged a tremendous international literature dedicated to its examination. However, Kolb disrecommends the use of its instrument, the LSI, for individual selection purposes as it cannot measure individuals with thorough precision. Moreover, the psychometric properties of the LSI have been the subject of criticism, and there is no unequivocal evidence in the literature that shows that it enhances academic performance. Although the LSI beneficiated from the critique to improve the reliability of the instrument, the test-retest reliability suggests that the LSI is rather volatile and the reliability coefficients for the four basic scales are not (enough) satisfactory.

While Honey and Mumford's Model can be used for personal development by drawing proposals to help individuals to fortify underexploited styles, studies prove it uselessness for individuals' selection on the basis of their learning styles as it exposes no enough distinctive scale scores to allow them to be categorized. Moreover, it labels individuals while most people show more than one strong style. And even if it has been extensively used in the professional field, it requires to be reformed to transcend shortcomings critics showed when evaluated by researchers and more proofs of its validity are essential in order to adopt it with confidence.

Entwistle' Approaches to Learning and Studying Inventory is an important aid for the discussion and diagnosis of the effective and ineffective strategies for learning in accordance with learners' actual approaches as a basis for redesigning instruction and assessment. However, the use of the instrument for adapting the pedagogic environment turned out to be difficult for non-specialists who lack an in-depth understanding of its underlying implications. The model has undergone extensive evaluation and the result showed satisfactory reliability and internal consistency but contested construct and predictive validity. Moreover, the external analysis confirmed the validity of the deep, surface and strategic approaches. All in all, this model needs to be redesigned, tested and revalidated for pedagogical interventions.

All the aforementioned things considered; the Felder-Silverman Learning style model reveals itself to be the most appropriate model for providing adaptivity and accurate instruction and assessment design in accordance with the learners' learning styles in e-learning environments.

Moreover, and according to the literature, Felder Silverman model turned out to be the most preferred model of learner style used in the learning theories and has been successfully implemented in many previous works when individually adapting the learning material. Author in [31] within the scope of adaptive education systems, inspected 69 studied published from 2005 to 2014. The results revealed that the Felder-Silverman learning style model was the most preferred model (42%), followed by the Kolb model (14.5 %). In another work, [32] examined integrating learning styles in adaptive e-learning systems by reviewing 51 studies published from (2004 to 2014). The results of this study show that the Felder-Silverman learning style model was the most preferred model (70.6%), and then the VARK model (9.8%). In a more recent study, [6] investigated the researches on learning styles used in e-learning environments published between the years 2001 to 2016. y When these studies were classified by considering the used learning style, it has been observed that Felder Silverman model was the most preferred model of learner style used in the learning theories (n= 33; 46.67%), followed by Kolb model (n=14; 19.71).

TABLE. VII. LEARNING STYLES MODELS' COMPARISON

Model	Family	Instrument	Instrument mechanism	Strengths	Weaknesses
Gregorc	Constitutionally based learning styles and preferences	Gregorc Style Delineator (GSD)	Rank a set of items	✓ Considerable internal consistency ✓ High test-retest reliability ✓ Moderate criterion-related validity	× Poor psychometric properties. × Static learning styles × Undemonstrated construct validity × Theoretical evidence of the pedagogical impact × Irrelevant for the individuals' assessment
Riding	Cognitive structure	Cognitive Styles Analysis (CSA) ***(1991)***	Select only one answer from two alternatives	✓ Simplicity ✓ Acceptable face validity	× Debatable conceptual issues × Weak internal consistency × Very low test-retest reliability × Questionable evidence of pedagogical impact × Unreliable instrument
MBTI	Stable personality types	Myers- Briggs Types Indicator (MBTI) ***(1962)***	Likert scale question	✓ Provides a view of the whole personality ✓ High-reliability coefficients ✓ Approved face validity	× Not learning specific × Complicated relationships between elements and scales × Weak stability of the learning styles × Contestable construct validity × No proof of any beneficial outcomes concerning the pedagogical impact.
FSLSM	Flexibly stable learning preferences	Index of Learning Styles (ILS) ***(1996)***	Select only one answer from two alternatives	✓ Learning specific ✓ Flexible and stable learning styles ✓ Detailed description of the learning style of a learner ✓ Widespread use ✓ Satisfactory convergent and discriminant validity ✓ Scarce reliability ✓ Convenient for instruction individualization	× Low predictive validity
Kolb	Flexibly stable learning preferences	Learning Style Inventory (LSI) ***(1976)*** Revised Inventory (R-LSI) ***(1985)*** Learning Style Inventory-v3 ***(1999)***	Rank a set of items	✓ Flexible and stable learning styles ✓ Reliable instrument ✓ Convenient for instruction individualization	× Unsuitable for individual selection × Deficient notion of a learning cycle × Doubtful psychometric properties × Controversial reliability × Disputed construct validity × Low predictive validity × Theoretically-based the pedagogical impact
H&M	Flexibly stable learning preferences	LS Questionnaire (LSQ) ***(1982)***	Mark a set of items	✓ Learning specific ✓ Helpful for individuals to fortify an under-used style ✓ Instrument translated into dozens of languages	× Individuals labeling × Useless for assessment/selection × Very criticized model design × Moderate internal consistency × Speculative validity × No empirical evidence of pedagogical impact
Entwistle	Learning approaches and strategies	Approaches to Studying Inventory ***(1981)*** Course Perception Questionnaire ***(1981)*** Revised Approaches to Studying Inventory ***(1995)*** Approaches and Study Skills Inventory for Students ***(1997)*** Approaches to Learning and Studying Inventory ***(2007)***	Rank a set of items	✓ Learning orientations assessment ✓ Course organization and instruction preferences evaluation ✓ Satisfactory reliability and internal consistency ✓ Confirmed validity of deep, surface and strategic approaches ✓ A basis for discussing the effective and ineffective strategies for learning ✓ A basis for redesigning instruction and assessment	× Complex model × Instrument with limited accessibility × Requires in-depth understanding × Test-retest reliability is not demonstrated. × Contested construct and predictive validity × Difficulties to transform the learning environment in accordance with the instrument results × No empirical evidence for the pedagogical impact

V. Conclusion

Through the review of learning style research, we found several models and assessment instruments that can be applied to university's education. Many of these have been adapted as online tests. The review shows that the Felder Silverman model is the most suitable for adaptive e-learning for the aforementioned and discussed reasons.

Future research and empirical studies will be done specifically to investigate the efficiency of these learning styles for optimal learning and teaching experience that leads to better learning outcomes. Moreover, all of the existing learning style instruments were built using only the textual form of information, which is considered more suitable for verbal learners than others. Consequently, and in order to increase the efficiency of the ILS instrument, we are thinking of constructing another form of the standard questionnaire that will be more convenient for a larger panel of users using the different forms of information. Furthermore, a future paper will be dedicated to how can we ideally detect learners' learning style in a hybrid manner by combining the use of the revised form of the questionnaire to initialize the model and automatic detection techniques to update it after each learning session.

References

[1] S. Graf, T. C. Liu, Kinshuk, N. S. Chen, and S. J. H. Yang, "Learning styles and cognitive traits - Their relationship and its benefits in web-based educational systems," Comput. Human Behav., vol. 25, no. 6, pp. 1280–1289, 2009.

[2] L. J. Ausburn, "The freedom versus focus dilemma in a customized self-directed learning environment: A comparison of the perceptions of adult and younger students," Community Coll. J. Res. Pract., vol. 26, no. 3, pp. 225–235, 2002.

[3] M. Šimko, M. Barla, and M. Bieliková, "ALEF: A framework for adaptive web-based learning 2.0," in IFIP International Conference on Key Competencies in the Knowledge Society, 2010, pp. 367–378.

[4] S. Cassidy, "Learning Styles: An overview of theories, models, and measures," Educ. Psychol., vol. 24, no. 4, pp. 419–444, Aug. 2004.

[5] R. M. Felder and L. K. Silverman, "Learning and teaching styles in engineering education," Eng. Educ., 1988.

[6] N. Kumar, Amit; Ahuja, Neelu Jyothi; Singh, "Learner characteristics based learning style models classification and its implications on teaching," Int. J. Pure Appl. Math., vol. 118, no. 20, pp. 175–184, 2018.

[7] L. Mei Ph'ng, "Teaching Styles, Learning Styles and the ESP Classroom," MATEC Web Conf., vol. 150, p. 05082, Feb. 2018.

[8] W. B. James and D. L. Gardner, "Learning styles: Implications for distance learning," New Dir. Adult Contin. Educ., vol. 1995, no. 67, pp. 19–31, 1995.

[9] F. Coffield, D. Moseley, E. Hall, and K. Ecclestone, Learning styles and pedagogy in post-16 learning A systematic and critical review. 2004.

[10] L. Curry, "An organization of learning styles theory and constructs.," Educ. Res. Inf. Cent., pp. 2–28, 1983.

[11] L. Curry, "Review of learning style, studying approach, and instructional preference research in medical education," Int. Perspect. Individ. Differ., vol. 1, pp. 239–276, 2000.

[12] M. A. Bokoros, M. B. Goldstein, and M. M. Sweeney, "Common factors in five measures of cognitive style," Curr. Psychol., vol. 11, no. 2, pp. 99–109, 1992.

[13] J. Chevrier, G. Fortin, M. Théberge, and R. Le Blanc, "Le style d'apprentissage: une perspective historique," Le style d'apprentissage, vol. 28, no. 1, 2000.

[14] L. Curry, "Patterns of Learning Style Across Selected Medical Specialties," Educ. Psychol., vol. 11, no. 3–4, pp. 247–277, 1991.

[15] N. Entwistle, V. McCune, and P. Walker, "Conceptions, styles, and approaches within higher education: Analytic abstractions and everyday experience," Perspect. thinking, Learn. Cogn. styles, pp. 103–136, 2001.

[16] R. Riding and I. Cheema, "Cognitive styles—an overview and integration," Educ. Psychol., vol. 11, no. 3–4, pp. 193–215, 1991.

[17] R. Riding and S. Rayner, Cognitive styles and learning strategies: Understanding style differences in learning and behavior. David Fulton Publishers, 2013.

[18] H. A. Witkin, R. B. Dyk, H. F. Fattuson, D. R. Goodenough, and S. A. Karp, "Psychological differentiation: Studies of development.," 1962.

[19] A. Paivio, Imagery and verbal processes. Psychology Press, 2013.

[20] E. Sadler-Smith and R. Riding, "Cognitive style and instructional preferences," Instr. Sci., vol. 27, no. 5, pp. 355–371, 1999.

[21] I. B. Myers, M. H. McCaulley, and R. Most, Manual, a guide to the development and use of the Myers-Briggs type indicator. consulting psychologists press, 1985.

[22] S. A. Girelli and J. E. Stake, "Bipolarity in Jungian Type Theory and the Myers--Briggs Type Indicator," J. Pers. Assess., vol. 60, no. 2, pp. 290–301, 1993.

[23] R. M. Capraro and M. M. Capraro, "Myers-briggs type indicator score reliability across: Studies a meta-analytic reliability generalization study," Educ. Psychol. Meas., vol. 62, no. 4, pp. 590–602, 2002.

[24] B. A Soloman and R. Felder, "Index of Learning Styles Questionnaire," 1999. [Online]. Available: https://www.webtools.ncsu.edu/learning styles/. [Accessed: 10-May-2019].

[25] S. El Lakkah, M. A. Alimam, and H. Seghiouer, "Adaptive e-learning system based on learning style and ant colony optimization," in 2017 Intelligent Systems and Computer Vision (ISCV), 2017, pp. 1–5.

[26] A. Y. Kolb and D. A. Kolb, "Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education," Acad. Manag. Learn. Educ., vol. 4, no. 2, pp. 193–212, Jun. 2005.

[27] Kolb and D. a Kolb, "The Kolb Learning Style Inventory 4.0," Exp. Based Learn. Syst. Inc., no. January 2005, pp. 1–72, 2013.

[28] K. Knight, "Book Reviews : The Manual of Learning Styles Peter Honey and Alan Mumford," Manag. Educ. Dev., vol. 14, no. 2, pp. 147–150, Jul. 1983.

[29] P. Honey and A. Mumford, The learning styles helper's guide. Peter Honey Publications Maidenhead, 2000.

[30] N. Entwistle, M. Hanley, and D. Hounsell, "Identifying distinctive approaches to studying," High. Educ., vol. 8, no. 4, pp. 365–380, 1979.

[31] Ö. Özyurt and H. Özyurt, "Learning style based individualized adaptive e-learning environments: Content analysis of the articles published from 2005 to 2014," Comput. Human Behav., vol. 52, pp. 349–358, Nov. 2015.

[32] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," Comput. Human Behav., vol. 55, pp. 1185–1193, 2016.

Performance Analysis of Double Gate Junctionless Tunnel Field Effect Transistor: RF Stability Perspective

Veerati Raju[1], Sivasankaran K[*2]
School of Electronics Engineering
VIT University, Vellore-632014, India

***Abstract*—This paper investigates the RF Stability performance of the Double Gate Junctionless Tunnel Field Effect Transistor (DGJL-TFET). The impact of the geometrical parameter, material and bias conditions on the key figure of merit (FoM) like Transconductance (g_m), Gate capacitance (C_{gg}) and RF parameters like Stern Stability Factor (K), Critical Frequency (f_k) are investigated. The analytical model provides the relation between f_k and small signal parameters which provide guidelines for optimizing the device parameter. The results show improvement in ON current, g_m, f_t and f_k for the optimized device structure. The optimized device parameters provide guidelines to operate DGJL-TFET for RF applications.**

***Keywords*—Junctionless tunnel FET; band to band tunnelling; High-k; RF stability; critical frequency**

I. INTRODUCTION

For the past four decades, the semiconductor industry is supplemented with CMOS devices due to the continuous growth of Semiconductor Technology. During this regime, the silicon device physical dimensions were reduced to nanometre domain and further scaling (Tens of Nanometre) is limited by Short Channel Effects (SCE) posed by CMOS devices [1-6]. To overcome such challenges, Multi-gate devices are proposed, which shown excellent immunity to SCE and yielded better scalable operations [7-9]. Beyond 30nm, these Multi-gate devices also suffer SCE's and to overcome them, Tunnel Field Effect Transistor (TFET) is proposed which has gained wider significance because of its low subthreshold slope and small leakage current [1-2, 10-14]. Moreover, the tunnel FET device suffers from low ON current and requires abrupt junctions for tunnelling [15]. To overcome the fabrication challenges posed by the MOS and TFET devices, a new transistor called Junctionless transistor (JL) with no doping gradients is proposed to achieve good ON and OFF states [16-17]. Even though the JL device has better scalable performance than the MOSFETs, still it suffers from the low subthreshold slope. To counter the above challenges, Junctionless Tunnel Field Effect Transistor (JL-TFET) is proposed, which exhibits better subthreshold slope of 24mV/decade and DIBL of 38mV/V as compared to conventional JLFET [18-19]. Further, most of the research is carried on investigating the analog performance metrics like Transconductance (gm), unity gain cut off frequency (fT), output conductance (gd) and Intrinsic gain (gm/gd) for the n-type and P-type Double Gate junction and Junctionless TFET [20-21]. Impact of geometrical variability's on the performance of JLTFET and Junctionless hetero structure TFETs (HJLTEFT) is investigated and proposed pocket oxide narrower source side HJTFETs (PNS-HJTFETs) for better performance [22]. The effect of the gate dual material (DMG) and gate engineering approach on the performance of DGJL-TFET is studied [23-25]. Comprehensive analysis on the 20nm HJLTFETs with high-k gate oxide material is presented [26]. The authors have investigated the influence of spacer on ION/IOFF ratio and gm of a DG JLTFET [27]. Impact of dual k spacer on the digital and analogue performances of JL TFET, formed with different substrates is analyzed [28-29]. Influence of parameter fluctuations caused by process variations on the RF stability of Double Gate Tunnel FET (DG-TFET) is reported by k.sivasankaran.et.al [30]. Influence of high-k material on the RF stability performance of Double Gate Junctionless FET is studied [31] and proposed an optimized structure for the better RF performance [32]. The impact of the high-k gate dielectric and dual spacer on the RF stability Performance of JLTFETs is not been studied before. Most of the studies [18-27] focused on analyzing the behaviour of JL-TFETs with and without high-k materials for improving the DC and analogue performance but not on the stability aspects of the device.

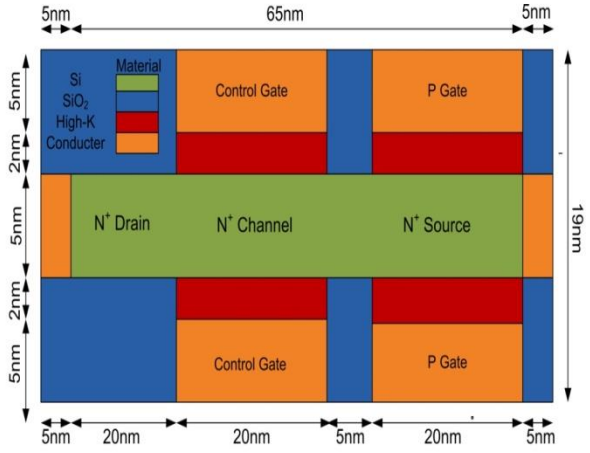


Fig. 1. Schematic Representation of DGJL-TFET Considered for Simulation.

*Corresponding Author

In this work, we have evaluated the RF Stability performance of the DGJL-TFET for different gate oxide materials and different isolation spacers. Further, the effect of geometrical variations towards the critical frequency is also analyzed for proposing the optimized device.

The remainder of the paper is organized as follows: Section 2 describes the DGJL-TFET considered for simulation. Section 3 gives the device Calibration and DC characteristics. RF Stability performance of the DGJL-TFET for different gate oxide materials is presented in Section 4. The results are discussed in Section 5. The optimized device structure is proposed in Section 6. Finally, the conclusion of the work is presented in Section 7.

II. Device Structure and Simulation

Fig. 1 shows the schematic of the DGJL-TFET. The device is heavily n-type doped 20nm long Si-channel with Source/Drain extension of 20nm [18]. The gate oxide thickness (t_{ox}) is 2nm and device doping profile is maintained at $1x10^{19}$ cm^{-3} for silicon body thickness of 5nm. The device is operated with two gates with different work functions: one gate called control-gate (CG) which is used to control the charge flow in the channel (ON and OFF of the device) by sweeping a control-gate voltage (V_{CG}) from 0V to V_{DD} and another one is P-gate (PG), used to convert the N^+ source of DGJL-TFET to P-type by using gate work function engineering for tunnelling operation. The spacer width, which isolates the CG and PG of the device, is 5nm. All the simulations are performed using a 5.15.32.R version of Silvaco Atlas [33]. A non-local band to band (BTBT) tunnelling model is used to estimate the TFET device performance [18] by considering the tunnelling along the lateral direction between source and drain. Due to heavy doping of the channel, the band gap narrowing (BGN) model is considered and because of the high impurity atom present in the channel, the Shockley-Read-Hall (SRH) recombination model is enabled. Both quantum confinement effect, as well as interface trap effects in TFETs on the non-local band to band tunnelling, are considered by including quantum confinement (QC) model developed by Hansch [33][34] and Schenk [33][35] trap-assisted tunnelling (TAT) model. The work function of Control Gate is taken as 4.3eV for switching the layer under it as intrinsic and platinum metal [36] with a work function of 5.93eV is considered for P-Gate to make the layer under it as P-type region.

III. Device Calibration

A. DC Characteristics

Fig. 2(a) shows the OFF-state charge carrier concentration profile of DGJL-TFET (Drain Source Voltage, V_{DS}=1V and Control Gate Source Voltage, V_{CGS}=0). From Fig. 2(a) it is evident that the device is behaving like N^+-I-P^+ device. Energy band profile of DGJL-TFET in OFF state is presented in Fig. 2(b). The energy gap between the valence band and conduction band is large, due to which tunneling probability of charge carriers through the tunneling region is negligible. Hence, the current flowing in the off sate is small and is only due to the leakage current flowing in N^+-I-P^+ diode.

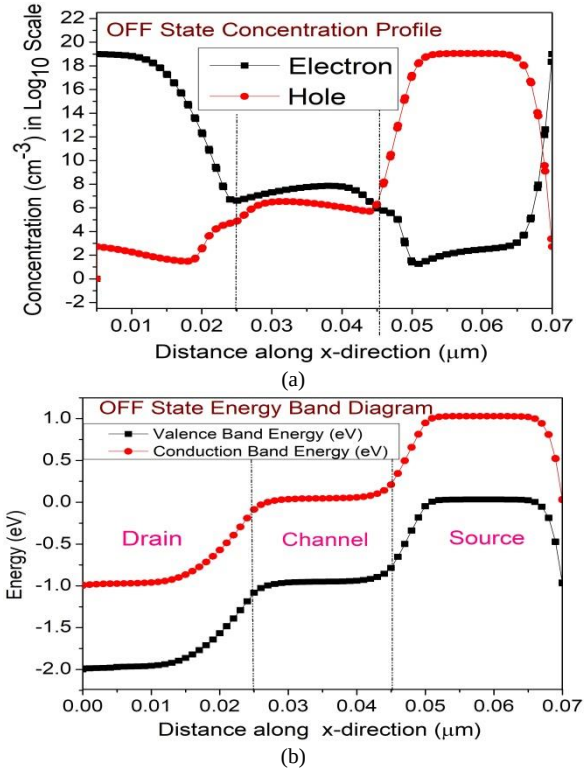


Fig. 2. (a) Electron and Hole Concentration of DGJL-TFET in OFF-State. (b) OFF State Energy Band Profile of DGJL-TFET.

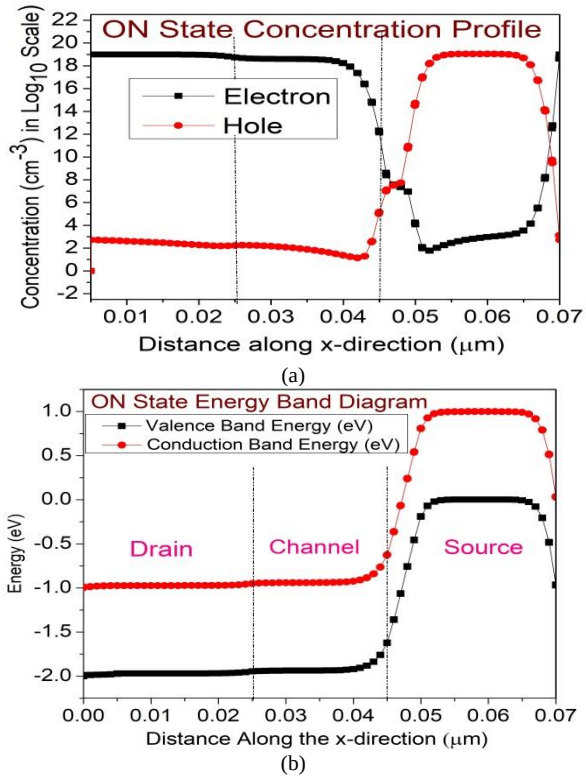


Fig. 3. (a) Electron and Hole Concentration of DGJL-TFET in ON-State (b) ON State Energy Band Profile of DGJL-TFET.

Fig. 3(a) shows the ON state (V_{DS} = 1V, V_{CGS} = 1V) charge concentration profile of DGJL-TFET. By looking at Fig. 3(a), it is evident that when a voltage of 1V applied between CG and source, the layer beneath the CG is converted to N^+, due to which electron concentration in the channel is increased. The rise in electron concentration is due to the increase of tunneling probability of the charge carriers moving from the source to channel. The energy band profile of DGJL-TFET in ON-state is shown in Fig. 3(b), it clearly shows that the narrow bandgap present between the channel and source, due to which tunneling width (λ) is small, subsequently the tunneling probability of the electrons flowing from source to channel is increased and thereby increasing the device ON current.

It is reported that for TFET devices, ON current can greatly improve with the use of high-k materials as gate dielectric [24]. So high k materials like TiO_2 (ε_r=80), HfO_2 (ε_r=25) and Al_2O_3 (ε_r=9) as gate dielectrics [18] are considered for studying their impact on stability of the device with a fixed physical thickness of 2nm.

B. Transfer Characteristics

The transfer characteristics of DGJL-TFET (depicted in Fig. 1) for various gate dielectric materials are shown in Fig. 4. It is observed that the gate oxide with high-k material gives higher ON current and improved subthreshold swing. The improvement in ON current with high-k gate dielectrics mainly due to the reduction of tunneling width (λ), in return, increases the nonlocal tunneling probability according to the WKB approximation, given by equation 1 and also due to the increase of gate coupling with the channel. From Fig. 4, it is observed that the ON current of 34.5μA/μm for a TiO_2 gate dielectric and 0.295μA/μm for Al_2O_3 gate dielectric, with both V_{CGS} and V_{DS} are at 1V.

$$T_t = \exp\left(-\frac{4}{3}\frac{\lambda\sqrt{2m^*}}{3qh(E_g+\nabla\phi)}(E_g)^{3/2}\right) \quad (1)$$

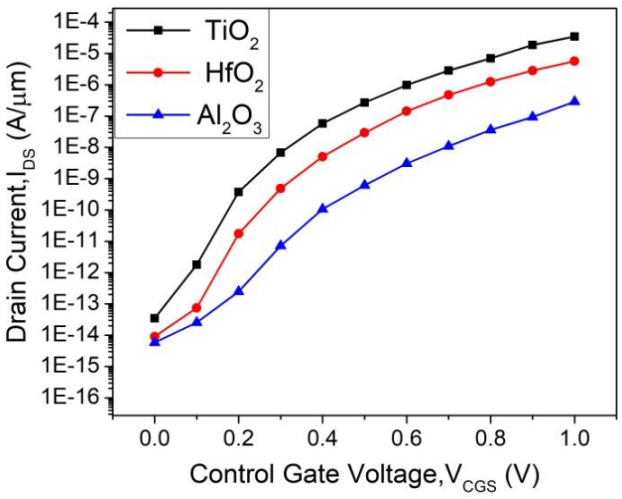


Fig. 4. Transfer Characteristics of DGJL-TFET for Various Gate Dielectric Material Ranging from k= 9 to k=80 with V_{CGS} Sweeping from 0 to 1V and V_{DS} at 1V.

IV. RF Stability Performance of DGJL-TFET

An important phase in the overall analysis of transistor is to identify its potential stability. This may be achieved by calculating the Stern stability factor (K), which specifies the device is unconditionally stable or conditionally stable [37-39]. If K value is greater than one, then the transistor is unconditionally stable, otherwise, it is conditionally stable and which leads to oscillations at some frequency. Equation 2 given by P.Stern in terms of Y parameters is used to calculate the K value of a transistor.

$$K = \frac{2\,\mathrm{Re}(Y_{11})\,\mathrm{Re}(Y_{22}) - \mathrm{Re}(Y_{12}Y_{21})}{|Y_{12}Y_{21}|} \quad (2)$$

Where Y_{11} and Y_{22} are admittances, Y_{12} and Y_{21} are known as transfer admittance of a two port network.

$$Y_{11} = \omega^2\left(R_{gs}C_{gs}^2 + R_{gd}C_{gd}^2\right) + j\omega(C_{gs}+C_{gd}) \quad (3)$$

$$Y_{12} = -\omega^2 R_{gd}C_{gd}^2 - j\omega C_{gd} \quad (4)$$

$$Y_{21} = g_{mi} - \omega^2 R_{gd}C_{gd}^2 - j\omega C_{gd} + \tau g_{mi} \quad (5)$$

$$Y_{22} = -g_{dsi} + j\omega(C_{gs}+C_{gd}) + \omega^2 R_{gd}C_{gd}^2 \quad (6)$$

Substituting equations 3 to 6 in equation 2 will further simplify the K, which is given by equation 7.

$$K \cong \frac{\omega(R_{gs}g_{ds}C_{gg}^2 + 2R_{g_d}g_mC_{gg}C_{gd} + C_{gg}^2)}{C_{gd}\sqrt{2\omega^2 g_m C_{gg}^2 + g_m^2}} \quad (7)$$

Critical frequency (f_k) is one of the key RF performance parameter, specifies at what frequency device attains unconditionally stability. By substituting the K=1 in Equation 7 along with some approximations is simplified to equation 8 which is used to calculate the f_k of the device in terms of various parameters [40].

$$f_k \cong \frac{f_T N}{\sqrt{g_{ds}g_m R_{gs}M^2 + NM(g_m R_{gd}+1)}} \quad (8)$$

where, f_T is the Unity-gain cut-off frequency, which is one of important metric of analog characteristics of the device, specifies the frequency at which the current gain reaches to unity and is expressed as

$$f_T = \frac{g_m}{2\pi C_{gg}} \quad (9)$$

$$M = \frac{C_{gs}}{C_{gg}} \quad (10)$$

$$N = \frac{C_{gd}}{C_{gg}} \quad (11)$$

$$Cgs = Cgsi + Cfext + Cfin \quad (12)$$

$$Cgd = Cgdi + Cfext + Cfint \quad (13)$$

The other parameters mentioned in equation 8 are output-conductance (g_{ds}), gate-source resistance (R_{gs}), gate-drain resistance (R_{gd}), gate-capacitance (C_{gg}), gate-source capacitance (C_{gs}) and gate-drain capacitance (C_{gd}) without taking into account of overlap capacitance [39]. The internal fringing field (C_{fint}) as well as external fringing field (C_{fext}) is expressed as:

$$C_{f\,int} = [\frac{W\varepsilon_{si}}{3\Pi}\ln(1+\frac{t_{si}}{2t_{ox}}\sin(\frac{\Pi}{2}\frac{\varepsilon_{ox}}{\varepsilon_{si}}))]e^{-(\frac{(V_{gs}-V_{FB}-2\phi_f-V_{ds})}{(3/2)\phi_f})^2} \quad (14)$$

$$C_{fext} = [\frac{2W\varepsilon_{ox}}{3\Pi}\ln(1+\frac{t_g}{2t_{ox}})] \quad (15)$$

ε_{ox} and ε_{si} are dielectric constants of gate oxide and silicon materials. Where W, t_{si}, t_{ox} and t_g are width, silicon thickness, oxide thickness and gate material thickness, respectively. Φ_f and V_{FB} are Fermi potential and flat band voltage, respectively.

The stability factor is evaluated using equation 7 for DGJL-TFET with various gate oxide materials, whose dielectric values range from 9 to 80 is shown in Fig. 5. From the results, we can observe that the DGJL-TFET with a low-k gate oxide material (k=9) is attaining unconditional stable at a lower frequency at 6 GHz than the high-k gate dielectric (k=80) at a frequency of 130 GHz. Lower f_k value is exhibited by low-k gate dielectrics is mainly due to low ON current, low gate capacitance and smaller gain associated with it.

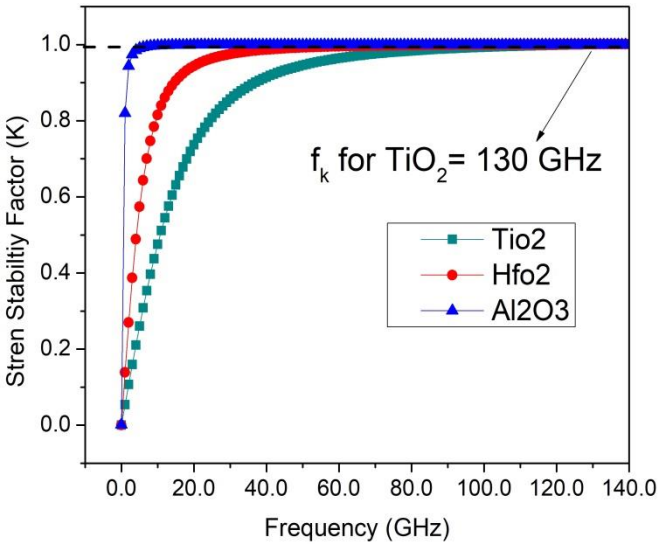


Fig. 5. Stern Stability Factor (K) of the DGJL-TFET Device for Various Gate Oxide Materials.

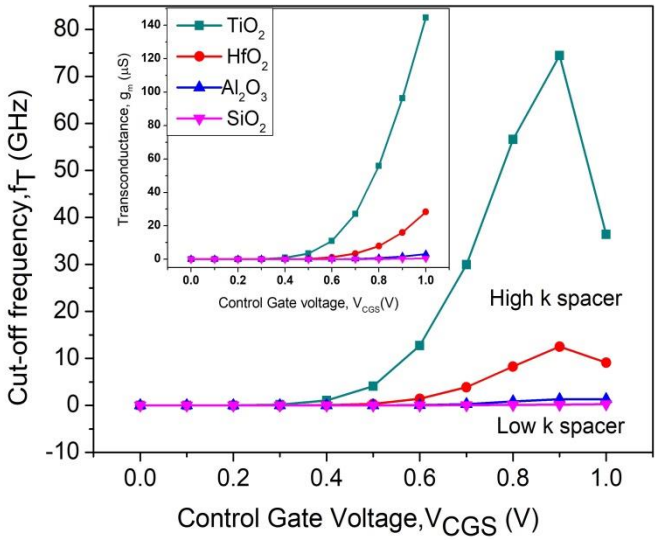
V. Results and Discussion

A. Impact of Gate Dielectric Material (k) on Critical Frequency (f_k)

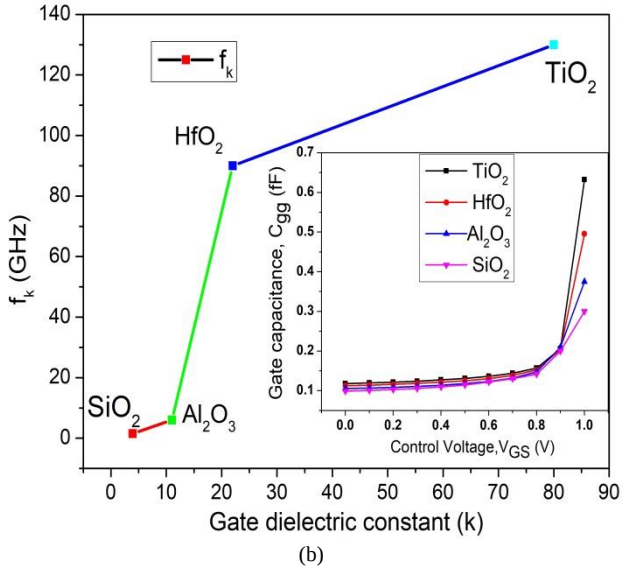
Fig. 6(a) shows the gate dielectric material influence on the cut off frequency (f_T) and transconductance (g_m).

Fig. 6(a) inset, it is observed that g_m is increasing with increase of dielectric constant of gate oxide material. As with the increase of gate dielectric constant, the tunneling width reduces due to which the tunneling probability increases, thereby enhancing the I_{ON} and thereby g_m. Because of the increase of relative permittivity of the gate oxide material with the rise of the gate dielectric constant, the C_{gg} also increases. Due to higher impact of g_m over the C_{gg}, f_T which is calculated from simulation values g_m and C_{gg} using equation 9 is increasing with increase in gate dielectric value.

As discussed earlier, from Fig. 6(a), due to improved performance of f_T, C_{gg} and g_m with a high-k gate dielectric, DGJL-TFET with high-k gate material (HfO_2, TiO_2) yielded higher f_k as compared to low-k gate material (SiO_2). From Fig. 6(b) f_k which is calculated using equation 8 is 1.5GHz for low-k gate oxide material (SiO_2) and for 130GHz for high-k gate oxide material (TiO_2). Various parameters which has effect on the stability and f_k for various oxide materials are presented in Table I.



(a)



(b)

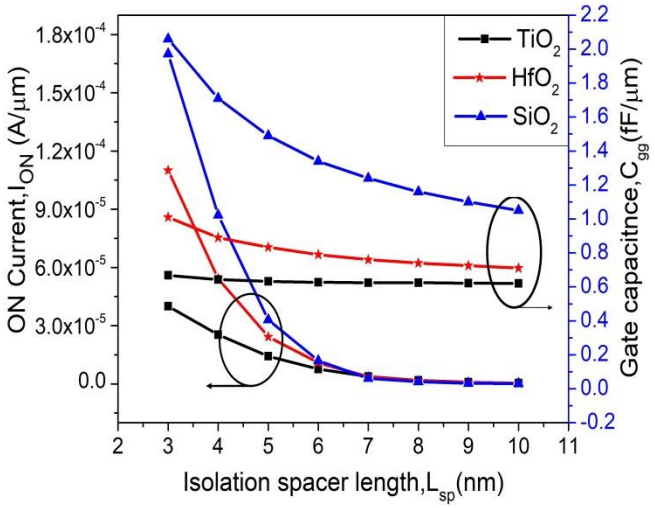
Fig. 6. Variation of (a) f_T and g_m (Inset) with Control Gate Voltage (V_{CGS}) and (b) f_k as Function of Gate Dielectric Constant (k).

TABLE. I. KEY PERFORMANCE METRICS OF DGJL-TFET FOR DIFFERENT GATE OXIDES AT VCGS= VDS=1V

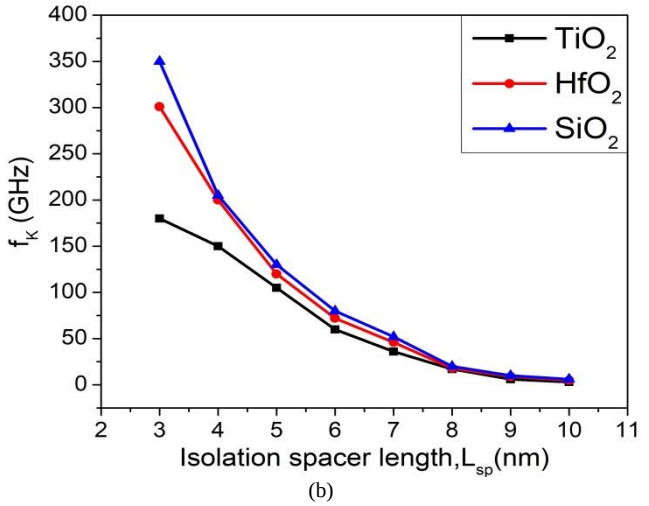
Gate oxide material	I_{on} (µA/µm)	g_m (µS)	C_{gg} (fF)	f_T (GHz)	f_k (GHz)
SiO_2	0.1	0.5	0.269	0.2	1.5
Al_2O_3	0.5	3.09	0.375	1.31	6
HfO_2	5.71	28.4	0.495	12.5	90
TiO_2	34.9	145	0.633	74.7	130

B. Impact of Isolation Spacer Length (L_{sp}) on f_k

In the previous subsection, it is observed that the ON current of the DGJL-TFET is improved with the aid of high-k gate dielectrics. The other way to increase I_{ON} is by scaling the isolation spacer length (L_{sp}), with the aid of charge plasma concept [24]. With scaling of spacer length (L_{sp}), the gate controllability over the source-channel region is increased, which reduces the tunneling barrier height, due to which the tunneling probability increases, subsequently increasing the I_{ON}. However, this improved I_{ON} is obtained at the expense of increased C_{gg}, which is shown in Fig. 7(a).



(a)



(b)

Fig. 7. Variation of (a) I_{ON} and C_{gg} and (b) f_k as Function oF isolation Spacer Length of DGJL-TFET.

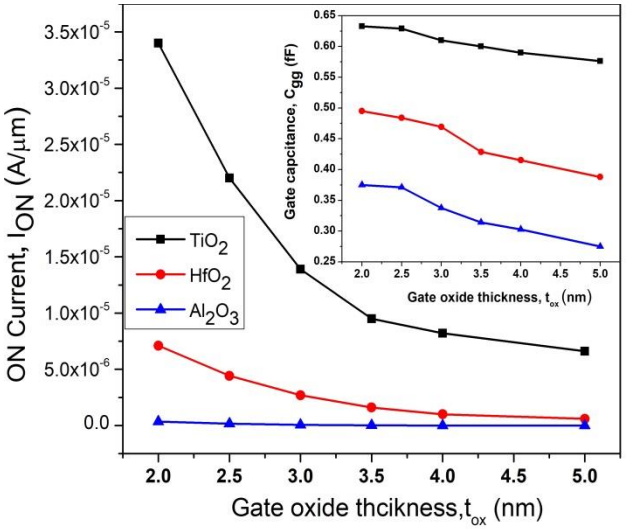
Fig. 7(b) shows the f_k variation with the scaling of isolation spacer (L_{sp}) of DGJL-TFET for different spacer material with 2nm TiO_2 as a gate dielectric. It is evident from Fig. 7(b) that, f_k is increasing with the scaling of the L_{sp}. While scaling the L_{sp} from 10nm to 5nm, the impact of it on f_k is not as much of for all the spacer materials, but further scaling of L_{sp} below 5nm the impact is large. This impact is largely attributed by interface defects at high-k oxide and Si interface, the induced trapped charges and capacitive fringing field's associated with high-k spacers. For TiO_2 spacer, f_k is varied from 3GHz to180GHz when L_{sp} scaled from 10nm to 3nm. However, a large variation of f_k nearly twice as that of the high-k spacer (TiO_2) is noticed with the low-k spacer (SiO_2). i.e., f_k is varied from 6 GHz to 350 GHz, when L_{sp} scaled from 10nm to 3nm.

C. Impact of Gate Oxide Thickness (t_{ox}) on f_k

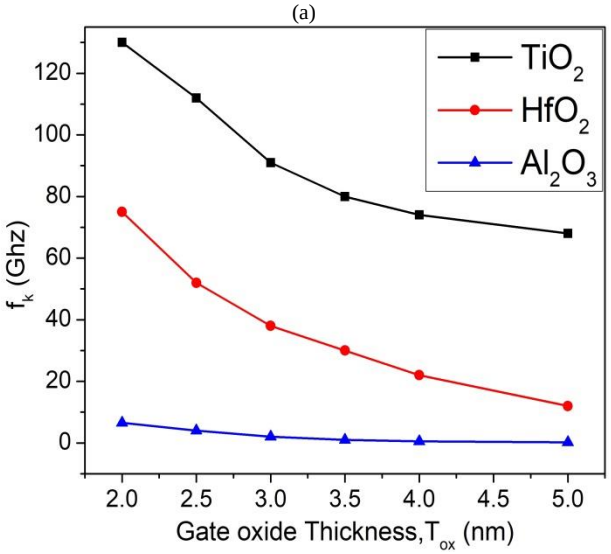
Tunneling process in T-FET devices is greatly affected by the variation gate oxide thickness (t_{ox}), this effect is noticed due to the variation of gate-capacitive coupling with variation of t_{ox}. As per WKB approximation, oxide thickness affects the tunneling probability by modulating the tunneling width. Equation 16 gives the dependency of tunneling width in terms of gate oxide thickness and other parameters. A thinner gate oxide will have less tunneling width (λ) and vice versa for thicker gate oxides.

$$\lambda = \sqrt{\left(\frac{t_{ox} t_{si} \varepsilon_{si}}{\varepsilon_{ox}}\right)} \quad (16)$$

The ON current of device with t_{ox}=5nm is smaller when compared to 2nm t_{ox} device, since 5nm oxide thickness devices have low capacitive coupling, due to which it has a lesser influence on the tunneling phenomena. Through simulation, it is also observed that irrespective of gate oxide material, DGJL-TFET exhibits improved I_{ON} with the scaling of the $t_{ox.}$ It is also observed that TiO_2 gate dielectric with t_{ox} =5nm has better I_{ON} (6.6 µA) than low-k gate dielectric (Al_2O_3) with t_{ox}= 2nm (0.35 µA) The simulated values effecting the stability and f_k with t_{ox} scaling for TiO_2 gate oxide material are given in Table II. Fig. 8(b) shows f_k variation with the scaling of t_{ox} for different gate oxide materials of a DGJL-TFET with 5nm SiO_2 isolation spacer. It is observed that, with the scaling of t_{ox}, DGJL-TFET with low-k gate dielectric has lower f_k then that of high-k gate dielectric. The simulation results illustrate that f_k for TiO_2 is varied from 68 GHz to 130 GHz and 0.2GHz to 6GHz for the Al_2O_3 with the scaling of t_{ox}. When t_{ox} scaled from 5nm to 2nm, nearly 50% increase in f_k for TiO_2 material is noticed and it is because of the large variations in I_{ON}, g_m, and C_{gg} with the scaling of t_{ox}. Fig. 9(a) illustrates a comparative impact of drain side spacer materials on the stability of the DGJL-TFET with 2nm TiO_2 as gate oxide and 5nm SiO_2 as isolation spacer and at an L_{dsp} of 15nm. It is observed that, drain spacer material has smaller impact on stability factor and almost same for all the spacers. As the drain spacer does not have tendency to affect the tunneling width, the current of the device and other stability parameters are virtually the same.



(a)



(b)

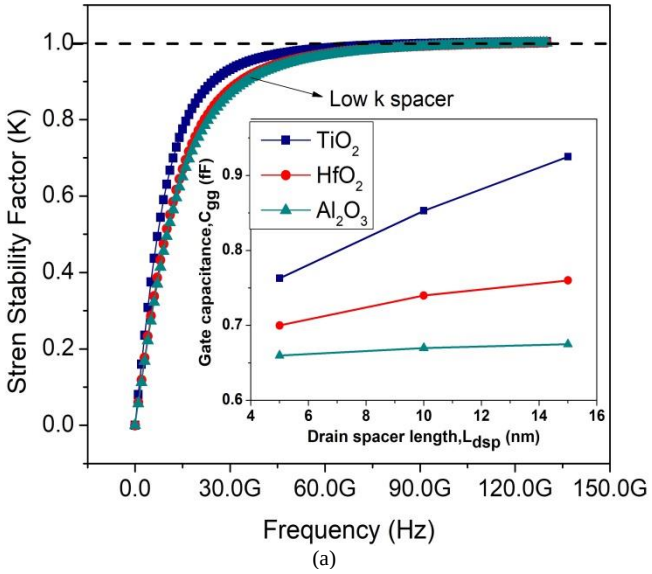
Fig. 8. Variation of (a) ON Current and C_{gg} and (b) Critical Frequency (f_k) with Respect to Scaling of Gate Oxide Thickness (t_{ox}) for different Gate Oxide Material of DGJL-TFET with 5nm SiO_2 as Isolation Spacer.

TABLE. II. KEY PERFORMANCE METRIC VALUES WITH THE SCALING OF TiO_2 GATE OXIDE OF DGJL-TFET

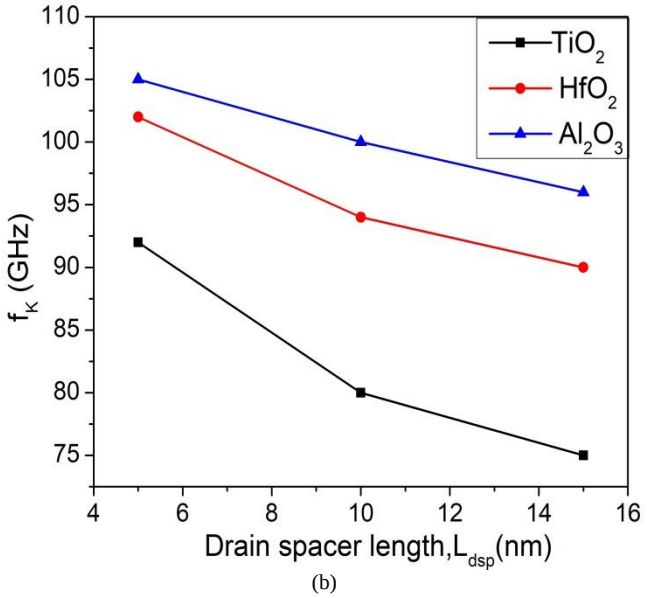
Gate oxide material=TiO_2	I_{on} (μA/μm)	g_m (μS)	C_{gg} (fF)	f_t (GHz)	f_k (GHz)
2 nm	34.9	145.0	0.63	74.4	130
2.5 nm	19.2	84.5	0.62	41.1	112
3 nm	14.0	64.4	0.61	28.9	91
3.5 nm	9.54	47.3	0.60	15.40	80
4 nm	8.27	38.1	0.59	11.8	74
5 nm	7.45	35.1	0.57	9.52	68

D. Impact of Drain Side Spacer and Drain Spacer Length (L_{dsp}) on f_k

From the Fig. 9(b), as L_{dsp} is scaled from 15nm to 5nm, f_k is increasing marginally for all the spacer materials and it is mainly because of the reduction spacer fringing fields and the capacitive area, thereby reducing the gate capacitance which is shown in the inset of Fig. 9(a). With the L_{dsp} scaling, C_{gg} is reduced largely for high k spacer and slightly varied for low k spacer. The parameters related to equation 3, like I_{ON}, g_m, f_t, are almost constant with spacer material and spacer length but the variation in C_{gg}, which is large for high k spacer is only impacting the f_k. Therefore f_k for TiO_2 spacer varied nearly 20 GHz, when L_{dsp} is scaled from 15 nm to 5nm.



(a)

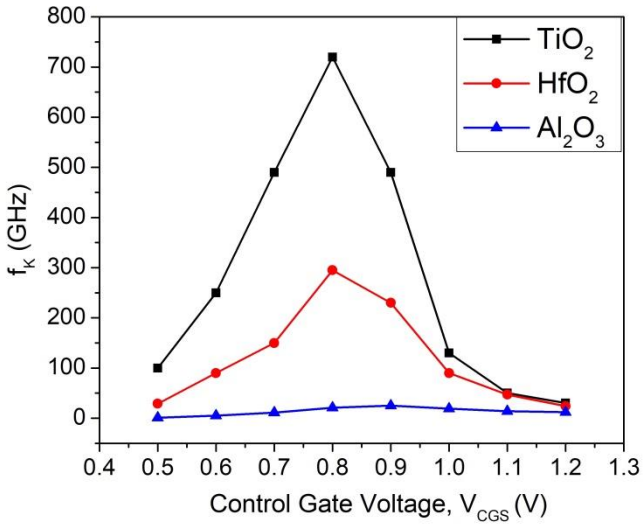


(b)

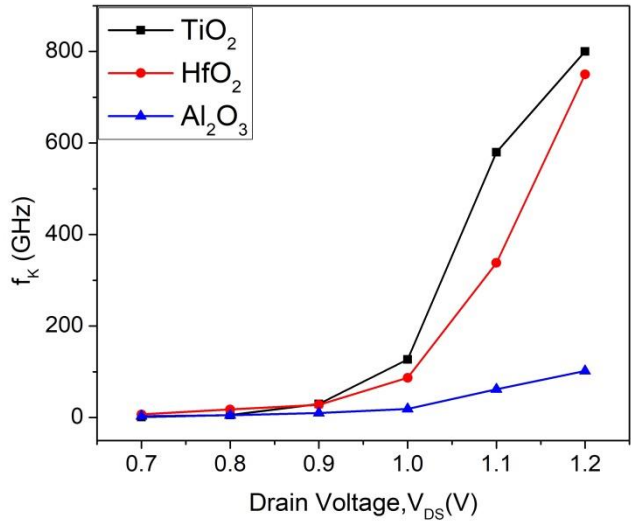
Fig. 9. (a) Stability Factor (K) Variation with Frequency for L_{dsp} 15nm and (b). f_k Variation with Scaling of L_{dsp} of a DGJL-TFET for different Spacer Materials.

E. Impact of Gate and Drain Bias on f_k

Fig. 10(a) shows the impact of control-gate voltage (V_{CGS}) on the f_k for 5nm isolation spacer DGJL-TFET for different gate oxides at V_{DS} of 1V. As gate bias V_{CGS} varied from 0.4V to 0.8V, f_k is increased for all gate oxides. since the device current, g_m and C_{gg} asscoiated with high- k gate oxide materials is large. Therfore the f_k of TiO_2 material is increased at a much higher rate than that of Al_2O_3 material. At higher gate bias, i.e., when V_{CGS} is increased beyond0.8V, f_k for Al_2O_3 lies in the same order but for HfO_2 and TiO_2 material it decreases to a larger extent. The decreasing in f_k for TiO_2 material is due to the large varition in C_{gg} and smaller variations in I_{ON} and g_m with control gate bias beyond 0.8V. The lowest value of f_k for TiO_2 material is 25 GHz at gate voltage of 1.2V. Finally, it is noticed that higher the control-gate voltage, lower the f_k value.



(a)



(b)

Fig. 10. Variation of f_k with (a) Control Gate Voltage and (b) Drain Voltage.

Fig. 10(b) shows f_k variation with respect to drain bias scaling for different gate oxide materials of DGJL-TFET with 5nm as isolation spacer. As drain voltage V_{DS} changing from 0.7V to 0.9V, f_k is of the same order and the insignificant difference is noticed for all the gate oxide. But for V_{DS} above 0.9V, the f_k of high-k gate dielectric (TiO_2) is much higher when compared to low-k gate dielectric (Al_2O_3). This is attributed by the decrease of output conductance (g_{ds}) and an increase of I_{ON} with the increase of drain voltage for TiO_2 gate oxide material. So, higher drain bias and high-k gate oxide material will yield higher f_k and vice versa.

VI. RF Performance of Optimized DGJL-TFET

The device geometry parameters like gate oxide thickness, spacer material, spacer length, gate dielectric material, and supply voltages are identified from the preceding sections for the optimization of the device for better RF stability performance is shown in Fig. 11. From the preceding section results, TiO_2 material is taken as gate dielectric material for the optimized DGJL-TFET structure with t_{ox} of 2nm. TiO_2 material is chosen as drain spacer and isolation spacer of a length 15nm and 3nm respectively. The structure is simulated with supply voltages at V_{CGS}=1.2V and V_{DS}=0.8V.

Fig. 12 shows the stern stability factor (K) for the optimized structure. The device attaining stability at a lower frequency and yielding lower f_k of 17.5GHz without degrading the I_{ON}. The low value of f_k is due to larger capacitance and larger fringing fields associated with TiO_2 gate oxide material. Since the optimized device exhibiting lower f_k, hence at a lower frequency,the device becomes unconditionally stable thereby making it a best suitable device for high frequency applications.

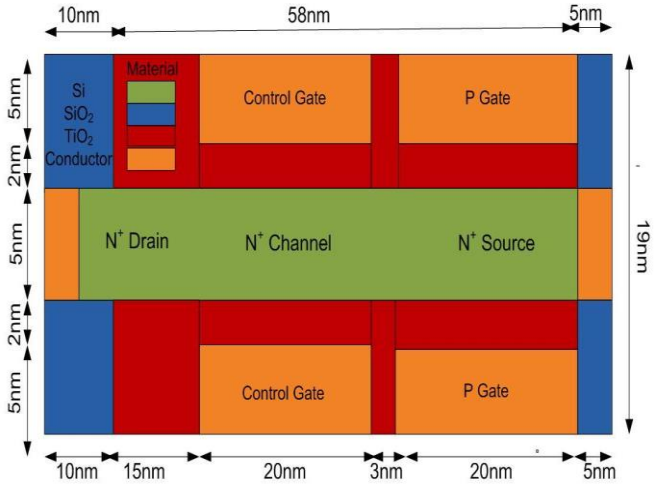


Fig. 11. Optimized DGJL-TFET with TiO_2 as Gate Oxide.

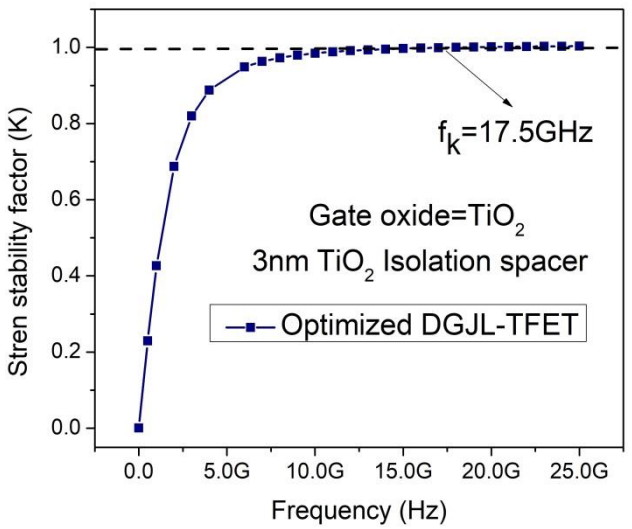


Fig. 12. Stern Stability Factor (K) as a Function of Frequency for optimized DGJL-TFET.

VII. Conclusion

In this paper, we have investigated the influence of high-k gate dielectrics / high-k spacer on the RF stability performance of 20nm channel DGJL-TFET. From the simulation results, it is observed that the device with low-k gate dielectric attains stable at lower frequencies, but yielding low ON current. It is also noticed that high-k as an isolation spacer is responsible for the obtaining lower f_k. In addition, the effect of geometrical variability's towards f_k is studied and results showed decreasing trend with the increase of the spacer length, gate oxide thickness, but an increased trend for drain voltage. Finally, the optimized structure is proposed for yielding the better stability and lower f_k, by which we can avoid the additional circuit. The proposed DGJL-TFET exhibits f_k of 17.5GHz. In summary the DG-JTFET is promising device, suitable for low power and analog/RF applications.

Acknowledgment

This work is supported by SERB (DST), Government of India Grant No: ECR/2017/000220.

References

[1] Reddick, William M., and Gehan AJ Amaratunga. "Silicon surface tunnel transistor", Appl. Phys. Lett , 67(4):494-496. 1995.

[2] Appenzeller, J., Y-M. Lin, J. Knoch, and Ph Avouris. "Band-to-band tunneling in carbon nanotube field-effect transistors", Phys. Rev. lett, 93(19):196805. 2004.

[3] Wang, P-F., K. Hilsenbeck, Th Nirschl, M. Oswald, Ch Stepper, M. Weis, D. Schmitt-Landsiedel, and W. Hansch. "Complementary tunneling transistor for low power application". Solid-State Electron,48(12):2281-2286, 2004.

[4] Boucart, Kathy, and Adrian Mihai Ionescu. "Double-Gate Tunnel FET With High-$\kappa $ Gate Dielectric". IEEE Trans. Electron Devices, 54(7):1725-1733, 2007.

[5] Koswatta, Siyuranga O., Mark S. Lundstrom, and Dmitri E. Nikonov. "Performance comparison between pin tunneling transistors and conventional MOSFETs". IEEE Trans. Electron Devices,56(3): 456-465, 2009.

[6] Nikam, Vishwanath, Krishna K. Bhuwalka, and Anil Kottantharayil. Optimization of n-channel tunnel FET for the sub-22nm gate length regime. IEEE Dev. Research Conference. 2008; 77-78.

[7] Colinge, Jean-Pierre. "Multiple-gate soi mosfets". Solid-state electron, 48(6):897-905, 2004.

[8] Sangeeta Mangesh, Pradeep Chopra and Krishan K. Saini, "A Trapezoidal Cross-Section Stacked Gate FinFET with Gate Extension for Improved Gate Control" International Journal of Advanced Computer Science and Applications (IJACSA), 10(1),2019. Doi:https://dx.doi.org/10.14569/IJACSA.2019.0100125

[9] Chen, Qiang, Keith A. Bowman, Evans M. Harrell, and James D. Meindl. "Double jeopardy in the nanoscale court [mosfet modeling]". IEEE Circuits and Devices Magazine, 19(1):28-34, 2003.

[10] Bhuwalka, Krishna Kumar, Stefan Sedlmaier, Alexandra Katharina Ludsteck, Carolin Tolksdorf, Joerg Schulze, and Ignaz Eisele. "Vertical tunnel field-effect transistor", IEEE Trans. Electron Devices, 51(2): 279-282, 2004.

[11] Zhang, Qin, Wei Zhao, and Alan Seabaugh. "Low-subthreshold-swing tunnel transistors." IEEE Electron Device Lett 27(4): 297-300, 2006.

[12] Boucart, Kathy, and Adrian Mihai Ionescu. "Length scaling of the double gate tunnel FET with a high-k gate dielectric". Solid-State Electron, 51(11):1500-1507, 2007.

[13] Björk, M. T., J. Knoch, H. Schmid, H. Riel, and W. Riess. "Silicon nanowire tunneling field-effect transistors". Appl. Phys. Lett, 92(19):193504, 2008.

[14] Damrongplasit, Nattapol, Sung Hwan Kim, and Tsu-Jae King Liu. "Study of random dopant fluctuation induced variability in the raised-Ge-source TFET". IEEE Electron Device Lett 34(2):184-186, 2013.

[15] Tirkey, Sukeshni, Dheeraj Sharma, Dharmendra Singh Yadav, and Shivendra Yadav. "Analysis of a novel metal implant junctionless tunnel FET for better DC and analog/RF electrostatic parameters". IEEE Trans. Electron Devices, 64(9):3943-3950,2017.

[16] Colinge, Jean-Pierre, Chi-Woo Lee, Aryan Afzalian, Nima Dehdashti Akhavan, Ran Yan, Isabelle Ferain, Pedram Razavi. "Nanowire transistors without junctions". Nat Nanotechnol, 5(3):225, 2010.

[17] Lee, Chi-Woo, Aryan Afzalian, Nima Dehdashti Akhavan, Ran Yan, Isabelle Ferain, and Jean-Pierre Colinge. "Junctionless multigate field-effect transistor". Appl. Phys. Lett, 94(5):053511, 2009.

[18] Ghosh, Bahniman, and Mohammad Waseem Akram. "Junctionless tunnel field effect transistor". IEEE Electron Device Lett, 34(5):584-586, 2013.

[19] Bal, Punyasloka, M. W. Akram, Partha Mondal, and Bahniman Ghosh. "Performance estimation of sub-30 nm junctionless tunnel FET (JLTFET)". J. Comput. Electron, 12(4):782-789, 2013.

[20] Akram, M. W., and Bahniman Ghosh. "Analog performance of double gate junctionless tunnel field effect transistor." J. Semicond, 35(7):074001, 2014.

[21] Dutta, Umesh, M. K. Soni, and Manisha Pattanaik. "Simulation study of hetero dielectric tri material gate tunnel FET based common source amplifier circuit". Int. J. Electron. Commun. 99:258-263,2019.doi https://doi.org/10.1016/j.aeue.2018.12.004.

[22] Khorramrouz, Fayzollah, Seyed Ali Sedigh Ziabari, and Ali Heydari. "Analysis and study of geometrical variability on the performance of junctionless tunneling field effect transistors: Advantage or deficiency?". J. Nano Dimens 9(3):260-272, 2018.

[23] Bal, Punyasloka, Bahniman Ghosh, Partha Mondal, M. W. Akram, and Ball Mukund Mani Tripathi. "Dual material gate junctionless tunnel field effect transistor". Journal J. Comput. Electron, 13(1):230-234, 2014.

[24] Abadi, Rouzbeh Molaei Imen and Seyed Ali Sedigh Ziabari, "Improved performance of nanoscale junctionless tunnel field-effect transistor based on gate engineering approach" Appl. Phys. A: Mater. Sci. Process, 122(11): 988, 2016.

[25] Priya, G. Lakshmi, and N. B. Balamurugan. "New dual material double gate junctionless tunnel FET: Subthreshold modeling and simulation". Int. J. Electron. Commun. 99:130-138, 2019. doi: https://doi.org/10.1016/j.aeue.2018.11.037.

[26] Rahi, Shiromani Balmukund, and Bahniman Ghosh. “High-k double gate junctionless tunnel FET with a tunable bandgap”. RSC Adv. 5(67);54544-54550, 2015.

[27] Singh, Sapna, and Sudakar Singh Chauhan. “TCAD simulations of double gate junctionless Tunnel field effect transistor with spacer”. In 2017 International Conference on Computing, Communication and Automation (ICCCA) 1441-1444, 2017.

[28] Raushan, Mohd Adil, Naushad Alam, and Mohd Jawaid Siddiqui. “Performance enhancement of junctionless tunnel field effect transistor using dual-k spacers”. J. Nanoelectron. Optoelectron,13(6):912-920, 2018.

[29] Raushan, Mohd Adil, Naushad Alam, Mohd Waseem Akram, and Mohd Jawaid Siddiqui. “Impact of asymmetric dual-k spacers on tunnel field effect transistors”. J.Comput. Electron, 17(2):756-765, 2018.

[30] Sivasankaran, K., and P. S. Mallick. “Stability performance of optimized symmetric DG-MOSFET”. J. Semicond, 34(10):104001, 2013.

[31] Raju, Veerati, and K. Sivasankaran. “Impact of high k spacer on RF stability performance of double gate junctionless transistor”. Int. J of numer. Model. 32(1):e2481, 2019.

[32] Pon, Adhithan, Arkaprava Bhattacharyya, B. Padmanaban, and R. Ramesh. "Optimization of the geometry of a charge plasma double-gate junctionless transistor for improved RF stability." Journal J. Comput. Electron, 18: 906, 2019. https://doi.org/10.1007/s10825-019-01340-4

[33] Silvaco, Version 5.15.32.R., 2009. [Online]. Available http://www.silvaco.com

[34] Hänsch, W., Th Vogelsang, R. Kircher, and M. Orlowski. “Carrier transport near the Si/SiO2 interface of a MOSFET”. Solid-State Electron 1989; 32(10):839-849.

[35] Schenk, A. “A model for the field and temperature dependence of Shockley-Read-Hall lifetimes in silicon”. Solid-State Electron, 35(11):1585-1596, 1992.

[36] Lide, David R. “CRC Handbook of Chemistry and Physics CRC”. Boca Raton 2008.

[37] Rollett, J. “Stability and power-gain invariants of linear twoports”. IRE Trans. Circuit Theory, 9(1):29-32, 1962.

[38] Ku, W. H. “Unilateral gain and stability criterion of active two-ports in terms of scattering parameters”. Proceedings of the IEEE, 54(11):1617-1618, 1966.

[39] Sivasankaran, K., D. Kannadassan, K. Seetaram, and P. S. Mallick. “Bias and geometry optimization of silicon nanowire transistor: radio frequency stability perspective”. Microwave and Optical Technology Letters, 54(9):2114-2117, 2012.

[40] Sarkar, Angsuman, Aloke Kumar Das, Swapnadip De, and Chandan Kumar Sarkar. “Effect of gate engineering in double-gate MOSFETs for analog/RF applications”. Microelectron. J 43(11):873-882, 2012.

A Robust Optimization Approach of SQL-to-SPARQL Query Rewriting

Ahmed Abatal[1], Mohamed Bahaj[2], Soussi Nassima[3]
Mathematics and Computer Science Department
Hassan I University, Faculty of Sciences and Techniques Settat, Morocco

***Abstract*—In order to ensure the interoperability between semantic web and relational databases, several approaches have been developed to ensure SQL-to-SPARQL query transformation direction, but all these approaches have the same weakness. In fact, they convert directly the input SQL query to its equivalent SPARQL one without any pre-processing phase enabling the optimization of this input query filled by users before starting the conversion process. This weakness has motivated us to add a pretreatment phase aiming to optimize the most important SQL statements which seem to have the biggest impact on the effectiveness of the transformed queries. Our main contribution is to enrich these rewriting systems by adding an optimization layer that integrate a set of simplification rules of Left, Right and Full Outer Join in order to avoid, firstly unnecessary operations during the conversion process, and secondly SPARQL queries with a high complexity due to Optional patterns obtained from outer join in this conversion context.**

Keywords*—SQL-to-SPARQL; outer join optimization; query transformation; SQL simplification; query optimization layer

I. Introduction

In the last decades, the semantic web [10] has emerged as an extension of the classic web aiming to exploit the full web potential by providing a common framework for knowledge to be shared across applications. It is a W3C recommendation that offers an easier way to search, share, reuse and combine information. It allows machines to understand the semantics of data on the web in order to conceive a globally-extended knowledge base that links data from different sources and ensure a better cooperation between computers and people. The semantic web uses numerous technologies to achieve the previous goals: RDF (Resource Description Framework) [4] as a flexible and standard data model for representing information on the web and make it machine readable, OWL (Web Ontology Language) [3] as the famous language of knowledge representation for creating structured ontology and SPARQL query language [5] for querying data from RDF graphs.

However, the majority of web data is stored in relational databases, which motivate the web researchers to develop a set of methods aiming to offer a better interoperability context between the both systems. In this light, some approaches have been made regarding SQL to SPARQL query transformation in order to facilitate data extraction for relational users by querying RDF stores with SQL language, but unfortunately, all these approaches have the same weakness in their proposed systems since they convert directly the input SQL query to its equivalent SPARQL one without any optimization phase enabling the pre-processing of this SQL query before starting the mapping process.

This weakness has motivated us to operate in this topic so as to remedy this gap and establish an intermediate step aiming to improve existing SQL-to-SPARQL mapping approaches by optimizing outer join clauses that leads to generate equivalent SPARQL query with Optional patterns responsible of the high complexity of the output query (SPARQL complexity evaluation PSPACE-hard [9]).

The remainder of this paper is organized as follows: Section II presents the key contribution of some related works. Section III gives some theoretical background related to the current topic. Section IV describes the main rules using by our optimization layer. Section V presents the functional architecture of the conceived layer and the proposed algorithm that aims to simplify all outer join types (left, right and full outer join) before starting the SQL-to-SPARQL conversion process. In Section VI we expose the java application implementing our solution and a comparison results summarizing the executing time of generated SPARQL queries in the optimized and direct mode. Finally, Section VII concludes our work and suggests some future extensions of this topic.

II. Related Works

Since the exponential emergence of semantic web in the last decades, researchers in the four corners of the earth have been interested in the interoperability between the semantic web and relational world, considered as one of the most used database management system until today, without physical transformation of data by elaborating a conversion context of their query languages (SQL and SPARQL). Several researches were particularly interested of SQL-to-SPARQL translation direction such as RETRO, SQL2SPARQL and others.

SQL2SPARQL method [2] operates in this light aiming to convert a classic and simple SQL queries into an equivalent SPARQL ones combining the transformation rules presented already in other works so as to realize a dynamic mapping. In addition, RETRO method [6] provides interoperability between relational database and RDF Stores by translating basic and composed SQL query (Union, Intersect and inner join in its simple form) to a semantically equivalent SPARQL one. In the same context, R2D method [8] proposes a mechanism integrating SQL-to-SPARQL translation by converting SQL queries, with pattern matching and aggregation, into the SPARQL equivalent ones. Similarly, the researchers of [1] explain their proposing approach of

interrogating RDF data using SQL queries via an algorithm that convert each clause of SQL queries (simple and complex ones without outer join) into an equivalent SPARQL ones.

Based on the previous analysis, we note that all these approaches have the same and common weaknesses: firstly, they operate just on simple SQL input queries without any consideration of outer join operations. Furthermore, the first work established in SQL-to-SPARQL conversion direction that takes into consideration the outer join in their simple and nested form is presented in [11]. Secondly, all existing solutions convert directly the input SQL query to its equivalent SPARQL one without any pre-processing phase enabling the optimization of outer joins and avoiding unnecessary operations during the conversion process.

To the best of our knowledge, this paper is the first work developed in this topic treating a detailed simplification of the input SQL queries before starting the SQL-to-SPARQL transformation process, more precisely; the simplification of outer joins clauses generating OPTIONAL patterns responsible of the high complexity of SPARQL queries.

III. Preliminaries

This section introduces our work by giving some theoretical background of the different operations used in this optimization approach.

A. SQL Outer Join

In relational terms, the data is distributed among several tables. In fact, we can imagine a table containing a foreign key to another one. In this case, we need to use joins in order to retrieve information from these both tables in a single query and exploit the full power of relational data-bases to achieve results that combine data efficiently from multiple tables. Joins are a powerful construction of SQL language, but they have to be handled carefully, because a small missed join can easily broke down a database server.

Outer join is one of relational join types used via SQL language aiming to regroup data from two or more tables by returning all rows from at least one of these tables indicated in FROM clause. We can distingue three types of outer joins as illustrated in Fig. 1:

- Left Outer Join: returns all rows from the left table and the matched rows from the right table.
- Right Outer Join: returns all rows from the right table and the matched rows from the left table.
- Full Outer Join: return all rows of left and right tables when the join condition is respected.

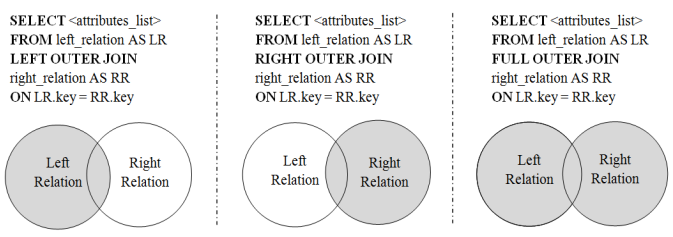


Fig. 1. Descriptive Schema of SQL Outer Joins.

The Left and right outer join are called one-sided outer joins because they preserve only one relation of its joined ones. Whereas, the full outer join is called two-sided outer join [12] because it preserves information from both relations.

As an example, at database level, we suppose that we have two tables: Customer and Order. In fact, we aim to conceive a list of customers with the total amount of their orders by joining these two tables with a left outer join operator so as to return all customers whether they placed any order or not. The final SQL query is conceived as follows:

```
SELECT Customer.fullName, Order.amount
FROM Customer
LEFTOUTERJOIN Order
        ON Customer.cust_id = Order.ord_id
```

B. SPARQL Optional Operator

The main part of SPARQL SELECT query is specified using a graph pattern. In fact, several studies show that 45% of SELECT queries are specified only by graph patterns [7]. They allow users and applications to query RDF data where the entire query pattern must match for there to be a solution. However, RDF has a semi-structured data; this is why SPARQL [5] is able to make queries that allow information to be added to the solution where the information is available, but they do not fail when some part of the query pattern does not match.

In order to realize the previous aim, SPARQL language uses the Option-al operator that combines a pair of graph patterns so as to extend the solution if the patterns matching have been succeeded. Else, the whole query does not fail if the optional pattern match fails.

The equivalent SPARQL query of the previous SQL one is given below:

```
SELECT ?full_name ?amount
WHERE {
?cust_id :Customer ?full_name.
  OPTIONAL {  ?cust_id :Order ?amount   }}
```

IV. Optimization Rules

In this section, we describe the main rules using by our optimization layer in order to simplify SQL queries with outer join operations (in this paper, we consider all outer join types).

Before starting the optimization process, as mentioned in the previous work [11], the input SQL query (having left outer join else we convert the other join types to a left one) has to respect some semantic rules:

- Condition 1: checking the validity of the join condition. In fact, shared variables between join relations must be bound to the same values.
- Condition 2: ensuring that the left outer join succeed by verifying the columns of the right relation which must in no case be all nulls, else the left outer join is reduced to a simple selection of the left table elements. For example, if this condition is not verified for SELECT * FROM R1 LEFT OUTER JOIN R2 ON (R1.a = R2.a),

our system will reduce it to a simple selection as SELECT * FROM R1.

- Condition 3: Before the evaluation of the main Left Outer Join clause, all containing clauses have succeeded. Hence, the attributes of their right relations must be NOT NULL.

Regarding the optimization process, we have used a set of algebraic equivalence rules based on null-rejected condition so as to simplify, whenever possible, the outer join operation and avoid generating Option-al patterns after the SQL-to-SPARQL conversion process.

In fact, a condition is said null-rejected in attribute set A for an outer join operation if it evaluates to False or Unknown on every tuple in which all attributes in A are null.

We consider the previous tables Customer and Order. For example, in the following query (Eq. 1) that aims to make a list of all customers who lives in Casablanca city:

$$\sigma \text{ City} = \text{'Casablanca Customer'} \quad (1)$$

The condition on customer's city (City = 'Casablanca') reject nulls on the attribute City and on any superset of City (*sch(Customer)*).

In this study, we use relational operators on a condition C such as a selection (σ_C), left outer join ($\overset{C}{=\!\triangleleft}$), right outer join ($\overset{C}{\triangleright\!=}$) and full outer join ($\overset{C}{=\!\bowtie\!=}$). The Inner join, returning records that have matching values in both tables, is denoted. The set of attributes referenced by a condition C is called the schema of C, and denoted *sch(c)*.

A. Left Outer Join Simplification

The Left outer join operation can be converted to an inner join one if and only if the WHERE condition is null-rejected on the right relation schema (Eq. 2 in rule 1), else we use the second rule (Eq. 3) to guarantee this equivalence that is presented as a union of the Inner Join and Minus be-tween the both joined relations.

Rule 1: if C_1 is null-rejected on $sch(C_1) \subseteq sch(R_2)$, then:

$$\sigma_{C1}(R_1 \overset{C2}{=\!\triangleleft} R_2) := \sigma_{C1}(R \overset{C2}{\triangleright\!\bowtie} R_2) \quad (2)$$

Rule 2:if the previous condition is not checked, then:

$$R_1 \overset{C2}{=\!\bowtie\!=} R_2 := (R_1 \overset{C2}{\bowtie} R_2) \cup (R_1 \setminus R_2) \quad (3)$$

It is not necessary that the condition of null-rejecting is checked just on relation schemas, but also on any set of attributes satisfying the conditions.

B. Right Outer Join Simplification

Regarding the Right Outer Join operation, its simplification (Eq. 4) is obtained by permuting the joined relation so as to have a left outer join expression and continue to use the rules 1 and 2 defined in the previous paragraph.

Rule 3 :

$$\sigma_{C1}(R_1 = \overset{C2}{\triangleright\!\!=} R_2) := \sigma_{C1}(R_2 \overset{C2}{\triangleright\!\!=} R_1) \quad (4)$$

C. Full Outer Join Simplification

Regarding the Full Outer Join, the test is applied to each side of operation. In fact, if the null-rejected condition is checked on R1 schema then the operation is reduced to a left outer join one (Eq. 5 in rule 4), else if it is checked on R2 schema then we replace the full join with the right outer join (Eq. 6 in rule 5).

Rule 4: if C_1 is null-rejected on $sch(C_1) \subseteq sch(R_1)$, then:

$$\sigma_{C1}(R_1 \overset{C2}{=\!\triangleleft\!=} R_2) := \sigma_{C1}(R_1 \overset{C2}{=\!\triangleleft} R_2) \quad (5)$$

Rule 5: if C_1 is null-rejected on $sch(C_1) \subseteq sch(R_2)$,

$$\sigma_{C1}(R_1 \overset{C2}{=\!\triangleleft\!=} R_2) := \sigma_{C1}(R_1 \overset{C2}{\triangleright\!=} R_2) \quad (6)$$

After performing the simplification of two-sided outer join to one-sided outer join, we process with the same manner as left and right ones.

V. Strategy Overview

A. Functional Architecture

In order to avoid direct transformation of input SQL queries to the equivalent SPARQL ones in existing conversion systems, we have proposed to add an optimization layer to these mapping systems in order to bridge the previous gap by conceiving a functional architecture, schematized in Fig. 2 and composed of five components: Query Analyser & Corrector, Is Null Rejected, Full Outer Join Simplifier, Right to Left Outer Join Converter and Outer Join Optimizer.

B. Optimization Algorithms

Query Analyser & Corrector: This step is very helpful especially in the case when the SQL queries were built based on user input that can be scanned and analyzed in order to correct syntactic errors, if they exist, before starting the optimization process.

Outer Join Optimizer: This component is considered as the main one in our proposed architecture. In fact, it operates on an SQL query filled by users and having an outer join clause in order to return the optimized SQL query at the end of Algorithm 1.

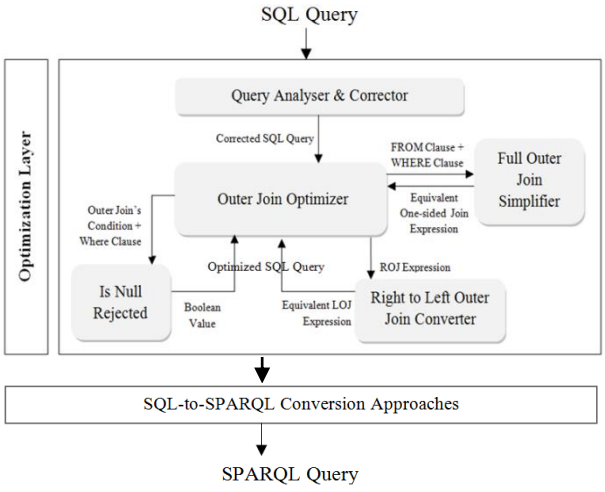


Fig. 2. Functional Architecture of our Optimization Layer.

Firstly, our algorithm extracts each clause of SQL query separately by parsing it to a binary tree, and then we check if the FROM clause contains full outer join expression then we call FullOuterJoinSimplifier in order to replace it by a one-sided join whenever possible. Else if the FROM clause contains a right outer join expression then we use the Right2LeftOJConverter component aiming to convert this expression to a Left outer join one before starting the optimization process using the rules defined in the previous section.

Secondly, we check if the where clause contains a null-rejected condition using the sub component *IsNullRejected* described subsequently, then we replace the one-sided outer join operator with the inner join one; Else we modify the FROM clause before returning the output SQL query.

Algorithm 1: Outer Join Optimizer

Input: SQL query with outer join clause(s), q_{in}
Output: Optimized SQL query, q_{out}
Begin
Tree SQLtree ← parse(q_{in})
q_{in}^{SELECT} ← SQLtree.getSelectClause()
q_{in}^{FROM} ← SQLtree.getFromClause()
q_{in}^{WHERE} ← SQLtree.getWhereClause()
Operator ← q_{in}^{FROM}.getJoinOperator()
if(IsFullOuterJoin(Operator) = True) **then**
/*Replace the FOJ by a one sided OJ whenever possible*/
q_{in}^{FROM}← FullOuterJoinSimplifier(q_{in}^{FROM}, q_{in}^{WHERE})
End if
If (IsRightOuterJoin(Operator) = True) **then**
q_{in}^{FROM}← Right2LeftOuterJoinconverter(q_{in}^{FROM})
End if
/*we check if the where clause contains a null rejected condition*/
R_R ← q_{in}^{FROM}.getRightRelation()
/* LAR_R : List of the right relation attributes*/
LAR_R ← parseList(R_R)
If (IsNullRejected(q_{in}^{WHERE}, LAR_R) = True) **then**
/*Replace the one-sided outer join operator with the inner join*/
q_{in}^{FROM}.setJoinOperator('INNER JOIN')
Else
R_L ← q_{in}^{FROM}.getLeftRelation()
R_R ← q_{in}^{FROM}.getRightRelation()
q_{in}^{FROM} ← q_{in}^{FROM} + 'UNION' + R_L + 'MINUS' + R_R
End if
q_{out} ← q_{in}^{SELECT} + q_{in}^{FROM} + q_{in}^{WHERE}
Returnq_{out}
End Algorithm

Full Outer Join Simplifier This is the main component in our system as presented in the Algorithm 2. In fact, it takes as input an SQL query qin containing an outer This sub component takes as input the Full outer join expression and the where clause of the SQL query in order to return the equivalent one-sided join expression. Firstly, we extract the left and right relations of the input join expression and then we parse them to a list of attributes for an ulterior use. In the next step, we check if the null-rejected condition is checked on the left relation schema then the operation is reduced to a left outer join one (Rule 4), else if it is verified on the right relation schema then we replace the full join with the right outer join (Rule 5).

Algorithm 2: Full outer join simplifier

Input: Full outer join expression (Exp_{in}), Where conditions (WhereClause)
Output: Equivalent one-sided join expression, Exp_{out}
Begin
R_L ← Exp_{in}.getLeftRelation()
R_R ← Exp_{in}.getRightRelation()
OnCond ← Exp_{in}.getOnCondition()
$schR_R$ ← parseList(R_R)
$schR_L$ ← parseList(R_L)
If (IsNullRejected(WhereClause, $schR_R$) = True) **then**
Exp_{LOJ} ← R_R + "Right Outer Join" + R_L + OnCond
Elseif(IsNullRejected(WhereClause, $schR_L$) = True)**then**
Exp_{LOJ} ← R_R + "Left Outer Join" + R_L + OnCond
End if
Return Exp_{LOJ}
End Algorithm

Is Null Rejected Takes as input the where clause and the attribute list of one joined relation in order to test if this clause contains a null-rejected condition or not on the given schema relation (the second input) and return a Boolean value. Algorithm 3 presents below the detailed instructions to realize this goal.

Algorithm 3: Is Null Rejected

Input: Where condition (WhereClause), Set of relation attribute ($AttrList_R$)
Output: Response as aBoolean value
Begin
int i ← 0
List WCL ← parseList(WhereClause)
while (i < WCL.size()){
cond ← WCL[i]
attribute ← cond.getAttribute()
if (attribute.isIncluded($AttrList_R$) = True) **then**
If (cond is like a joined tables attribute with 'IS NOT NULL' condition
OR
cond is evaluated to False or Unknown for the generated null tuples) **then**
Return True
End if
end if
i ← i +1
End while
Return False
End Algorithm

Right to Left Outer Join Converter the main aim of this sub-component is to convert a ROJ expression to an equivalent LOJ one applying this swapping ruleRelation_L Right Outer Join $\text{Relation}_R \Leftrightarrow$ Relation_R Left Outer Join Relation_L as presented in Algorithm 4 presented below.

Algorithm 4: Right to Left Outer Join Converter

Input: ROJ expression, ROJexp
Output: Equivalent LOJ expression, LOJexp
Begin
LeftRelation = ROJexp.getLeftExpression()
RightRelation = ROJexp.getRightExpression()
OnCond ← ROJexp.getOnCondition()
LOJexp ← RightRelation + 'Left Outer Join' + LeftRelation + OnCond
Return LOJexp
End Algorithm

VI. Implementation

In order to improve the effectiveness of our proposed approach, we have developed a java application offering to relational users an efficient tool to query RDF stores with optimized SPARQL queries. The experiments were carried out on a PC with 8 GB RAM, intel Xeon X7460 2.7 GHz.

We present below some examples of SQL queries supported by our system and its equivalent SPARQL ones obtained by a direct conversion and optimized one. In the example illustrated in Fig. 3, we have operated on an SQL query with left outer join aiming to select researchers having supervisors. Thus, the equivalent SPARQL query is composed obviously of an Optional pattern that corresponds to the left join in the input query. However, if we analyze deeply the input query, we note that it contains a null rejected condition (Supervisor.sup_id IS NOT NULL) in where clause that leads to transform left outer join operation to an inner join one (Rule 1). Consequently, the SPARQL equivalent query conceived from the optimized SQL one is composed by a simple graph pattern instead of an optional pattern responsible of the high complexity of SPARQL queries (Fig. 4).

In order to compare our optimized approach with others converting SQL queries to SPARQL ones without any preprocessing phase, we have used JENA framework and an RDF file with 5 million triples to execute a set of SPARQL queries generated by our application via the direct and optimized way, and then we have compared the execution time of each one. The comparison results are summarized in Fig. 5, which ensure the effectiveness of our proposed work that generates optimized SPARQL queries as shown in the diagram.

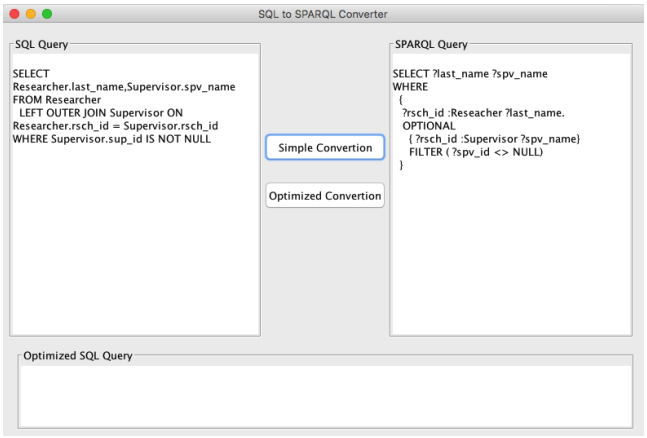


Fig. 3. Direct Conversion Example of SQL Query with Left Outer Join.

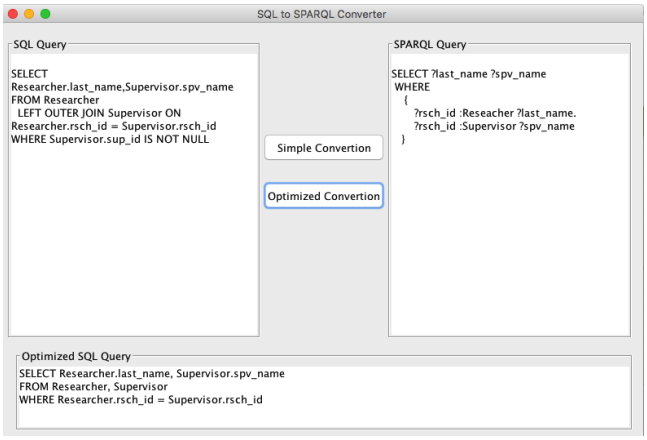


Fig. 4. Optimized Conversion Example of SQL Query with Left Outer Join.

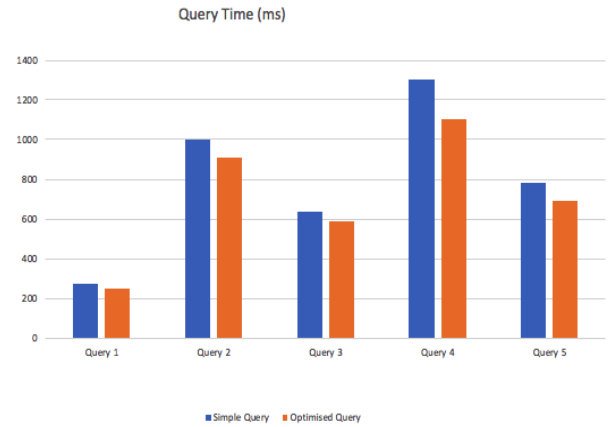


Fig. 5. Diagram of SPARQL Queries Execution Time via the Simple and Optimized Conversion Mode.

VII. Conclusion

In this paper, we have contributed in the enhancement of existing SQL-to-SPARQL conversion approaches by adding an optimizer and pretreatment layer to these systems in order to simplify all outer join types (left, right and full) in SQL queries aiming to avoid the generation of optional pattern(s), whenever possible, in the output SPARQL queries responsible of its high complexity. In addition, we have implemented our algorithm in order to improve its performance on a real data and make our strategy easily and effortless exploitable by the target audience.

The major limitation of our approach that is based only on null rejected properties of join condition so as to simplify SQL queries and avoid the generation of optional pattern(s), whenever possible, in the output SPARQL queries responsible of its high complexity.

In the future work, an obvious extension of the approach is to add more simplification rules to our Optimizer component and to integrate this framework into the relational database management system in order to offer relational users a direct and optimized extension to the semantic world.

References

[1] Alaoui, L., Abatal, A., Alaoui, K., Bahaj, M., & Cherti, I. (2015, July), "SQL to SPARQL Mapping for RDF querying based on a new Efficient Schema Conversion Technique", International Journal of Engineering Research and Technology, Vol. 4, No. 10.

[2] Antal, M., Anechitei D., & Cuza, A. I. (2012), "SQL2SPARQL".

[3] Bechhofer, S. (2009), "OWL: Web ontology language", In Encyclopedia of Database Systems, Springer US.

[4] Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J. J., & McBride, B. (2014), "RDF 1.1 concepts and abstract syntax", W3C recommendation, 25(02).

[5] Harris, S., Seaborne, A., & Prud'hommeaux, E. (2013), "SPARQL 1.1 query langu-age", W3C Recommendation, 21.

[6] Rachapalli, J., Khadilkar, V., Kantarcioglu, M., & Thuraisingham, B. (2011), "RETRO: a framework for semantics preserving SQL-to-SPARQL translation", The University of Texas at Dallas, 800, 75080-3021.

[7] Picalausa, F., & Vansummeren, S. (2011, June), "What are real SPARQL queries like? ", Proceedings of the International Workshop on Semantic Web Information Manage-ment, p. 7, ACM.

[8] Ramanujam, S., Gupta, A., Khan, L., Seida, S., & Thuraisingham, B. (2009), "In Seman-tic Computing", IEEE International Conference, pp. 303-311.

[9] Schmidt, M., Meier, M., & Lausen, G. (2010, March). "Foundations of SPARQL query optimization", In Proceedings of the 13th International Conference on Database Theo-ry ACM, pp. 4-33.

[10] Shadbolt, N., Berners-Lee, T., & Hall, W. (2006), "The semantic web revisited", IEEE intelligent systems, 21(3), 96-101.

[11] Soussi, N., & Bahaj, M. (2017), "Semantics preserving SQL-to-SPARQL query trans-lation for Nested Right and Left Outer Join", Journal of Applied Research and Tech-nology.

[12] Pérez, J., Arenas, M., & Gutierrez, C. (2006, November), "Semantics and Complexity of SPARQL", In International semantic web conference , Springer, Berlin, Heidelberg, pp. 30-43.

Impact of Scrum and Tactic Workflow Management System on Organization Performance (A Study on Animation Studios in Pakistan)

Abdul Wahab Khan[1], Usman Khan[2]
Maaz Bin Ahmad[3]
College of Computing and Information Sciences
PAF- Karachi Institute of Economics and Technology
Karachi, Pakistan

Farhan Shafique[4]
Dept. of Computer Science
FUUAST
Karachi, Pakistan

***Abstract*—Assessing and considering the efficiency effect of scrum management system is a significant and troublesome issue for a researcher. We prescribe that you best comprehend the adequacy of scrum management system applications by breaking down at the data handling level in the animation studio. As we going to concentrate on the progression of data with others in the association procedure is moderately not coupled, to oblige the estimation issue. In light of the aftereffects of this research, it will reason that in the wake of actualizing the scrum management system framework, the organization will turn out to be progressively proficient and compelling underway exercises. Also, its exhibition will upgrade and the greater part of the issues will resolve effectively which prompted better profitability and a superior notoriety in the market. The study additionally attempted to underline the effect of animation studio the scrum management system. It is proposed to determine how the scrum management system causes a movement studio to work adequately. Research proposes that activity studios ought to give adaptability to execute the executives data frameworks, ought to likewise be noticed that to advance control of the organization's market and obtain sufficient programming and a proper program through correspondence media associations to meet the worldwide scrum management system business condition in the developing business sector development and extension.**

Keywords*—MIS (Management Information System); scrum management system; deployment; performance; animated studios

I. Introduction

The universe of Information System (IS) in associations gives various cases of fruitful IS execution that gives advantages to the two business and representatives working for them. These advantages incorporate better gainfulness and better performance of a complete organization [8] [12] [15]. This is also true for effective business processes or individual-level work routes [6] [10] [13].

Based on resource-based view (RBV) theory, a few scholastics recommend that interior business procedures could be essential factors that connect IT capacity and organizational performance [7] [15]. IT capabilities, although it is considered a key asset [21] Can't make an incentive in a vacuum, its part in supporting vital business forms is impacted by extraordinary effects [1] [18] Academics also say that such unusual effects are important for the implementation of IT capabilities, in which conditions are to be changed by changing conditions [15], [17]. In the present period Business administrators are unmistakably progressively included about the impact of rivalry as contrast with a couple of years back. They ought to need to react to restricting dangers not only from local sources yet territorial as well, national and worldwide sources; they have likewise attempted to locate every single accessible open door in the national and worldwide condition in a split second. Deregulation has expanded focused weight for associations, endurance, and flourishing.

The scrum management system likewise ascertains the side idea of data stream and Configuration of associations around the choice focuses. Execution benchmarks are a piece of any great arrangement as well. For reasons unknown, Depending on the accessibility of the pertinent administration data framework as deciding quality as different parts of the arranging procedure.

A. Background

Scrum management system has a significant job in rivalry among organizations. Organizational capacities, for example, the general fitness of an organization to orchestrate its composite HR and different assets adequately to accomplish corporate execution [11]. Organizations can get a contending predominance by picking up or developing authoritative capacities that are commendable, extraordinary, can't be copied in a perfect world and are not commutable in uncommon mixes. Given that the capacities of the association slope to show course confirmation, arbitrary vulnerability, and social trouble, a contention advantage created by size can be continued for more term of time [16] [20].

Limit of IT has been portrayed as the capability to prepare and execute assets of IT mixture with different assets and capacities of the association [4]. Under resource-based view (RBV) IT abilities that displays the features of interesting attributes, Appropriability, non-reproducible, and non-substitutability, could be a wellspring of driving act [18]. Data System researchers have altogether tried the impact of IT limit

on the exhibition of the organization. Organizations that has more IT limit slanted to exceed rivals in a scope of execution gauges dependent on points of interest and costs [4]. Besides, there is expanding confirmation that contention benefits often depend on either organization to make the limit of their IT limit [5].

In animation studios, there are numerous offices and each division works like a column, they ought to have legitimate data so they can settle on a choice on schedule and perform better. Scrum management system will assist them with making better choices and perform well by giving them appropriate data on schedule.

B. History of Waadi Animation

Waadi Animations began in 2014 with ten visual specialists and twenty PCs. Of the twenty frameworks, ten were utilized as rendering hubs, while the rest were distributed as workstations for the illustrators, special visualizations, craftsmanship and sound. They utilized i7 Intel processors, with 16GB frameworks and 2GB realistic cards. To spare costs, they changed over a framework into a server. Later when they started building up their whole activity, they fabricated servers and made hubs as the group developed.

The principal film took eighteen months to finish and in spite of the fact that the last item was a first of its sort in Pakistan, it was of disappointing quality. The underlying driver of this was related with fumble of time and misallocation of assets. Cutoff times were not met and assignments were not being concluded. Basic choices couldn't be made in light of the fact that tests couldn't be audited in time which further postponed the criticism procedure. Issues and protests couldn't be settled in due time and would consider antagonistically the group's presentation. The procedure was Adhoc and being a beginner group, they learned by committing errors as they went.

Another issue that was confronted was the utilization of appropriate innovation. Having no earlier information or experience with respect to what PC frameworks would have the option to hold up under the heaviness of prepared livelinesss. At first, the procedure was run on CPU-based rendering. CPU-based renders would mean a casing would be delivered each hour. As they advanced and with the headway in innovation, they moved to GPU-based rendering which was quicker and practical. Presently, each edge could be delivered in a short time. This implied procedures and undertakings that used to take a year, could now be finished in four months.

C. Implementation

The team is divided into Art, Composite, Modeling, Animation, VFX, Lighting and Render, all in constant liaison with the IT team. The process is broken down into fortnightly sprints.The process of decision making begins with calling team meetings.In those meetings, every member is debriefed on their daily workload. - SCRUM Methodology in the supervision of Head of department and Lighting Supervisor. This can take place between the heads of each department where strategies are outlined for how a script will be processed into an animated film. The goals of each sprint are finalized and then, these are delivered to each department. Each member gives brief answers to three essential questions:

1) What I completed yesterday?
2) What I aim to complete today?
3) What obstacles I face in completing my tasks?

Some departments run independent of one another and do not need to wait upon others to complete allotted tasks.

In those instances, the team leader will decide to pull upcoming deadlines closer to reduce lag and direct attention towards secondary tasks.Although this keeps labor costs low, it is inefficient as later on, it can become burdensome on the team as they have to do overtime to catch up on missed deadlines. But, with the use of the software, we can track the progress of each team member.

As each task is being worked on, the software allows team leaders to assign small, realistic goals within realistic deadlines and then tracks the time spent by each team member in completing it. With accurate up to date information about what percentage of the workload has been completed, team leaders can make decisions about how to plan for the coming fortnights.

D. Research Problem

Information technology is widely viewed as a key reason for the growth of any business [5]. Information is very important and also describe the cable of any company (public limited business, private limited companies, local business, multi-national industries, franchising, retail outlets, and chains). This guides you into more intelligent way and enables the right job to be completed at the right time [22].

Expert of information system discussed that companies must build the good information system to accomplish companies goals [17] [4] In essence, the concept of IT capacity underlines IT resources are shared resources, together with the importance of mobilization and deployment, and other resources and capabilities [4].

a) Research Questions

- RQ1: Does the deployment of scrum management system has an impact on organizational performance?

b) Research Objectives

- RO1: To study the impact of deployment of scrum management system over organizational performance.

c) Hypothesis

- H1: Deployment of scrum management system positively impacts on organizational performance.

E. Contribution of Study

The Animation studio sector goes around different issues and challenges, which impact on the organizational performance directly. We see much lacking in the animation studio system; related to administration, production activities, project planning, project deadlines, and communication issues, etc. There is literature available that talks about different

aspects of organizational performance. Different researches has done primarily on scrum management system in business sectors, hospitals, and education sectors but we did not find any research related to scrum management system deployment in Animation Studios in Pakistan.

It is therefore believed that this research will provide solid groundings to work over this main sector as well specifically in developing countries. Currently, this study caters only two variables i.e. deployment of scrum management system as an independent variable while Organizational Performance (animation studios) as a dependent variable due to limitation of time. But this will provide a road map for others to follow this gap after this research.

The study under consideration variable and it impacts, it is aimed to come with logical groundings to present statistical data to top management, enforce them to take reality base actions in animation studios in Pakistan for the organizational performance.

F. Scope and Limitation of the Study

Animation Industry is an extremely huge industry and furthermore called the cash multiplier industry. This specific examination will concentrate on the Animation Studios of Pakistan. The absolute populace is the staff working in Animations Studios in Pakistan, incorporates creation and non-generation (administrative and managerial) staff. The constraint of the examination is that we have not many activity studios in Pakistan. The complete number of animation studios of Pakistan is 10 with around 1000 creation and around 200 non-generation staff. The assortment of information will be done through a survey and from creation staff just cooking the both gender orientations staff on an accommodation examining premise. The present investigation has one autonomous variable. Scrum management system Deployment in Animation Studio. While one ward variable as the authoritative execution of movement studios. Numerous components influence authoritative exhibition. In any case, this examination is concentrating just on one variable that is thought to be very solid for the liveliness business of Pakistan. The extent of this investigation won't be constrained to this exploration, yet it could be conveyed forward with a similar variable on different areas of Asia and the remainder of the world.

II. Literature Review

Scrum management system indicates that communication is a must for the performance of the organization and management functions linked to the external environment. The scrum management system set a link to communication, making activities and responsibilities possible for administration and managers.

Concern over scrum management system and improved handling result in reduced bottlenecks associated with management processes. The managers were reorganized for years, so the traditional production information for the performance has had limited control value. However, In many companies, this is the only type of data that is regularly collected and analyzed. Management needs a variety of product information regarding the external environment, such as electricity, cooling systems and technology development.

The scrum management system is a functional system to collect, compare, analyze and distribute external and internal information to the company in a very efficient and effective way [19]. The scrum management system can be defined as supporting the executive branch to provide a competitive advantage that must support the organization's goals [14] [9]. Explains that scrum management system works as a planning system for collecting, processing, storage and transmission functions needed to manage the data format information provided.

The scrum management system uses formal strategies to provide appropriate information to all levels of management based on external and internal data and helps them to arrange coordinate and control the auspicious and powerful choices of the exercises they participate in.

An efficient scrum management system framework regularly utilizes PC and other sophisticated technology to process data that mirrors the everyday activities of the organization. In light of the above, scrum management system framework is a coordinated manual PC framework that gives information to help the tasks of administrations and the decision making elements of an organization. Scrum management system framework is additionally an accumulation of individuals, strategies and gadgets sorted out to change over data from interior and outside sources into information and impart such information in a proper frame to administration and management by any means levels.

The scrum management system framework additionally incorporates the association of people, procedures, and contraptions to change over information from inside and outside information sources into data and convey this data in any capacity to the executives and the board in a fitting structure.

A proficient scrum management systemprocess information that mirrors the association's every day exercises by utilizing PCs and other refined advancements. Considering this, scrum management systemis a planned framework structure that gives assistance as far as data to settle on better basic leadership exercises in an association. The scrum management systemstructure likewise incorporates the association of people. Contraptions and procedures to change over information from outer and inner information sources into data and impart this data to organization and the board using any and all means levels.

A. The Software

The software used in the animation industry to manage projects and production tracking simultaneously is called SouthPaw TACTIC. It is a LINUX-based open source software which can be installed on a server and is accessible via both Windows and Mac systems. An operator can connect to it via a T-SQL database that records all employee access information and allow specific permissions to each user i.e. rights for file generation, modification and sharing. TACTIC works through the database which sits on a server, with a

portal front end. A user enters their designated username and password, can view their daily progress. All files they have worked on will be accessible via the portal. Because everything is connected via a centralized system, all complaints and issues can be immediately resolved, performance indicators can be measured and allow greater control over projects which further helps team leaders plan and strategize for the future.

Through TACTIC, the team leaders can assign specific tasks to entire teams by grouping users and placing deadlines on the teams. Alerts are systematically sent out to team leaders and members as deadlines approach. But apart from the production, the software also links this data with other departments such as Administration, Project Management, Human Resources, Marketing, Accounts and IT.

B. Tactic

In the event that something is completing, there must be a work process included. Strategy work process is a work process motor planned around the conviction that each organization, regardless of whether they know it or not, is utilizing work process to do what they do.

In the event that a chief plunks down with a pen and a napkin, they ought to have the option to draw the work process they see around them consistently. In numerous organizations, the manner in which procedures and systems are done is a successful and full grown procedure. As a rule, off-the-rack the executives programming's demands their bearing of sorting out work processes, tossing out what is natural and generally effective. What we are keen on is taking what as of now exists in an organization and carrying it to a more elevated level of ability.

Strategy work process isn't intended to upset; it is intended to incorporate and improve work process pipelines. It utilizes these pipelines to outwardly speak to the inward procedures of an organization, enabling upper administration to put and allocate needs and observing how it will influence the whole work process, continuously. As a pipeline is fabricated, code is being created out of sight that brings ground-breaking computerization at a director's fingertips. We realize that a work process isn't simply from A to B. The TACTIC pipeline underpins the intricacy that can be normal in a genuine workplace through computerized procedures and choice trees.

Everything that occurs in an undertaking is followed through errands, notes, and resources. Strategy work process handles this data through a perplexing information framework and presents this information through Views custom-made to various client types. This joined with an adjustable notice framework guarantees that everybody is seeing the correct data. Strategy tracks computerized resources through the creation procedure progressively, changing information into accessible resources as you work. It robotizes the naming, arranging, following and looking through computerized resources, making work process association quick and programmed. Along these lines, the work process motor guarantees that conveyance is constantly smooth and that records never get lost or disappear. TACTIC workflow is web-based and is accessible from anywhere with an internet connection. When a workflow is hosted in an effective web solution, significant obstacles are overcome in the area of remote people and teams. Two people from across an ocean find themselves using the same ways to work and communicate as two people in the same room. TACTIC Workflow brings automated work processes such as validation, triggers and decision trees with human project management together under a single workflow engine. It steps into an existing or brand new work environment and looks to enable companies through the many tools it has to offer. The result of all of this is that people spend less time wondering what to do and more time doing what needs to be done.

The goal is to advocate the use of an scrum management system in film production in Pakistan to enhance performance and productivity while simultaneously cutting back on the amount of time production can take.

- Advantages of TACTIC
- Provides up to date statistics of the task at hand to all concerned team members.
- Monitoring of performance can be easily measured.
- Allocate and assign reasonable and visible deadlines.
- Can cut costs down by nearly 40% Regular issues.
- Connectivity issues: regular internet outages.

The software needs to be constantly connected to a secure, upgraded and customized LAN but connectivity issues due to low-quality hardware or power outages at the ISP can cause a breakdown in correspondence between nodes and team leaders

- Database administration costs

The server houses the entire database of employee records, script and animation render files. In order to actively maintain this vast collection of data, the system requires skilled IT labor, such as programmers, database administrators as well as systems engineers for the upkeep of the servers.

C. Discussion of Variables

a) Deployment of scrum management system: Scrum management systemmeets the diverse needs by querying the system, system analysis, system modeling, and decision support systems and other systems.

Scrum management system supports the control of management, strategic planning, operations and data flow control between departments. scrum management system assists animation production in distribution of animation tasks in forms of concept art, color keys, modeling, uv's, rigging, animation, light/shade/rendering, VFX and composite, the status of a particular scene or shot or an asset on a discrete level and reference to a variety of production related documents.

Scrum management systemguides Lead of every department namely modeling, animation, lighting, and VFX by providing them with the operational data they need for planning, scheduling and make their participation possible in decision-making at the operations level to control the project

deadlines. Scrum management system assists supervisors to conduct short-term planning, goal setting, and production control. Planning and management control tools support it, respectively.

Scrum management system assists management at the highest level in goal setting, strategic planning and the evaluation of animation production plans and their implementation. The role of scrum management system is to generate information, make better communication, identify the problem and helps in the decision-making process. scrum management systemis a crucial part of the management, administration, and operations of an organization.

b) Organizational Performanc: scrum management systemhas a very crucial role in the organization; this will have an impact on the function, performance, and productivity of the organization. The impact of scrum management systemon the function under management support, scrum management system manages animation production tasks, schedule meetings and deadlines for tasks, and production control and personnel become more efficient. The monitoring of overall production becomes controllable. The production heads are kept alert by providing certain information for tracking the overall performance and project completion percentage. This results in predicting and controlling the timelines and taking necessary and important decisions when required. Reporting system based on the information provided by the Leads, Head of departments and supervisors helps in tracking, managing and controlling the tasks workflow and saving valuable time.

Scrum Management System in animation production helps the organization in understanding the life and complexity of the project. The scheduling of tasks for the production clearly shows the estimated timeline and approximate delivery time of the project. It helps in dealing with the potential bottlenecks of the projects, they can be in the form of artist's needs for the specific job or the technology barrier or can also be a hardware resource required to meet the need of the current project. Scrum management system require systematic business operations to achieve an effective system design. This prompts a stream of activities that muddle the outline of the framework. This enhances business administration by bringing discipline into their activities, as everybody must take after and utilize frameworks and techniques. This process gives a high level of professionalism in business operations.

Business goals and objectives are resulted by scrum management system goals and objectives. This aides in pushing the whole association one way toward goals and objectives, giving applicable data to the association and down the line to the working artist.

A well-designed system with an emphasis on the Lead and supervisor affects the effectiveness of management. The information foundation motivates an enlightened lead and supervisors to use various management tools. This helps them to resort to exercises such as experimentation. Using computers allows you to use tools and methods that you cannot use manually. Impact on management ability to perform. This greatly increases the effectiveness of decision-making.

Scrum management System in the animation production process plays a very important role and is considered to be a backbone for achieving the goals of the project. It's not just merely a system for managing tasks but a workflow of precision. The heavy paperwork of supervisors is transferred to a computerized system, making it easier for the human mind to work better. It should be noted that most of the workforce is involved in this activity in the organization. This Scrum management System has a direct impact on this cost.

D. Hypothesis Development

This study aims to target animation studios of Pakistan for studying the effect of understudy variables on the organizational performance of Animation Studios. Animation Studios Performance is considered as a dependent variable while one factor scrum management system deployment is considered as an independent variable; to study their impact on Animation Studios Performance.

Ho: Deployment of Scrum management System positively impacts on Organizational Performance of animation studios.

H1: Deployment of Scrum management System does not positively impacts on Organizational Performance of animation studios.

E. Summary

At current, animation industry of Pakistan faces different challenges that playing a major role in dissatisfaction with their production. This study will results over the impact of MID Deployment on Organizational Performance of animation industry of Pakistan.

III. METHODOLOGY

This research targets the animations industry of Pakistan for studying the effect of Scrum management System deployment on organizational performance. This section portrays the foundation to the study, factors linkages with hypothesis, strategies for gathering information, depicting programming that would be utilized for breaking down the information gathered, setting objective populace and test size, explore approach and structures of factual models. This part will lead this investigation towards the outcomes and talk stage.

A. Conceptual Framework

In this study, We have taken two variables; one is independent while the other is a dependent variable. I aimed to establish the link of an independent variable with the dependent variable Fig. 1.

a) Independent Variable

- Deployment of scrum management system.

Deployment of scrum management system is considered as an independent variable for my study.

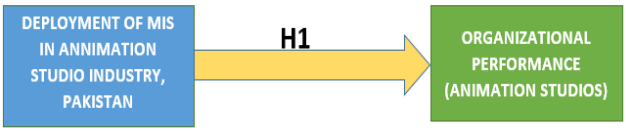


Fig. 1. Conceptual Frame Work.

b) Dependent Variable

- Organizational Performance of Animation Studios

Organizational performance of animation studios concerning the deployment of scrum management system is considered as a dependent variable in my current study.

c) Data and sampling frame work: There is a total of ten animation studios currently working in Pakistan with around staff of 3,000 employed over there. Out of these 3,000 employees, around 2,500 is production staff while around 500 employees are Administrative, marketing, and other support departments. Data has been aimed to collect from three organizations and 52 employees are the sample size calculated from Rao soft calculator with 10% margin of error and confidence level 85% for this study. Data is collected through a self-administered questionnaire.

B. Measurement of Variables

The survey instrument which is used for data collecting is based on five-point Likert scale for Section 2, adopted as per previous literature and published studies. The scale of the study is set as Strongly agree=5, Agree=4, Neutral=3, Disagree=2 and Strongly disagree=1. There are four variables in this section of the instrument i.e. Promotional policies, employee empowerment, Working environment, and Employee satisfaction, each having four statements to inquire responses from the respondent. Refer Table I for a detailed description of instrument adaptation.

C. Statistical Models and Analysis

The aforementioned research will be using Correlation analysis for examining degree of association between variables. To study the impact of understudy independent variables on job satisfaction and to test the hypothesis set earlier in the study, this will be done through multiple regression analysis. Regression analysis is a statistical tool when to find the impact of one variable over the other To implement these models and perform descriptive analysis on the collected data, a statistical package for social sciences (SPSS) version 21 will be used.

D. Diagnostic Analysis

Before approaching to collect data through a survey instrument, pilot testing of the instrument has been done. Based on the sample size from the population i.e. 260, 10% of the sample size were asked to fill out the survey instrument. The results of these respondents were used for survey instrument validation and reliability purpose. The value of cron back alpha for these responses was 0.836 which as per the multivariate data analysis book, [10] value of cron back alpha ranges between 0.75 to 0.93 is appropriate; [4] if cron back alpha is higher than 0.70, the constructs used in the study are reliable.

E. Robustness Analysis

Since the value of the Corn back alpha for the instrument received is greater than 0.70, we consider the survey instrument to be reliable and proceed further to get responses from the remaining sample size of the study. Refer Table II for Corn back alpha test results.

F. Summary

This section is the way to drawing closer the previously mentioned research on finding the effect of sending of Scrum management System on hierarchical execution of movements studios in Pakistan. This part unmistakably distinguishes the populace for the exploration and dependent on that example size were determined, explaining the extent of the investigation. This section likewise explains on information gathering systems and methods, displaying diverse factual models and investigations to be utilized for assessing the said speculation. Approval of the instrument is likewise done in this stage.

TABLE. I. SCALES OF THE STUDY

No.	Variable	Items / Statements
1.	Deployment of Scrum management System	Q1. Scrum management System will help me in my work. Q2. Scrum management System will increase my performance by providing correct information. Q3. Scrum management System will help me in decision making by providing information on time. Q4. A software work is better than manual work or verbal work. Q5. Scrum management System allows information in the sort of reports and screens to the manager and many business professionals. Q6. Scrum management System set a link to communication, making activities and responsibilities possible for administration and managers.
2.	Animation Studios Performance	Q7. Scrum management System will have an impact on the function, performance, and productivity of the organization. Q8. The monitoring of performance can be easily measured with the help of Scrum management System. Q9. Scrum management System will help all departments in Animation Studio of Pakistan to perform effectively. Q10. There can be a big role of Scrum management System in an animation studio's performance.

TABLE. II. RELIABILITY ANALYSIS

Reliability Statistics	
Cron Back Alpha	No. of Items
.762	10

IV. DISCUSSION

Past section of this study featured populace for this exploration, test size and information gathering method that have been received. This part gives a spellbinding investigation of the information gathered through surveys and performed examination with relapse strategy to see results of the outcomes dependent on the hypothetical model and speculation created in going before sections. Further, it likewise give dependability investigation of the instrument that is utilized for gathering the information. This part additionally features discoveries of the Linear relapse test for finding the effect of the autonomous variable on the needy variable and dependent on these outcomes; basic investigation over the outcomes accomplished has been performed relating them with look into targets, inquire about questions and suggesting extension for future examines.

A. Respondent Profile

The respondent for this study has been selected from different organizations currently working in Animation Studio Industry, their profile is mentioned below indicating a number of male & female respondents.

Fig. 2 indicates that out of 52 respondents, 32 respondents i.e. 62% of the sample population were male while 20 respondents i.e. 38% of the sample population were female. All (52 respondents) valuable responses have been taken into account for conducting this research (Table III).

B. Descriptive Statistical Analysis

Descriptive statistics are computed in order to assess the statistical properties of the research variable/constructs. Descriptive statistical analysis is presented in Table IV.

TABLE. III. RESPONDENT GENDER COUNT

Gender	Count	
Female	20	
Male	32	
Total	52	

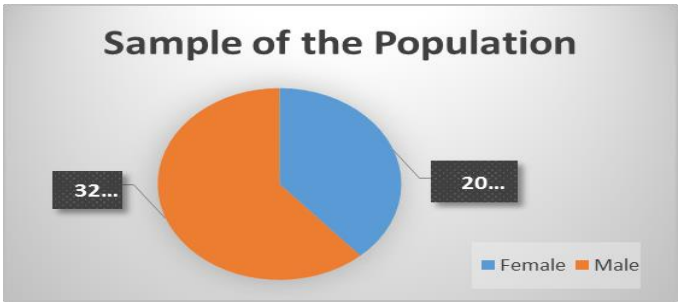


Fig. 2. Graphical Representation of Respondent Gender Count.

TABLE. IV. DESCRIPTIVE STATISTICAL ANALYSIS

Variables	N	Mean	Std. Deviation
	Statistic	Statistic	Statistic
Deployment_ Scrum management System	52	4.3017	.46307
Performance_Increase	52	4.0433	.49190

The aforementioned Table IV shows descriptive statistics of independent and dependent variables with their responses. The sample size of the respondent were 52 and there is no missing data in all the construct responses. Deployment of scrum management system has Mean=4.30 and Standard deviation SD=.46307 while Performance Increase has Mean=4.04 with Standard Deviation SD=.49190.

C. Reliability of the Constructs

In order to examine the reliability of the constructs, the Cornbach alpha values are calculated and stated in Table V.

The aforementioned Table V show that reliability within the construct of Deployment of scrum management system ($\alpha = 0.804$) is the highest followed by Performance Increase ($\alpha = 0.728$). Since both the values of Cornbach alpha is higher than 0.70, therefore, it is safe to assume that the constructs used in this study are reliable [2] [3].

D. Regression Analysis

Regression is a statistical method used for analyzing the nature of the relationship between variables. The analysis includes what type of impact exist between variables; positive or negative linear or non-linear. The impact of the Deployment of Scrum management System on Organizational Performance has been examined through regression analysis. The regression results are summarized in Table VI.

Results of Table VI indicate that the Deployment of Scrum management System has a significant positive impact on Organizational Performance with ($B=0.533$, $p<0.05$). The coefficient of the predictor variable is statistically significant at confidence interval 95%.

TABLE. V. RELIABILITY ANALYSIS

Variables	Mean	Std. Deviation	Cornbach's Alpha
Deployment_ scrum management system	4.3017	.46307	0.804
Performance_Increase	4.0433	.49190	0.728

TABLE. VI. REGRESSION RESULTS (COEFFICIENTS) DEPENDENT VARIABLE: PERFORMANCE_INCREASE

Model	Unstandardized Coefficients		Std. Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	1.752	.562		3.115	0.003
Deployment__ Scrum management System	.533	.130	.502	4.099	.000

TABLE. VII. REGRESSION RESULTS (MODEL SUMMARY)

Model	R	R Square	Adjusted R Square	Std. The error of the Estimate
1	.502[a]	.252	.237	.42980

To answer the aforementioned RQ1, the hypothesis was developed. Regression results suggest that the Deployment of scrum management system has a positive and statistically significant impact on organizational performance with Beta Value = 0.533, p=0.000 which is less than 0.05; therefore null hypothesis is rejected and Alternative (claimed) hypothesis is accepted.

V. CONCLUSION

Information system arrange information according to client prerequisites consequently. Associations live and inhale on monetary information. As people, we utilize and exchange it to look after request. So the board data frameworks give us how we can utilize PC produced scientific models to empower advancements in information the executives and empower individuals to work adequately. It is a territory of between disciplinary research between business science and software engineering.

While working with enormous information or distributed computing as a space, programming like TACTIC enables huge gatherings of artists to envision new foundations while keeping up a smooth work process. The product permits the centralization of innovation that aids basic leadership, arranging and laying out procedures as opposed to film studios attempting to center a great deal of their interests in IT all alone. It is a direct result of a straightforwardness of the consistently changing pipeline through and through that upper administration can control the creation team towards accomplishing higher economies of scale.

Open source software is code that is available to the public, free of charge. Developers can inspect, make changes and constantly monitor the application's performance. This allows users trained in both Macintosh and Windows operating systems to communicate seamlessly without needing to buy new machines or hardware. This appears cost effective but requires the services of trained IT professionals to either be added to the payroll or outsourced, as well as provide the incentive to constantly undergo additional training.

A key test that ought to be tended to is thinking of a procedure to manage the lofty expectation to learn and adapt while presenting another strategy for attempting to both administration and the IT division of a film organization. With regards to dynamic situations, programming prologue to an Adhoc procedure changes the hierarchical culture. Some adjust rapidly, while others will set aside somewhat more effort to make up for lost time with their companions. The business impacts how we assess the innovation speculation also.

A. Summary of Main Findings

Numerous associations are organized dependent on practical territories. This is frequently reflected in an authoritative outline. Ordinarily useful zones incorporate funds, HR, advertising, fabricating, and so forth. In this situation, we have seen that the execution of the TACTIC the executives data framework in a film organization enables upper administration to incorporate liveliness data from different sources, give simple access to execution data in abridged structure, think about notable and current action and aid aggregate investigation with simple to-utilize devices.

B. Limitations of the Study

The animation business is extremely huge and furthermore called the cash multiplier industry. This specific examination will concentrate on the Animation Studios of Pakistan. The all out populace is the staff working in Animations Studios in Pakistan, incorporates generation and non-creation (administrative and authoritative) staff. The constraint of the investigation is that we have not very many movement studios in Pakistan. The all out number of animation studios of Pakistan is 10 with around 1000 generation and around 200 non-creation staff. The assortment of information will be done through a poll and from creation staff just, cooking the catering both genders staff on an accommodation inspecting premise. The present investigation has one autonomous variable. Scrum management System Deployment in Animation Studio. While one ward variable is the authoritative execution of liveliness studios. Numerous elements influence Animation Studios Performance. In any case, this investigation is concentrating just on one variable that is thought to be very sound for the movement business of Pakistan. The extent of this examination won't be constrained to this exploration, yet it could be conveyed forward with a similar variable in different areas of Asia and the remainder of the world.

REFERENCES

[1] ARAGON-CORREA J and SHARMA S (2003) A contingent resource-based view of proactive corporate environmental strategy. Academy of Management Review 28(1), 71–88.

[2] Isabella Mayer Firm Qualitative Research with A Focus On Qualitative Data Analysis . International Journal of Sales, Retailing and Marketing.

[3] Nancy Leech , George Arthur Morgan , Karen Barrett SPSS for Intermediate Statistics: Use and Interpretation. Isbn: 978-1-84872-999-5 January 2015.

[4] Muhammad Nasir Khan, Dr. Tariq Jalees Human Resource Management Practices and Employee Performance in Pakistan. Vol 12, No 2 (2017).

[5] BHARADWAJ A (2000) A resource-based perspective on information technology capability and firm performance: an empirical investigation. MIS Quarterly 24(1), 169–196.

[6] BHATT G and GROVER V (2005) Types of information technology capabilities and their role in competitive advantage: an empirical study. Journal of Management Information Systems 22(2), 253–277.

[7] BURTON-JONES, A., & GALLIVAN, M. J. (2007). Toward a deeper understanding of system usage in organizations: a multilevel perspective. MIS Quarterly 31(4), 657–679.

[8] DEHNING B and RICHARDSON VJ (2002) Returns on investments in information technology: a research synthesis. Journal of Information Systems 16(1), 7–30.

[9] DEVARAJ, S., & KOHLI, R. (2003). Performance impacts of information technology: is actual usage the missing link Management Science, 49(3), 273–289.

[10] Joseph F. Hair Jr. William C. Black. Barry J. Babin Rolph E. Anderson. Multivariate Data Analysis Hair Black Babin Anderson 7th edition.

[11] GABLE, G. G., SEDERA, D., & TAIZAN CHAN (2008). Re-conceptualizing Information System Success: The IS-Impact Measurement Model. Journal of the Association for Information Systems, 9(7), 1–32.

[12] GRANT R (1991) The resource-based theory of competitive advantage. California Management Review 33(3), 114–135.

[13] HENDRICKS, K. B., SINGHAL, V. R., & STRATMAN, J. K. (2007). The impact of enterprise systems on corporate performance: A study of ERP, SCM, and CRM system implementations. Journal of Operations Management, 25(1), 65–82.

[14] IGBARIA, M., & TAN, M. (1997). The consequences of information technology acceptance on subsequent individual performance. Information & Management, 32(3), 113–121.

[15] KENNETH LAUDON, JANE P.LAUDON (2003). Essentials of Management Information System. 5th edition. Prentice-Hall, Inc, New Jersey.

[16] MELVILLE, N., KRAEMER, K., & GURBAXANI, V. (2004). Review: Information technology and organizational performance: An integrative model of IT business value. MIS Quarterly, 28(2), 283–322.

[17] PORTER M (1985) Competitive Advantage. Free Press, New York.

[18] STOEL D and MUHANNA W (2009) IT capabilities and firm performance: a contingency analysis of the role of industry and IT capability type. Information & Management 46(3), 181–189.

[19] WADE M and HULLAND J (2004) Review: The resource-based view and information systems research: review, extension, and suggestions for future research. MIS Quarterly 28(1), 107–142.

[20] BARNEY J (1991) Firm resources and sustained competitive advantage.

[21] BARRY, S. (1998):Introduction to Computer Information System. New York, Rinelart and Winston.

Atmospheric Light Estimation using Particle Swarm Optimization for Dehazing

Padmini. T.N[1]
Department of Embedded Technology
School of Electronics Engineering
Vellore Institute of Technology (VIT)
Vellore, India

Shankar. T[2]
Department of Communication Engineering
School of Electronics Engineering
Vellore Institute of Technology (VIT)
Vellore, India

***Abstract*—For the past decade, many researchers have been working towards the improvement in the visibility of single hazy images, using the haze image model. According to the haze image model, the hazy-free image is restored by estimating the atmospheric light and transmission from a hazy image. The objective of this proposed work is to improve the perceptibility by decreasing the density of haze in the hazy images. The research work was carried to estimate the optimal value of atmospheric light by tuning the weights using a bioinspired technique called Particle Swarm Optimization (PSO) based on the objective of minimizing the fog density. We have selected a fitness function or objective function which incorporates all statistical features to differentiate a clear image from the hazy image. The results are validated with the state-of-the-art, by measuring fog density of the restored image using Fog Aware Density Evaluator (FADE). Also, the results are validated by measuring the Peak signal to noise ratio (PSNR) and structural similarity index (SSI) using ground truth images from Foggy Road image database (FRIDA). This research work demonstrates better results qualitatively and quantitatively.**

***Keywords*—Hazy images; particle swarm optimization; dark channel prior; transmission; atmospheric light**

I. Introduction

Fog, haze, smoke, rain, and snow are the factors which deteriorate the perceptibility of outdoor images. The conditions of the weather vary primarily in size, type, and concentration of the particles in the space. Opaqueness is formed in the lower atmosphere, due to fine suspended particles. Haze affects the perceptibility because it produces a unique gray hue [1]. The difference between haze and fog is that the particles of haze are smaller than fog droplets but slightly larger than air molecules. Similarly, the difference between cloud and fog is that the former exists at a higher altitude than the latter which is at the ground level. As we hinder ourselves to ground level vision, clouds are of less pertinence to us [1].

Fig. 1 shows the effects of scattering of light due to suspended particles in the atmospheric. According to Koschmieder, the light from the sun falls on the scene and gets reflected back to the camera is known as direct transmission. But due to suspended particle like haze, fog and smoke in the atmosphere the light gets scattered in different directions and some portion of the light reaches the camera is known as airlight. Dehazing algorithms were developed based on enhancement and restoration based methods. Restoration based methods use the haze image model which is expressed in equation (1) as follows:

$$H(x,y) = f(x,y)t(x,y) + A(1-t(x,y)) \quad (1)$$

where $H(x, y)$ is the hazy image, $f(x, y)$ is the clear image, A is the atmospheric or ambient light and $t(x,y)$ is transmission which is expressed in equation (2) as follows,

$$t(x,y) = e^{-bd(x,y)} \quad (2)$$

Transmission refers to the light that is reaching the camera from the scene and its value varies between 0 and 1. The exponentially decaying light depends on depth $d(x, y)$ and scattering coefficient b.

Thus from the Koschmieder equation (1), we can conclude that the hazy image has low contrast by the attenuation term f(x,y)t(x, y) and whiteness is blended with the image because of the airlight term A(1-t(x, y)).To obtain the haze-free image, contrast of the image should be improved and colour correction has to be done [2]. Most of the researchers use restoration methods to restore the haze-free (f(x,y)) parameter using the haze image model by calculating the transmission (t) and atmospheric light (A).

In Section 2, the survey of past work in haze removal is discussed. Section 3 explains our proposed work of dehazing using PSO, Section 4 describes outcome of our work. Finally, Section 5 ends with a conclusion.

II. Related Work

The colour and contrast of the images get deteriorated drastically under inclement weather conditions. With the increase in distance between the camera and the object, the level of degradation also increases. Image de-hazing problem was solved by using multi-image technique initially. When only a single deteriorated image is available, then the task becomes more challenging. There have been several algorithms proposed in the past few years for removal of fog using a single image. Information regarding the depth is estimated in these algorithms and it can be determined by categorizing the algorithm into two: interactive or automatic restoration technique.

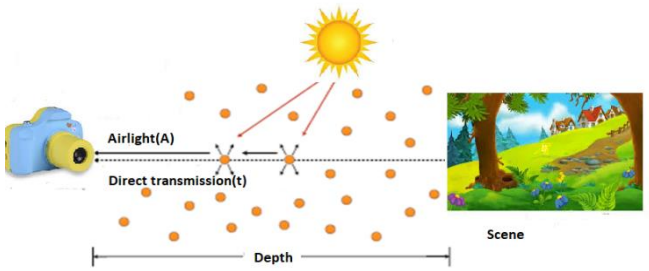


Fig. 1. Pictorial Representation of Effects of Atmospheric Scattering.

Based on the 3D model of the scene, Kopf [3] proposed a procedure applying which can remove the haze effects easily if the information about the depth is known at each pixel. His work is based on interactive restoration technique. Under the automatic restoration technique, there are several methods which have been proposed to remove the haze without the intervention of the user using the single hazy input image. Oakley and Bu [4] developed a model which gives a way to perceive the presence of airlight. It is assumed that the airlight throughout the image is constant. Contrast loss can be corrected easily once the level of airlight is determined. This algorithm falls short when there is no uniformity in airlight over the image. Tripathi [5] has discussed about the interactive and automatic methods in his review paper. Kim [6] improvised the method recommended by Oakley and Bu [4] so that it can be applied even when the airlight is variable over the image. The airlight is estimated by using the luminance image which is determined by the combination of R, G, and B colour component. The airlight map generated is used for restoring the image by getting subtracted from the foggy image. Region segmentation is performed for estimating the airlight in each region which is reflected by depth variation. This method falls short in covering a wide range of depth of the scene. Tan [7] proposed a method based on spatial regularization from the grayscale or single colour image. The image restored produces halos and looks saturated. Independent Component Analysis (ICA) is the concept which Fattal [8] used in his work. To estimate the transmission, Fattal considered ICA. Optical transmission is estimated in the hazy scene and based on this, to increase the scene perceptibility, the scattered light is eliminated. This technique cannot be used for a grayscale image as the colour information is required for restoration. Dense fog is colourless and hence this technique falls short when the fog is very dense. He [9] proposed an effective method involving soft matting and dark channel prior. It is an innovative approach adopted by most researchers which explain as follows, the natural clear images except sky region was found to have pixels with shades of black in any one of the R, G, and B channel. The algorithm uses Levin's soft matting way to improve the transmission map which is computationally expensive. Tarel and Hautiere [10] introduced a computationally effective algorithm known as fast visibility restoration in which airlight is expected to be the percentage between the local average and deviation from the average of the whiteness. The contrast is maximized for the resulting image and it is assumed to have a smooth depth map all over except along the edges. In the case of discontinuities in depth, the quality of the restored image is not so good. Zhu [11] proposed a method based on colour attenuation prior, where the estimation of the depth depends on the difference between brightness and saturation. Once depth is estimated transmission can be determined. But this method fails at edges. Dark channel prior fails to work at sky region. To overcome this problem Wang [12] and Li [13] segmented the sky and other regions separately from outdoor images to estimate and remove the haze from images. Recently, researchers are focusing on implementing optimization algorithms to improve the perception of hazy images based on haze image model. Guo [14] in his work as tuned the aerial perspective and lower bound of transmission for recovering scene radiance using Genetic algorithm. Zhang [15] used Ant colony optimization algorithm using k means clustering to evaluate atmospheric light. Optimization algorithms outperform other methods by obtaining visually compelling results. Recently, Singh [16] has reviewed dehazing algorithm explaining various techniques adopted by researchers in this field.

III. Proposed Method for De-hazing

A simple and efficient method is implemented by extending the work of Dark channel prior by tuning the parameters suitably using Swarm intelligence optimization technique, which is an exhaustive search method for obtaining an optimal value for atmospheric light (A). Also, while calculating the transmission, the aerial perspective ω and while restoring the haze-free image the lower bound of transmission t0 are suitably tuned as specified by Guo [14], using Particle swarm optimization PSO [17] so that the haze-free image is visually compelling. The step-by-step framework for dehazing is summarized in Fig. 2.

For the restoration process, hazy colour image $H^c(x, y)$ is required as the input. To obtain an optimal value for atmospheric light, initially dark channel is obtained from the input hazy image. As per He [9], the dark channel $H^{dark}(x, y)$ is given by,

$$H^{dark}(x,y) = \min_{\Omega(x)}\left(\min_{C\in(R,G,B)} H^{C}(x,y)\right) \quad (3)$$

where Ω(x) is a local square window, *min* is the minimum filter, $H^c(x,y)$ refers to colour channel R, G, B of the input image $H(x, y)$. Dark channel inclines towards zero for clear image. So the equation (3) can be written as

$$H^{dark}(x,y) \to 0 \quad (4)$$

For optimizing the atmospheric light *A*, we followed the weighted method proposed by Shiau [18]. According to Shiau, the atmospheric light is related to illumination of the restored images. If a hazy image as both dark and bright regions then computing the brightest pixel alone from the dark channel will not increase the visibility of the output image. In the weighted method, the threshold value T is obtained from the grayscale of the input hazy image which is used to divide the dark channel regions into bright and dark regions. We took the patch size as 3x3 for computing dark channel as per Shiau [18]. The pixel values higher than T will be under bright region and the other pixel are under the dark region.

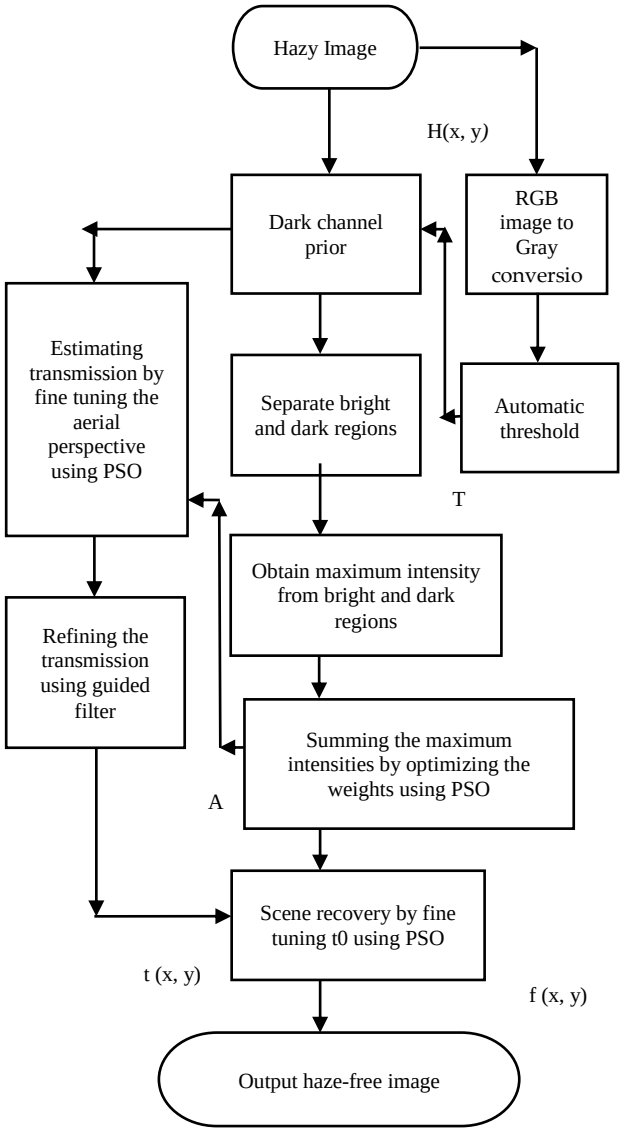


Fig. 2. Overview of our Proposed Method for Dehazing.

The maximum intensities in both the dark and bright regions are found out to compute the atmospheric light. This can be calculated as follows:

$$A_{brightf} = \max_{(x,y)\in bright} \{H^{dark}(x,y)\} \quad (5)$$

$$A_{darkf} = \max_{(x,y)\in dark} \{H^{dark}(x,y)\} \quad (6)$$

where $H^{dark}(x,y)$ is the dark channel of the input hazy image. Let (s_1, t_1) and (s_2, t_2) are the coordinates where $A_{brightf}$ and A_{darkf} are located. To obtain the atmospheric light, the corresponding coordinate pixels of the above said parameters in the input haze image are taken and is expressed as:

$$A = W_b * H^C(s_1,t_1) + W_d * H^C(s_2,t_2) \quad (7)$$

where $H^c(s_1, t_1)$ and $H^c(s_2, t_2)$ are the corresponding pixels of $A_{brightf}$ and A_{darkf} respectively in the input hazy image $H(x,y)$, W_b and W_d are the weights assigned to brighter and dark regions respectively. '*' indicates the multiplication symbol. The total weight,

$$W_b + W_d = 1 \quad (8)$$

In our proposed method, the weights W_d and W_b are optimized using PSO [17] satisfying the equation (8). Using the optimized weight, atmospheric light is estimated. Shiau [18] in his work used a fixed weighted value by considering the sum of the probability within dark region and within bright region.

Before implementing our research work, we tested using different fixed values of weights for calculating the atmospheric light without using optimization technique. Fig. 3 shows the restored image and the effect of atmospheric light when the weights are varied with different fixed values retaining the other parameters used for calculating atmospheric light by Shiau [18].

From the visual comparison of the images in Fig. 3, the significance of weights of atmospheric light can be determined. Hence an optimal weight will solve this problem by improving the quality of the image. Transmission can be obtained once an optimal value of atmospheric light is determined.

The transmission is estimated as per He's [9] approach as follows:

$$t = 1 - \omega \min_{\Omega(x)} \left(\min_{C\in R,G,B} \left(\frac{H(x,y)}{A} \right) \right) \quad (9)$$

where ω is a constant known as aerial perspective whose value lies in the range $0<\omega\leq 1$. Generally, ω is fixed as a small value to maintain naturalness of the scene. Based on He's [9] observation, the dark channel $H^{dark}(x,y)$ is very close to zero, the equation (9) was obtained which is as per the definition of the dark channel prior. In our proposed method the aerial perspective ω is chosen as an optimization parameter to fine tune to get an optimal solution. Since the transmission is assumed to be constant within the window, halo artifacts might occur and so for eliminating the halos, we use edge-preserving smoothing guided filter [19] to refine the transmission. Hence, the restored clear image f(x,y) can be expressed as.

(a) (b) (c) (d) (e) (f)

Fig. 3. Restored Images with Fixed Weights W_b and W_d for Airlight Calculation (a) Hazy Input Image (b) Restored with Wb =0.3 and Wd=0.7 (c) Restored with Wb =0.7and Wd=0.3 (d) Restored with Wb =0.5 and Wd=0.5 (e) Restored with Wb =0.1and Wd=0.9 (f) Restored with Wb =0.9 and Wd=0.1.

$$f(x,y) = \frac{(H(x,y) - A)}{\max(t(x,y), t_0)} + A \tag{10}$$

where *t0* refers to the lower bound of the transmission which is optimized using PSO to get a better quality of scene radiance at the end of the restoration process as discussed by Guo [14] who used genetic algorithm to tune the parameters.

A. Implementation of PSO for Dehazing

In this section, we propose the implementation of PSO for dehazing choosing an optimal weight (W_b ,W_d) for estimating atmospheric light (A) as specified in equation (7), the optimal value for aerial perspective (ω) as specified in equation (9) and optimal value for lower bound of the transmission (t_0) as specified in equation (10). Fig. 4 represents a complete framework of the implementation of an optimization technique to obtain the haze-free image.

In this research work, restoration is considered as an optimization problem and the details are summarized. The conventional PSO algorithm is an exhaustive search based algorithm which considers a collection of particles as Swarm. Let us consider there are N particles in the search space. The optimizing parameter can be represented as follows:

Weights for estimating the atmospheric light W_d = (W_{d1}, W_{d2},……, W_{dN}) and W_b = (W_{b1}, W_{b2},……, W_{bN}). Similarly, aerial perspective ω = (ω_1, ω_2, …… ω_N) and lower bound of the transmission t_0 = (t_{01}, t_{02},…… t_{0N}).

For each iteration, and for each particle, the optimizing parameters (W_d, W_b, ω, t_0) are used to perform the restoration process of the hazy image using the equation (7), (8), (9) and (10). After restoration, the images are evaluated using fitness function.

Since the objective is to minimize the fog density of the restored image, the fitness function or objective function is appropriately chosen in our proposed method to calculate the fog density 'D' of the restored image, referred as Fog aware density evaluator (FADE) proposed by Choi [20]. By fine tuning, the parameters like weights in estimating the atmospheric light, aerial perspective and lower bound of the transmission, the fitness function is minimized to obtain an optimal solution using Particle Swarm Optimization.

FADE computes the density of fog in an image, proposed by Choi [20], by considering the statistical features of 500 hazy and 500 clear images. The statistical features used to compute fog density is explained as follows.

Mean Subtracted Contrast Normalized (MSCN) coefficients obtained from outdoor clear scenes and from foggy images are computed as follows:

$$I(MSCN) = \frac{I_{gray}(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{11}$$

where I_{gray} is the gray scale of the input image, μ is the mean and σ is the standard deviation. As per Choi [20], Normalized Histogram was plotted using MSCN coefficients and found that for denser foggy images histogram incline away from Gaussian whereas the MSCN coefficient of clear images is very close to unit normal Gaussian. The variance of the MSCN coefficient is considered as one of the features. Similarly, the variance of the positive and negative mode of the vertical product of MSCN coefficients is considered as another feature for computing fog density. Apart from MSCN coefficients for measuring fog density, Choi [20] used other statistical features like sharpness which is described by the local standard deviation σ(i,j), the coefficient of variation of sharpness, Contrast energy obtained separately from gray, yellow-blue and red-green channel of the Image. Entropy is another feature used since very fewer details are contained in foggy images. The novel feature is the pixel-wise dark channel prior introduced by He[9] is used. It is found using the equation (3) discussed in dark channel prior topic. As the foggy region is affected by colour, to measure its perceptibility, the fog aware feature used is colourfulness and colour saturation. The degree of difference between gray and colour is colourfulness and the amount of whiteness blended with colour is saturation. With an increase in fog density on the image, the colourfulness and saturation decrease. This is because of the colour shift caused due to airlight scattering. Colourfulness is calculated from the standard deviation and mean of red-green and yellow-blue plane using the following expression:

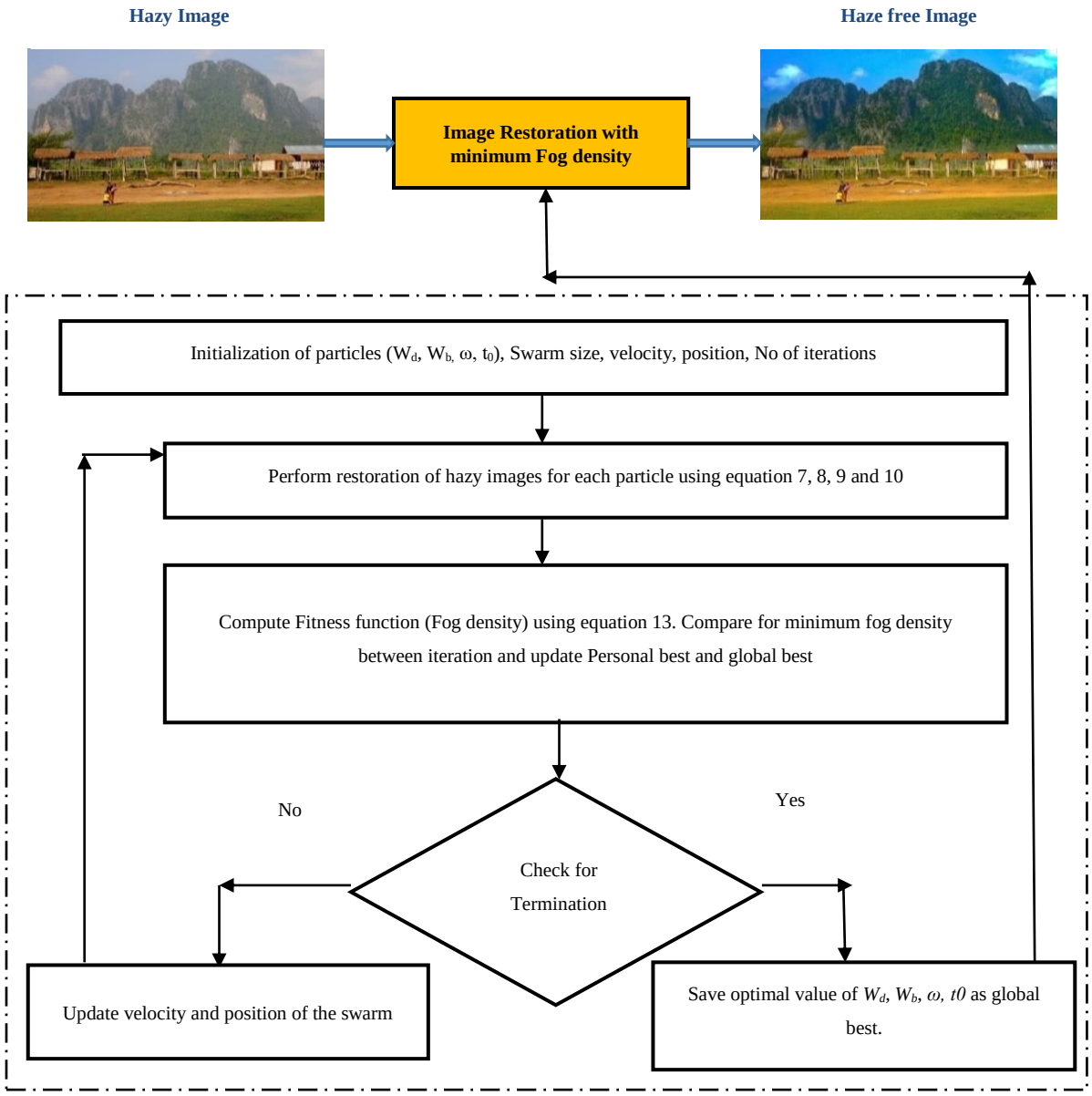


Fig. 4. Framework for Dehazing based on PSO.

$$CF = \sqrt{\sigma_r^2 + \sigma_y^2} + 0.3\sqrt{\mu_r^2 + \mu_y^2} \quad (12)$$

where $\sigma_r, \sigma_y, \mu_r, \mu_y$ are standard deviation and mean of red-green and yellow-blue plane respectively.These statistical features are used to fit as a Multivariate Gaussian model (MVG) to compute the fog density. Fitness function in our approach is chosen which considers all the statistical features used to differentiate between clear and hazy image is specified in equation (13),

$$D = \frac{D_{fft}}{D_{ft} + 1} \quad (13)$$

where D_{fft} is the Mahalanobis distance between MVG models obtained using statistical features acquired from 500 clear image or fog-free natural images and the test image. Similarly, D_{ft} is the Mahalanobis distance between MVG models obtained using statistical features acquired from 500 foggy images and the test image. Lower fog density is indicated by smaller D values.

Each particle calculates the fitness function 'D' using equation (13). The particle can move within the search space based on its current position x(t) and with a velocity v(t) for each iteration 't'. Each particle is capable of storing the personal best p(t) by comparing its fitness value with the previous iteration fitness value. The personal best position of i^{th} particle is obtained by finding the minimum fog density 'D' as follows:

$$P_i(t) = \arg\min[D_i(t)] \qquad i = 1,2,\ldots N \quad (14)$$

Also global best g (t) value is obtained by comparing the best fitness value obtained by all the particles in the search space.

$$g(t) = \arg\min[\, D(P_i(t))] \tag{15}$$

For every iteration, the personal and global best values are updated so that the optimal value is chosen when fitness function 'D' becomes minimum. Also, the velocity *v(t)* and particle position x(t) are updated using equation (16) and (17) until the iteration converges.

$$v(t+1) = wv(t) + c_1 r_1 (p(t) - x(t)) + c_2 r_2 (g(t) - x(t)) \tag{16}$$

$$x(t+1) = x(t) + v(t+1) \tag{17}$$

where w is inertia, c1 and c2 are accelerating coefficients, r1 and r2 are random values ranging between 0 and 1. Finally an optimal global best value of W_d, W_b, ω and t_0 are used to restore the haze-free image with minimum fog density.

IV. Result Analysis and Discussions

Qualitative and Quantitative evaluation comparing with the state-of-the-art is necessary to prove our results obtained are better, so that the visual quality of the restored image is compelling. For choosing the optimal parameters before proceeding with the proposed method we initially tested the parameters manually with different values so that the parameters vary with the fog density. After testing manually the different parameters, we finally selected weights for atmospheric estimation, aerial perspective and lower bound of transmission as the optimal parameters.

A. Parameter Settings

In the proposed method the work was carried using conventional Particle swarm optimization technique. Table I specifies the values set for different parameters like particle size, number of iterations used in our method, inertia weight and the value of accelerating coefficient and Table II specifies the lower and upper bound values of optimum variables chosen in our work. According to He [9], aerial perspective should not be one, so we have selected the upper bound to be 0.99. Similarly transmission lower bound should not be zero, so we have selected 0.0001.

TABLE. I. Parameters and Values

Parameters	Values
Particle Size	20
Iteration	20
Inertia weight(w)	0.5
Accelerating coefficient (c1,c2)	c1=2 and c2=2

TABLE. II. Lower and upper Bound Values of Optimum Variables

Optimum variables	Lower bound	Upper bound
Weights for atmospheric light estimation (W_d)	0	0.5
Aerial Perspective(ω)	0	0.99
Lower bound of Transmission(t0)	0.0001	1

B. Qualitative Evaluation

The proposed dehazing algorithm is implemented using MATLAB 2018 version and tested with the other dehazing algorithm developed by Tarel [10], He [9], Meng [21], and Sulami [22] using the source code provided by the authors. The natural outdoor hazy images are taken from the database [23], [24] to validate our performance. Fig. 5 and 7 are highlighted to show the outperformance of our method. The boxes highlighted are drawn to show that the haze-free image restored from hazy image shows detail information which wasn't visible in the input hazy image. Similarly, Fig. 6, 8 and 9 also shows the improvement in the visual quality of the image of our method. The colour and the contrast of the restored image are better than the input hazy image.

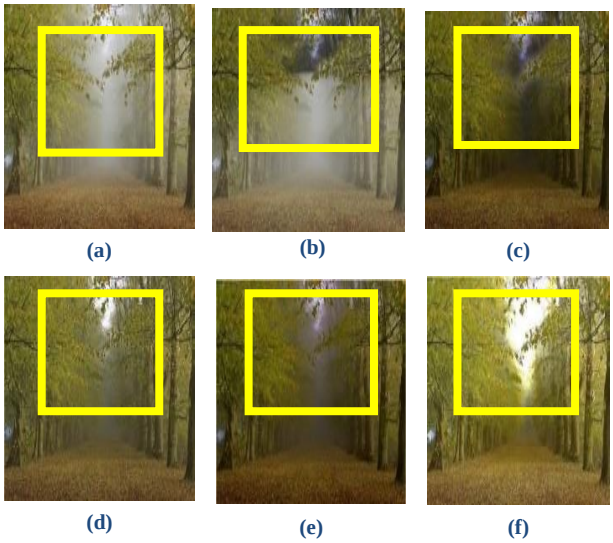


Fig. 5. Visual Comparison of (a) Forest.jpg Input Image; Dehazing Results Obtained by (b) Tarel [10]; (c) He [9]; (d) Meng [21]; (e) Sulami [22]; (f) Our Methodology.

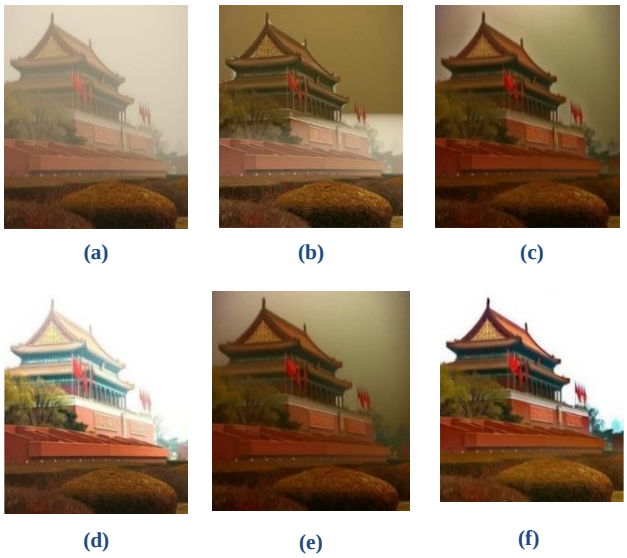


Fig. 6. Visual Comparison of (a)Tiananmen.png Input Image; Dehazing Results Obtained by (b) Tarel [10] (c) He [9]; (d) Meng [21]; (e) Sulami [22]; (f) Our Methodology.

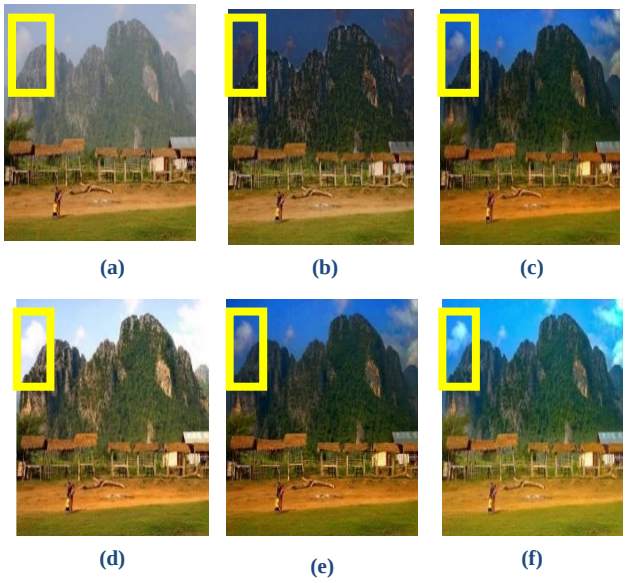


Fig. 7. Visual Comparison of (a) Mountain.png. Input Image; Dehazing Results Obtained by (b) Tarel [10] (c) He [9]; (d) Meng [21]; (e) Sulami [22] ; (f) Our Methodology.

Fig. 8. Visual Comparison of (a) Toys.jpg Input Image; Dehazing Results Obtained by (b) Tarel [10] (c) He [9]; (d) Meng [21]; (e) Sulami [22]; (f) Our Methodology.

Fig. 10 shows the plot with a decrease in fog density with respect to the number of iterations applied for the input hazy image shown in Fig. 3(a). The plot clearly indicates that within three iterations the fog density almost minimizes. We tested around fifty low resolution images and found that for most of the hazy images the fog density decreases between two to ten iterations. We took 20 as the maximum iteration.

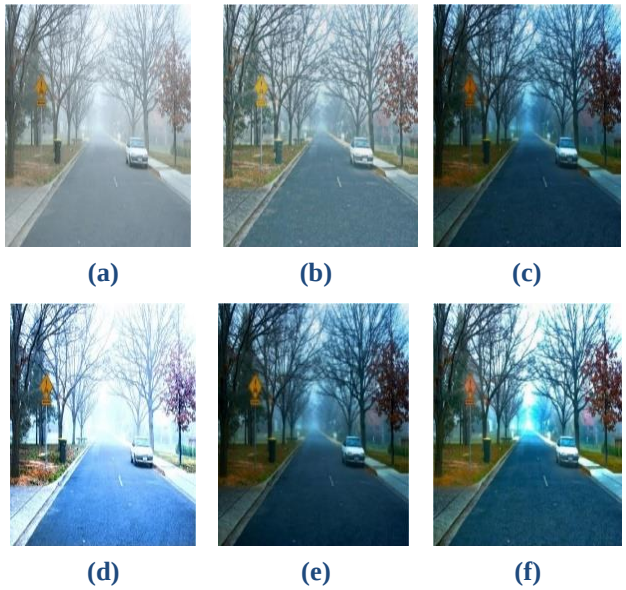


Fig. 9. Visual Comparison of (a) Sweden.jpg Input Image; Dehazing Results Obtained by (b) Tarel [10] (c) He [9]; (d) Meng [21]; (e) Sulami [22]; (f) Our Methodology.

Apart from foggy outdoor natural images, we also executed our proposed method on synthetic images obtained from Foggy Road image database (FRIDA) [24] which has images with different fog density to validate our proposed method with the state-of-the-art. Fig. 11(a) shows the fog-free synthetic images in the first column, which will be later used as ground truth images for quantitative evaluation.

Fig. 11(b) shows foggy images with different fog density added to fog-free synthetic images. The first row (second column) foggy image as a fog density of 1.5829. Second row (second column) with a fog density of 1.1607, the third row (second column) with a fog density of 1.5308, fourth row (second column) with a fog density of 1.4024 and fifth row (second column) with a fog density of 1.7072, respectively. Fig. 11(c) to Fig. 11(g) shows output restored haze-free images obtained by different dehazing technique implemented by various researchers are executed to differentiate our work on synthetic images. The hazy image will have low contrast and the colour will be faded. Our proposed method restores both colour and contrast in natural outdoor images and synthetic images. The comparison of various dehazing algorithms with our proposed method for qualitative analysis shows that our methodology is visually compelling. Since ground truth images are available in the FRIDA database we are able to perform quantitative analysis for the synthetic images by calculating Peak signal to noise ratio (PSNR) and structural similarity index (SSIM). The detail quantitative analysis or objective evaluation for synthetic images is further discussed in the next quantitative evaluation section.

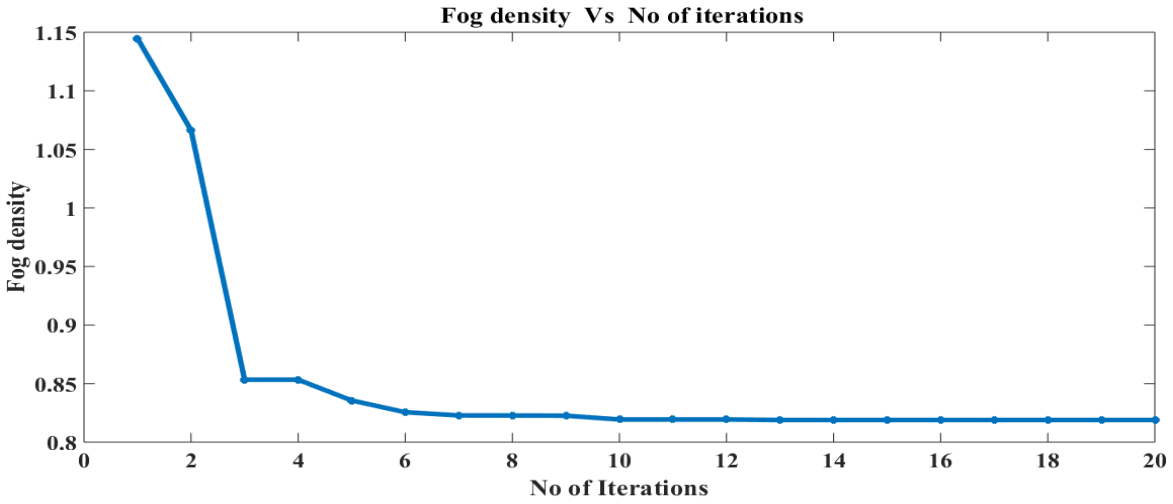


Fig. 10. Plot Indicating the Decrease in Fog Density with Respect to the Number of Iterations.

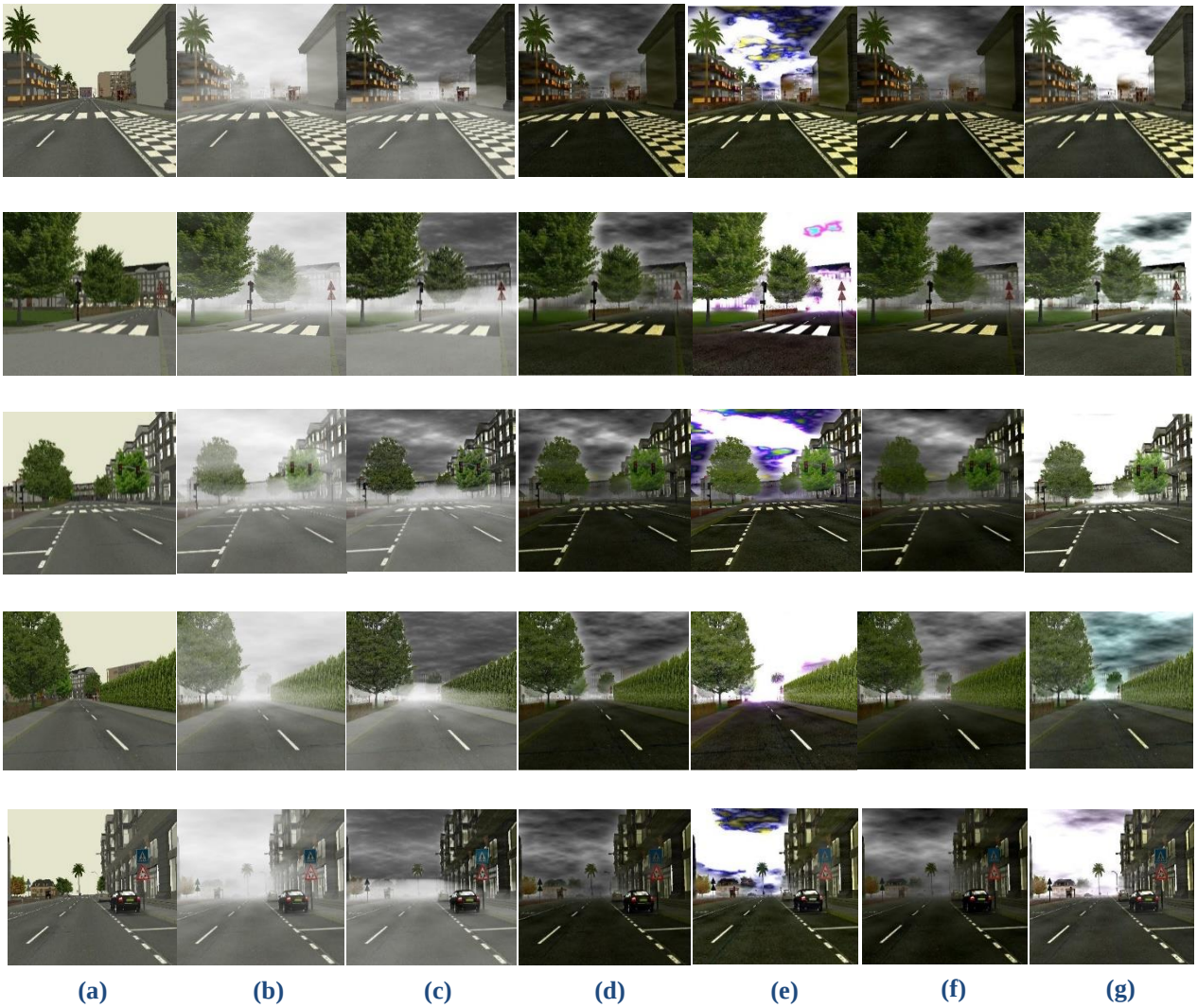


Fig. 11. Column Wise (a) Clear Synthetic Image (b) Foggy Image with different Fog Density. Results Obtained after Dehazing (c) Tarel [10] (d) He [9] (e) Meng [21] (f) Sulami [22] (g) Our Proposed Method.

C. Quantitative Evaluation

For quantitative evaluation, we calculated the fog density of the input image and output restored images using equation (13) proposed by Choi [20]. Using visibility measurement of Hautiere et al. [25], we evaluated three parameters e, r and Σ.

The objective or quantitative evaluation of our proposed method are summarized in Tables III, IV, V and VI for the images in Fig 5, 6, 7 8 and 9. From Table III, we can conclude that the fog density measured from restored image using FADE evaluator tool shows that the average fog density is comparatively less than the average fog density obtained by various dehazing techniques.

Similarly Tables IV and V shows that the average rate of visible edges 'e' and average ratio of mean gradient 'r_g' obtained are comparatively higher than the values obtained from various techniques implemented by researchers who worked in this field to remove haze from images. The saturated pixels from Table VI shows almost zero for all the figures indicating that there are no saturated pixels in restored image.

TABLE. III. FOG DENSITY EVALUATED USING FADE

Images	**Fog density (D)**					
	Input image	***Tarel [10]***	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Fig 5	0.389	0.365	0.190	0.182	0.222	0.203
Fig 6	1.345	0.529	0.383	0.497	0.434	0.369
Fig 7	0.583	0.242	0.232	0.328	0.242	0.218
Fig 8	1.229	0.398	0.286	0.338	0.258	0.253
Fig 9	0.963	0.378	0.215	0.473	0.224	0.220
Avg	**0.902**	**0.382**	**0.261**	**0.364**	**0.276**	**0.252**

TABLE. IV. RATE OF VISIBLE EDGES

Images	**Rate of visible edges 'e'**				
	Tarel [10]	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Fig 5	0.10895	0.32917	0.27279	0.2541	0.3504
Fig 6	0.38368	0.4532	0.47356	0.47356	0.5246
Fig 7	0.16431	0.15462	0.0911	0.11851	0.0542
Fig 8	2.0598	1.9668	1.9097	2.15886	2.1664
Fig 9	0.52933	0.58948	0.26659	0.5638	0.6429
Avg	**0.64921**	**0.698654**	**0.602748**	**0.713766**	**0.7477**

TABLE. V. RATIO OF MEAN GRADIENT

Images	**Ratio of mean gradient 'r_g'**				
	Tarel [10]	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Fig 5	2.1643	1.6439	2.3099	1.7616	2.5230
Fig 6	4.1775	3.2471	4.9593	3.2175	5.5012
Fig 7	1.3866	1.1098	2.6645	1.7922	2.1715
Fig 8	3.1120	1.6189	2.6765	1.7722	3.2172
Fig 9	2.3188	1.6494	2.1667	1.4151	2.4710
Avg	**2.63184**	**1.85382**	**2.95538**	**1.99172**	**3.17678**

TABLE. VI. PERCENTAGE OF SATURATED PIXELS 'Σ'

Images	**Percentage of saturated pixels 'Σ'**				
	Tarel [10]	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Fig 5	0.00	0.00	0.00	0.00	0.00
Fig 6	0.00	0.00	0.00	0.00	0.00
Fig 7	0.00	0.00	0.0005	0.00	0.00
Fig 8	0.00	0.00	0.00	0.0047	0.00
Fig 9	0.00	0.00041	0.00	0.0051	0.00

From the results shown in Tables III, IV, V, and VI, it is evident that decrease in fog density, as well as more edges obtained in the restored image, shows the quality of the image is better than the already existing methods. Also, the number of saturated pixels are almost zero which indicates that our method is better when compared with enhancement based methods. In the sky region, Meng's and He's outcomes poor image quality and Tarel's outcomes show very clear halos around the edges.

By using the synthetic images from FRIDA database, performance analysis of the different methods have been compared with the proposed method. The two evaluation metrics measured are PSNR and SSIM. The two objective evaluation results are summarized in Tables VII and VIII. From Tables VII and VIII we can confirm that the average PSNR and SSIM are comparatively better in our proposed method for synthetic images.

Comparing the input and the output image, as per our observation our methodology shows a better quality of restored image extracting the information properly at the end of dehazing using a single color images.

TABLE. VII. PSNR OBTAINED FROM SYNTHETIC IMAGES

Images from Fig 11	**PSNR (Peak signal to noise ratio) in dB**				
	Tarel [10]	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Row 1	9.620	11.526	10.778	10.065	15.137
Row 2	10.276	10.280	10.423	10.708	13.096
Row 3	9.961	11.146	12.220	10.098	13.882
Row 4	10.17	10.495	12.005	10.816	12.393
Row 5	10.17	10.949	12.454	9.603	16.021
Avg	**10.041**	**10.879**	**11.576**	**10.258**	**14.106**

TABLE. VIII. SSIM OBTAINED FROM SYNTHETIC IMAGES

Images from Fig 11	**SSIM (Structural Similarity Index Metric)**				
	Tarel [10]	***He [9]***	***Meng [21]***	***Sulami [22]***	***Proposed Method***
Row 1	0.345	0.419	0.322	0.453	0.638
Row 2	0.429	0.416	0.337	0.425	0.488
Row 3	0.469	0.384	0.440	0.356	0.486
Row 4	0.394	0.369	0.364	0.491	0.630
Row 5	0.624	0.371	0.415	0.36	0.614
Avg	**0.452**	**0.392**	**0.376**	**0.417**	**0.571**

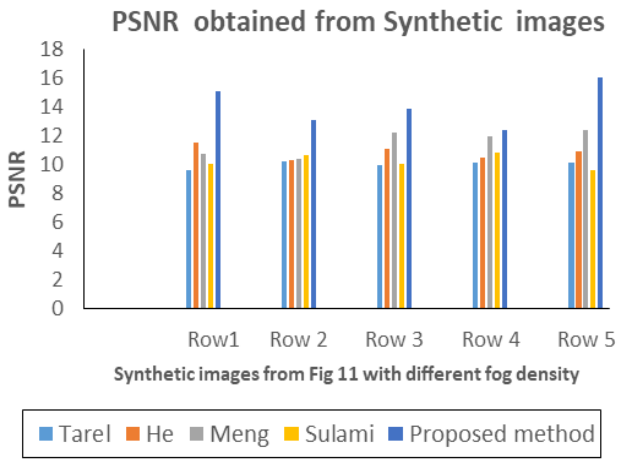


Fig. 12. PSNR Obtained from Synthetic Images.

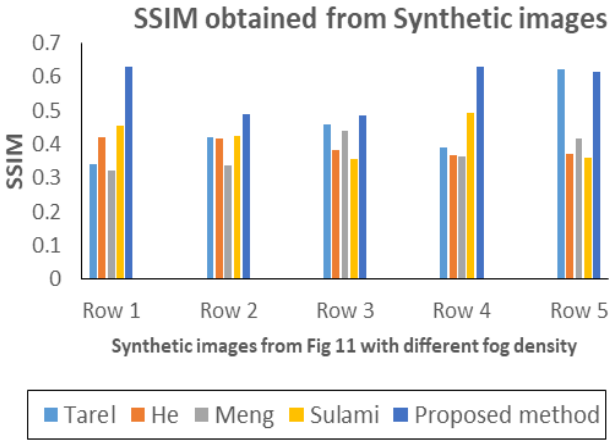


Fig. 13. SSIM Obtained from Synthetic Images.

The plots shown in Fig. 12 and 13 will give a clear interpretation of the results obtained by comparing the PSNR and SSIM of our proposed method with different methods.

V. Conclusion

This research demonstrates the atmospheric light estimation using Particle Swarm Optimization technique for restoring the haze-free image. This technique can exhaustively allow the search of a specific optimized value that can be used for effectively improving the visibility of the hazy regions. By choosing optimal weights for estimating the atmospheric light, haze can be removed from images more effectively. Also, the proposed method does not depend on salient objects in an image like sky regions to estimate the atmospheric light to remove haze thereby producing a perceptual quality image.

The proposed method showed variation in the results and the resultant image quality was perfectly suitable for video surveillance applications. PSNR, SSIM, rate of visible edges and restoration quality have been analyzed and the results have been compared with different methods. Since PSO is an iterative based method, the computational time is more and hence it is not suitable for applications like autonomous navigation. In future, the computational time can be reduced by implementing this technique in hardware like Field Programmable Gate Array (FPGA). However the proposed method works best for removing haze when compared to dense foggy images. Further there is a scope to develop defogging algorithms that can remove fog from images. Overall, the results obtained from the proposed method using PSO shows better results qualitatively and quantitatively.

References

[1] Nayar, S. K., & Narasimhan, S. G. “Vision in bad weather”. In Proceedings of the Seventh IEEE International Conference on Computer Vision (Vol. 2, pp. 820-827). IEEE.1999.

[2] Narasimhan, S. G., & Nayar, S. K. “Contrast restoration of weather degraded images”. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 713-724, 2003.

[3] Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., & Lischinski, D. “Deep photo: Model-based photograph enhancement and viewing” (Vol. 27, No. 5, p. 116). ACM.,2008.

[4] Oakley, J. P., & Bu, H. “Correction of simple contrast loss in color images”. IEEE Transactions on Image Processing, 16(2), 511-522, 2007.

[5] Tripathi, A. K., & Mukhopadhyay, S. “Removal of fog from images”: A review. IETE Technical Review, 29(2), 148-156,2012.

[6] Kim, D., Jeon, C., Kang, B., & Ko, H. “Enhancement of image degraded by fog using cost function based on human visual model”. In 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (pp. 64-67). IEEE.2008.

[7] Tan, R. T. “ Visibility in bad weather from a single image”. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE, 2008.

[8] Fattal, R. “Single image dehazing”. ACM transactions on graphics (TOG), 27(3), 72, 2008.

[9] He, K., Sun, J., & Tang, X. “Single image haze removal using dark channel prior”. IEEE transactions on pattern analysis and machine intelligence, 33(12), 2341-2353, 2011.

[10] Tarel, J. P., & Hautiere, N. “Fast visibility restoration from a single color or gray level image”. In 2009 IEEE 12th International Conference on Computer Vision (pp. 2201-2208). IEEE,2009.

[11] Zhu, Qingsong, Jiaming Mai, and Ling Shao. "A fast single image haze removal algorithm using color attenuation prior." IEEE transactions on image processing 24.11 3522-3533,2015.

[12] Wang, Wencheng, et al. "Dehazing for images with large sky region." Neurocomputing 238: 365-376,2017.

[13] Li, Yunan, et al. "Single image haze removal based on haze physical characteristics and adaptive sky region detection." Neurocomputing 182: 221-234,2016.

[14] Guo, Fan, Hui Peng, and Jin Tang. "Genetic algorithm-based parameter selection approach to single image defogging." Information Processing Letters 116.10: 595-602, 2016.

[15] Zhang, Wenbo, and Xiaorong Hou. "Estimation algorithm of atmospheric light based on ant colony optimization." Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence. ACM,2017.

[16] Singh, Dilbag, and Vijay Kumar. "A Comprehensive Review of Computational Dehazing Techniques." Archives of Computational Methods in Engineering 1-19.2018.

[17] Eberhart, Russell, and James Kennedy. "A new optimizer using particle swarm theory." MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. IEEE,1995.

[18] Shiau, Y. H., Chen, P. Y., Yang, H. Y., Chen, C. H., & Wang, S. S. “Weighted haze removal method with halo prevention”. Journal of Visual Communication and Image Representation, 25(2), 445-453, 2014.

[19] He, Kaiming, Jian Sun, and Xiaoou Tang. "Guided image filtering." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[20] Choi, L. K., You, J., & Bovik, A. C. “Referenceless prediction of perceptual fog density and perceptual image defogging”. IEEE Transactions on Image Processing, 24(11), 3888-3901, 2015.

[21] Meng, G., Wang, Y., Duan, J., Xiang, S., & Pan, C. “Efficient image dehazing with boundary constraint and contextual regularization”. In Proceedings of the IEEE international conference on computer vision (pp. 617-624), 2013.

[22] Sulami, M., Glatzer, I., Fattal, R., & Werman, M. “Automatic recovery of the atmospheric light in hazy images”. In 2014 IEEE International Conference on Computational Photography (ICCP) (pp. 1-11). IEEE,2014.

[23] Ma, Kede, Wentao Liu, and Zhou Wang. "Perceptual evaluation of single image dehazing algorithms." 2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015.

[24] Tarel, Jean-Philippe, et al. "Improved visibility of road scene images under heterogeneous fog." 2010 IEEE Intelligent Vehicles Symposium. IEEE, 2010.

[25] Hautiere, Nicolas, et al. "Blind contrast enhancement assessment by gradient ratioing at visible edges." Image Analysis & Stereology 27.2: 87-95, 2008.

Intelligent Pedagogical Model with Kinesthetic-Static Immersion based on the Neuro-Linguistic Programming Approach (NLP)

Simón Choquehuayta Palomino[1], José Herrera Quispe[2], Luis Alfaro[3], Blas Choquehuayta Llamoca[4]
Universidad Nacional de San Agustin de Arequipa, Peru

***Abstract*—In this paper, the authors propose a teaching/learning pedagogical model, based on an approach that uses neurolinguistic programming, educational data mining and haptic interaction. It also uses the theory of learning styles, which are identified with data mining techniques, clustering and the Farthest First algorithm, as well as a test of Neurolinguistic Programming. Depending on the results obtained, the teaching/learning strategies are defined, and the activities of an educational coaching are suggested, with the purpose of boosting the students' attention in the classroom, stimulating their communicative and psychomotor skills. The proposal was evaluated with a sample of students of regular basic education, to whom an instrument was applied before and after carrying out the tests of applying the plan of the teaching/learning activities. For this purpose, a multifunctional learning kit was constructed, which is a didactic and playful resource, applicable to the student's psychomotor area. The kit contains an application and a hardware device called "Tusuna-pad 1.0", which was implemented in the Unity games engine and was programmed using the C# language. The pedagogical model was validated with the participation of students of Regular Basic Education, considering pedagogical and computational aspects, results that were validated and duly analyzed. Finally, conclusions and recommendations for future work were established.**

***Keywords*—Haptic interaction; viral immersion; learning styles; neuro-linguistic programming; educational data mining**

I. Introduction

Every year, during the learning process, educators encounter students who belong to generations of the digital age, most of whom are familiar with computers and technological devices from society, such as computational supports or devices. For example, for Gramingna [1], games with virtual environments have the following characteristics: a) they allow interaction between the user and the device; b) when playing, users are involved in a simple way; c) they foster interest, making learning posible; d) they allow the trial to be associated with the reward, and e) both hemispheres are activated, promoting new ways of thinking.

Related to the pedagogical aspect, Sanchez [2] states that students learn in many ways: seeing and listening, reflecting and acting, reasoning logically and intuitively, memorizing and visualizing, constructing analogies and playing. Indeed, leisure activities [3] have been the driving force and one of the most effective modes of learning since the appearance of man. Through them it is possible to learn different things, some pleasant, others conflictive, such as experimenting, making mistakes [4] and putting oneself in the place of others, above all by doing and playing [5], so that the motivated student can adapt to the educational context and the learning is flexible enough. These are conditions that the student needs in order to be open-minded and to explore the knowledge of a certain thematic line through a computing environment.

II. Motivation

In the Arequipa region, few research studies attempt to solve educational problems with data mining; and if they do exist, they are focused on higher education [6][7], and there are no studies for EBR, therefore the use of data mining techniques is proposed to detect the predominant learning style of the student [8][9], according to the NLP approach. Depending on the predominant style, teaching strategies and coaching activities should be applied in a playful interactive environment with haptic-kinesthetic immersion based on NLP, understanding that psychomotor communicative skills have their implications in body expression Choquehuayta [10].

Failure to contribute to the strengthening of psychomotor skills in time could have an impact on learning difficulties. In the case of mathematics, if one does not work on body laterality in the school stage between 2 and 5 years of age, the child will not be able to start adding or subtracting several quantities from right to left [11]. It has been proven that the psychomotricity of kinaesthetics increases the reality of the subject and brings some action to the data and meanings [12]; and this helps us notably with the comprehensive assimilation of the information of any content. It should be noted that in studies presented at the 2005 Harvard Conference [13], music video games can improve concentration skills as well as psychomotor communication skills; for example, a recent exploratory study of 1500 high school students showed significant increases in reading fluency and math fluency, as well as significantly improved attention [14]. Based on the rationale, is EAIHK-NLP likely to enable EBR students to strengthen psychomotor communication skills, thereby enhancing attention?

III. State of the Art Review

Juan Olague and Socrates Torres [15] developed the project "Application of data mining techniques and learning content management systems for the development of a computer learning system of computer programming" [16], whose main objective was the description of learning styles in the students of the career of Computer Science and of course

they elaborate a software that would facilitate to learn the subject of computer programming better, once its style has been detected. Gutiérrez and Solórzano [17] cite a phrase from Confucius that says: "I hear and I forget; I see and I remember; I do and I understand"; they consider that this phrase manifests the effectiveness of learning based on experiential learning that is characterized by sustaining that "one learns by doing", and where it is considered that the student should actively participate in his or her own learning process in such a way that his or her capacity to learn to learn is strengthened, understanding one's own way of learning and the processes that are required for it.

The use of data mining in education is encouraged to describe and predict learning styles with the work "identification of dominant learning styles in computer students" [18], and in Durán et.al. [19], present the work "Data Mining to describe Learning styles". In this work they used the instrument created by [20], which consists of 44 questions with two answer options each, in which the student must necessarily choose one of them, and then perform an analysis of the data in Weka using the Clustering technique and the FarthestFirst algorithm, and for the display of data, they used the Torta graph. The second work of the previous paragraph used the instrument of the Felder-Silverman model [21]. This only considers 4 dimensions and not 5 as the first one. Cumbirayco [22], in her work entitled "Neurolinguistic programming in students' strategic learning" at the Universidad César Vallejo, concludes that the applicability of NLP makes great sense, since it is an effective and simple tool that allows a significant change in the student's academic performance.

IV. Object of the Research

To propose a learning pedagogy based on the learning approach of NLP, with the purpose of contributing to the improvement of psychomotor communicative skills, consequently enhancing the attention of EBR students.

V. Topics Involved in the Research

A. Educational Data Mining (EDM)

EDM can transform the raw data collected by learning systems into useful information that can be used to make informed decisions and answer research questions [23]. EDM is an emerging discipline at the intersection of data mining and pedagogy. While pedagogy provides the intrinsic knowledge of the learning process, data mining provides the techniques of information analysis and modeling. In Fig. 1, we present the elements that make up its structure.

B. Neuro-Linguistic Programming

The NLP, through analogy with the computer, uses the universal patterns of communication and perception that we have to recognize and intervene in diverse processes: learning, therapy, other conflicts [25]. Neuro-linguistic programming offers us invaluable tools to give students a personalized, efficient education, seeing them as individuals in order to be able to provide them with the required information in an effective manner.

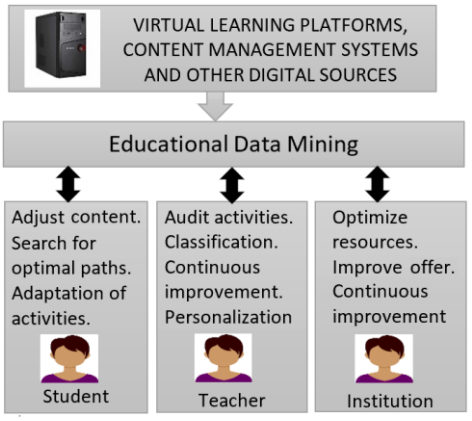


Fig. 1. Elements in Educational Data Mining Source: Ballesteros J, Sánchez D., & García R. [24].

C. Haptic Interaction

Haptic interaction considers both the tactile (touching) and kinesthetic (movement) [26]. By means of kinesthesia the muscular movement, weight and position of the corporal parts are perceived. (C. Colwell, H. Petrie, A. Hardwick and S. Furner, 1998) write about the close link between perception and action in haptic functioning. This means that in order to perceive haptics, the user must move actively in computer interaction [27].

D. Play and its Importance in Education

Play is considered as one of the most enjoyable and significant activities of the human being, which can also be used as a learning strategy for the acquisition of knowledge, both in the classroom and virtual environments [28]. For example, Rocha [29] in his work concluded that the application of play in the educational field considers motivations, interests, and the evolutionary characteristics of learners. Torres [30] also points to play as a learning strategy that leads students to the world of knowledge. Chacón [31] forcefully reveals that play has an educational objective that includes moments of pre-reflective action and of abstract-logical symbolization or appropriation of what has been lived in order to achieve the objectives of curricular teachings.

1) Educational video games: These are interactive multimedia materials through which one or several topics can be learned. An important characteristic is that the knowledge is acquired in an implicit way, that is to say, the players do not realize that while playing they are acquiring a series of concrete skills, but that they are acquiring them in the natural course of the videogame [32].

VI. Methodological Design and Elements Applied in the Proposal

The method used for the development of this work is a quasi-experimental descriptive study [33], with a tendency to action research [34] in information systems, which has become one of the main methods of qualitative research in the field of information systems and software engineering [35].

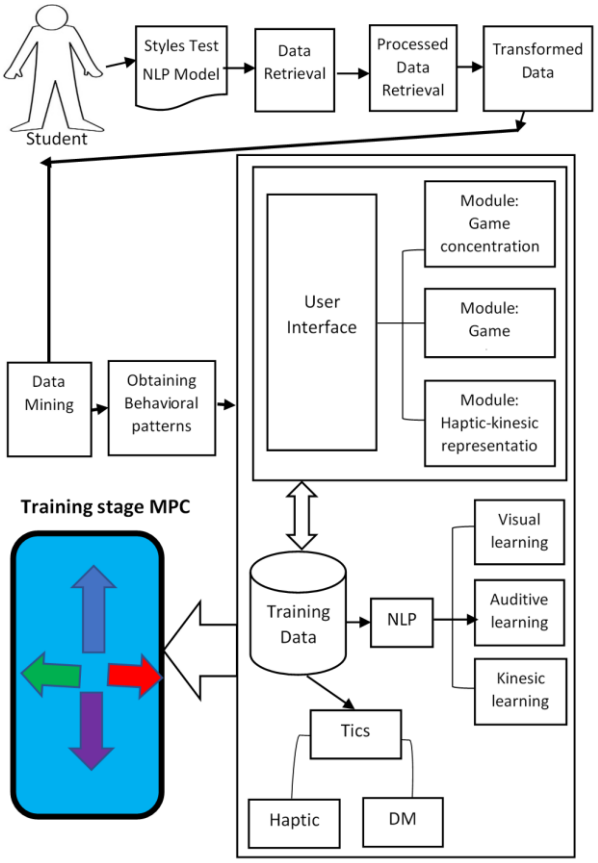


Fig. 2. General Outline of the Proposal Source: Taken from [43].

The DM process was applied for the discovery of knowledge and, based on the results, the computational pedagogical material (MPC) was implemented. Fig. 2 presents the general scheme of the methodology applied. According to Matas and Dalmau [36] it is suggested that the proposals should be schematized.

VII. Procedural Design of the Proposal

From the many processes as shown in Fig. 2, the student is considered first for the DM process and data collection. Applying the information collection instrument, the NLP learning styles test [37], which is made up of 40 questions, each with three alternative solutions, which represent a specific style, the student can only mark one answer. Table X shows a part of the 40 questions of the questionnaire modified in form and content for students of the 5th cycle (5th and 6th grade) of EBR at IE 40134 primary education. For the validation of the proposal we worked with the card shown in Table I, the same that was used for the pre-test and post-test of the experiment. This instrument was applied to 90% (92 students) of the population of 108 students.

With the information provided by the test per student, a database was generated to submit it to the KDD process (Knowledge Discovery in Database) [38]. Of course, the data mining technique "cluster analysis" was applied, based on the common characteristics of a certain number of groups; this type of analysis is used to classify individuals to identify homogeneous subgroups within the sample of respondents. From the process, it was found that the student group under study has a high predominance of the kinesthetic-saptic learning system. As a result, the majority of students interact in virtual environments that are generally related to leisure. The multifunctional kit called "Tusuna-pad 1.0" was implemented, which is an application and device with protocol to the PC. This MPC was used in 6 learning sessions of educational coaching (1 per month); with the intention of strengthening psychomotor communication skills, especially regarding attention. To this effect, the heading of Table I of the development of the instructional plan was applied.

TABLE. I. Heading Applied for Validation

Criteria / Points		Rating levels / Punctuation	Value
Jump with both feet on the mat, located on the floor.	1	It is not driven and does not put both feet at the same time in case there is a left-right arrow	
	2	It flexes the trunk and drives with both legs. It does not fall on the mat.	
	3	It drives and stands on the left-right arrows with both legs, while coordinating the simultaneous extension of arms.	
	4	It is energized by stepping on the circles with arrows as appropriate	
Perform the jump and move the trunk from the body to the compass.	1	Make 90° movement	
	2	Make trunk movement up-down	
	3	Make foot movement with great agility	
	4	Make movement to put the feet and legs in 180° with great agility	
Retains body coordination and balance on the mat	1	Coordinates a fluid movement from the legs and trunk to the arms	
	2	There is flexion and extension of the knee and leg when stepping on the circles of the mat	
	3	It balances on your leg without having much inclination to the ends of the mat.	
	4	The sequence of movement of the feet and legs is rhythmic	
Participation	1	Only one or two people actively participate	
	2	At least half of the students actively participate	
	3	At least 3/4 of the students actively participate	
	4	All students participate enthusiastically	
Interaction quality	1	Very little interaction: very short conversation; some students are distracted or disinterested	
	2	It favors coexistence and contributes to the resolution of conflicts in a peaceful way	
	3	Students show being versed in the interaction; lead lively discussions centered on the game	
	4	Demonstrate attitudes of tolerance and respect for the possibilities and limitations of others in game	
Total			

Source: Own Elaboration

VIII. Organization of the Logical Structure Pertaining to the Multifunctional Learning Kit

Prototype contextualized to our region with peculiarities and terms characteristic of the Andean culture. It is aimed at

children from 8 to 12 years old and is used in innovation and physical education classrooms (psychomotor area).

A. Navigability Scheme of the Multi-Functional Kit

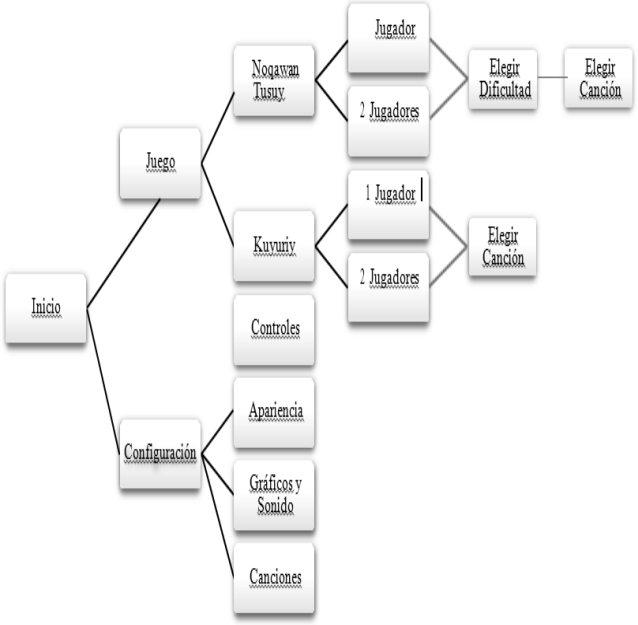


Fig. 3. Navigability Scheme Source:Own Elaboration.

B. Diagrams

In this section we present the user case diagram according to Fig. 4, and the administrator case diagram Fig. 5. Fig. 10 shows the hardware device corresponding to the kit.

C. Interface

It consists of two modules as shown in Fig. 6. The elements of the prototype were implemented in the Unity game engine and its C# coding.

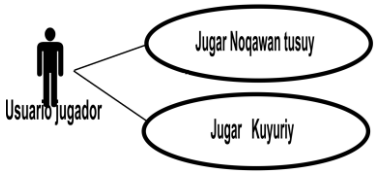


Fig. 4. Usage Case Diagram Source: Own Elaboration.

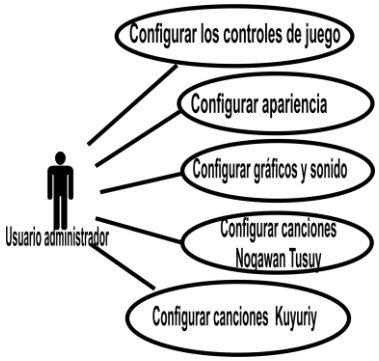


Fig. 5. Usage Case Diagram Source Own Elaboration.

Fig. 6. Youruna-Pad Learning Kit Modules Source: Own elaboration.

Fig. 6 refers to the "Kuyurity" module that results from an innovation of the "Twister" game, which does not have an interface and a firm platform, which does not evaluate the levels of attention of the user player, nor does it have reports of the scores or statistics, elements that were implied in the proposed model. The screen of the "Kuyurity"game is observed in Fig. 7, and then the final game report with the respective ratings and statistics concerning the attitudinal aspect. In Fig. 8, Rapid Rhythm, the level of attention is divided into the qualifications based on Tables VIII and IX.

According to this table the logic of the application is described, with respect to the qualification that the user obtains, when the application is executed. The final percentage is obtained by means of a simple rule of three, where the total quantity of arrows, equivalent to 100%, results from the sum of the perfect points plus the failed points.

The report of Fig. 8, achieved by the user, attention level and rhythm, is based on the percentages and data of Tables VIII and IX. For example percentage calculation: A final report of a player, at the end of the piece of music in the user's environment. You get the total number of arrows: 238 + 85 = 323, then apply a simple rule of three to obtain Percentage

323 …….100%

238…….. X% (1)

X= 73%

According to the equation (1) and Table IX, 73% are in the last range, which is equivalent to a Divided Attention Level and Rapid Rhythm qualitatively qualifying as "Good".

Fig. 7. "Kuyuriy" Module Source: Own Elaboration.

Fig. 8. Results Screen of the "Kuyuriy" Game Source: Own Elaboration.

Fig. 9. Tusuna-Pad Hardware Device Source: Own Elaboration.

Fig. 9, shows a photo of the pad or mat, which connects to a Pc, which of course becomes a dance platform, an improved model of what was presented at TISE 2011 [10] called "Chaqui-pad". In order to work with this MPC, classroom session plans were planned, known as the Pedagogical Practice Proposal (PPP), material presented in Table VI, which allowed for the organization and planning of learning objects and activities, as well as for the evaluation of the student's psychomotor communication skills and abilities.

IX. Results Analysis

A. Processes for Detection of Learning Style based on the NLP Model

a) Selection stage: 108 students were selected from IE 401344

b) Stage of data cleansing and elaboration: For this purpose, exploratory tasks were carried out analyzing each of the cards registered in a table, achieving the elimination of inconsistencies, through manual corrections, such as the existence of unmarked alternatives, definition of double marking and other typographical errors, including the elimination of cards such as damaged and blank tests, supposedly generated by some nonconformity to the test.

c) Data mining: Once the Excel file has been saved: (IE40134. XLSx) with the tabulated PNL test data, we found a record of 108 students and 44 fields (columns) to be analyzed quantitatively and qualitatively by means of the Weka tool. Fig. 10 reports stored data ready to be exported to CVS format (format that allows to separate fields separated by commas in Excel) to load it into a text editor; in this case, it was the Editplus application; it could also be loaded into Wordpad but this lacks other functions that the case requires. Fig. 11 presents the file in Editplus.

	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	VISUAL	AUDITIVE	KINESIC	RESULT
2	b	c	a	c	c	c	a	c	c	b	c	a	a	a	c	a	9	11	20	KINESIC
3	a	c	a	b	c	c	c	c	c	b	c	a	c	a	c	c	8	11	10	KINESIC
4	a	c	c	c	c	c	c	c	c	c	b	c	b	b	b	c	10	5	25	KINESIC
5	a	b	b	c	a	c	a	c	c	c	a	c	a	c	b	b	10	16	14	AUDITIVE
6	a	b	a	b	b	c	c	c	c	c	b	a	a	c	b	b	15	11	14	VISUAL
7	c	a	c	b	a	c	c	a	a	a	a	a	c	a	b	b	10	11	19	KINESIC
8	c	c	b	c	b	b	c	c	c	c	b	a	a	c	a	a	9	5	26	KINESIC
9	b	a	b	a	a	a	c	c	c	b	a	c	c	c	c	c	11	15	14	AUDITIVE
10	b	a	c	b	b	b	c	b	b	a	c	a	b	b	b	a	14	13	13	VISUAL
11	c	b	b	c	b	a	c	b	b	a	c	c	a	c	c	c	14	10	16	KINESIC
12	b	a	b	b	c	a	a	c	c	b	a	c	a	b	b	a	17	10	13	VISUAL
13	c	b	b	c	a	c	b	c	c	c	b	c	b	a	c	a	12	10	18	KINESIC

Fig. 10. Tabulated Test Data in Excel Source: Own Elaboration.

```
1   @Relation encuesta
2
3   @attribute P1 {a,b,c}
4   @attribute P2 {a,b,c}
5   @attribute P3 {a,b,c}
6
7   @attribute P38 {a,b,c}
8   @attribute P39 {a,b,c}
9   @attribute P40 {a,b,c}
10  @attribute VISUAL numeric
11  @attribute AUDITIVE numeric
12  @attribute KINESIC numeric
13  @attribute RESULT {VISUAL, AUDITIVE, KINESIC}
14
15  @data
16  a,c,b,c,a,c,b,a,c,b,a,b,c,c,b,a,c,c,b,c,c,b,c,ac,c,b,c,a,a,a,c,a,9,11,20, KINESIC
17  a,c,a,a,c,b,b,c,c,b,c,c,a,c,b,c,a,c,b,c,c,a,c,b,c,c,a,c,b,c,c,s,b,c,c,b,a,c,b,c,a,8,11,21, KINESIC
18  C,c,c,c,b,c,b,c,b,c,c,a,c,a,c,b,c,b,c,c,a,c,b,c,a,c,b,c,c,c,b,c,c,b,c,b,10,5,25, KINÉSIC
19  c,c,a,a,c,b,a,b,b,c,a,a,c,b,b,c,c,bc,a,a,a,c,a,a,a,b,a,c,a,c,a,c,b,b,c,b,10,16,14,KINÉSIC
20  a,c,c,b,a,b,a,b,c,b,b,c,b,b,c,a,a,a,c,a,,b,c,a,b,c,b,b,c,c,c,c,b,a,a,c,b,b,15,11,14VISUAL
21  a,c,c,c,a,c,a,c,c,b,c,a,c,c,b,b,c,c,b,b,c,c,c,a,b,b,a,c,c,a,b,a,a,a,c,a,b,b10,11,19,KINÉSIC
22  c,a,c,b,c,b,c,c,c,c,b,c,c,c,c,c,c,c,c,c,c,c,c,b,b,c,c,b,b,b,b,c,c,c,c,b,a,a,c,a,a,9,5,26,KINÉSIC
23  c,a,c,a,c,b,c,a,b,a,a,c,b,b,c,b,c,b,a,b,a,a,a,a,b,a,b,a,a,a,,,c,cb,b,a,c,c,c,c,c,c,11,15,14,AUDITIVE
24  a,c,b,a,b,c,a,a,c,c,b,c,c,c,a,c,b,b,a,a,b,a,c,b,b,b,c,,c,b,a,a,c,b,a,14,13,13,VISUAL
25  c,c,b,c,a,c,a,a,b,c,a,b,c,c,a,a,b,b,c,c,b,b,b,c,b,b,c,b,b,c,b,a,c,b,c,c,14,10,16,KINESICO
```

Fig. 11. File in CVS Format to Generate Arff. File Source: Own Elaboration.

A. Results of the DM process

The results were guided by the five questions posed by Han and Kamber [39]; Fig. 12 shows the data clustering, which is described as follows

- Which dataset was considered relevant to discovering the knowledge?

108 students in the 5th cycle as a result of preprocessing.

- What kind of knowledge is to be discovered??

The type of knowledge to be discovered is the des criptive type. Specifically, cluster analysis was applied (Witten and Frank, 1999) to identify homogeneous subgroups within the surveyed student population. For this, the DM tool was used, and the Cluster option was selected. [40] In particular, the FarthestFirst algorithm was chosen because it is a problem of k centres where the maximum distance between a tuple and its centroid is intended to be minimum [41]. In order to detect the dominant learning styles, it was decided to keep the number of clusters (0.1) to be generated in two. As can be seen in the results, there is no need to expand this number given the high percentage of tuples included in the first cluster.

- What measures can be used to estimate patterns of interest?

From this knowledge it was decided to replace each of the attributes included in the two centroids of the clusters cast by Weka with the code of the learning style with which it is linked (Table IV), with which the centroids were expressed as.

```
Correctly Classified Instances    53          49.0741  %
Incorrectly Classified            55          50.9259  %
Instances
Kappa statistic                   0
Mean absolute error               0.419
Root mean squared error           0.4573
Relative absolute error           100     *
Root relative squared error       100     %
Total Number of Instances         108
    =====Detailed Accuracy By Class
    TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
    0          0         0         0         0         0.5    VISUAL
    0          0         0         0         0         0.5    AUDITIVE
    1          1         0.491     1         0.658     0.5    KINESIC
  Weighted Avg.  0.491    0.241     0.491     0.323     0.5

    === Confusion Matrix===
       A b c   <---classified as
      0 0 31 I a = VISUAL
      0 0 24 I b = AUDITIVE
     0 0 53 I C = KINESIC
```

Fig. 12. Determiation of the Dominant Kinesthetic Style Sourde: Own Elaboration.

Cluster Centroid 0:

k, v, a, a, a, k, k, k, a, k, a, k, k, k, a, k, v, v, k, a, a, v, k, a, k, a, k, v, k, v, v, v, k, v, a, v, a, k, k, v

Cluster Centroid 1:

a, v, k, v, v, v, k, k, k, a, k, a, v, v, k, a, v, k, v, a, k, v, a, k, v, v, v, k, a, a, a, k, v, v, v, k, k, k, v, k

In order to determine the combination of learning styles, represented by each centroid, the number of occurrences of each style was counted. Result:

Cluster Centroid 0: V = 11 A = 12 K = 17

This centroid represents the most dominant kinesthetic style of the group, followed by auditory and visual.

Cluster Centroid 1: V = 16 A = 9 K = 15

So this centroid represents Visual, kinesthetic and auditory style (see Table II).

- What background knowledge will be useful?

Based on the characteristics defined by Richard Blander, the correspondences between the learning styles and the possible answers to the forty questions of the NLP model test were determined. The information in Fig. 13 can easily be expressed as a rule-based hierarchy in order to represent the hierarchies of concepts obtained from our knowledge of the domain.

- Visualization of discovered patterns

Fig. 14 shows the patterns discovered in a bar graph, the resulting clusters as a function of the percentage of tuples included in each cluster. 52% of the population under study constituted the first cluster and 48% the second cluster.

TABLE. II. LEARNING STYLES AND THEIR CORRESPONDING REPLACED CODE

Learning styles	Code
Visual	V →b
Auditive	A →a
Kinesthetic	K →c

Source: Own source

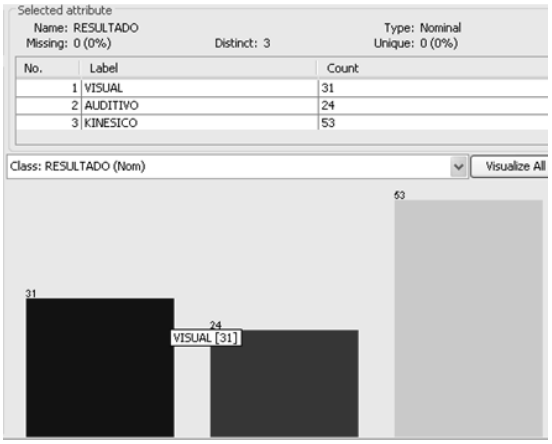


Fig. 13. Result of Kinesic Channel Predominance Source: Own Elaboration.

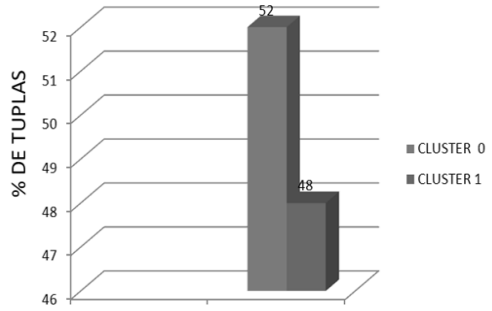


Fig. 14. Visualization of Clustering Results Source: Own Elaboration.

B. Results

As can be seen in Table III, the dominant style is kinesthetic with 49% (53 students), followed by visual 29% (31 students and auditory 22% (24 students), as well as the grouping of the different learning styles present in Fig. 15.

When it comes to evaluating an educational application, it is suggested to evaluate under two dimensions or criteria: computational and pedagogical.

TABLE. III. NUMERICAL AND PERCENTAGE RESULT OF THE KDD PROCESS

V Cycle Students	Applied Tests	Learning Styles – NLP Model		
		Visual	Auditive	Kinesic
108	108 (100)	31 (29%)	24 (22%)	53 (49%)

Source: Own elaboration

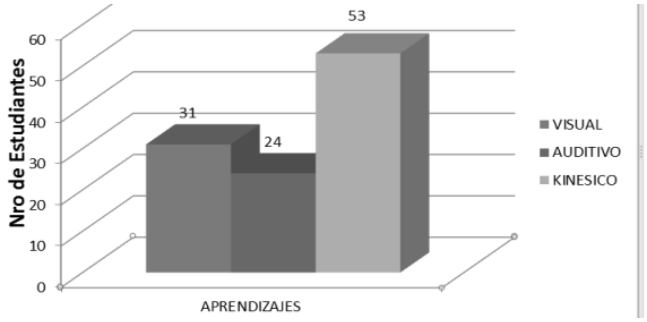


Fig. 15. Students Distributed by Learning Styles Source: Own Elaboration.

Regarding the evaluation of computational design, undergraduate students of the professional career of Educational Sciences specialising in educational informatics (fifth year) of the National University of San Agustín of Arequipa-Peru took part, using the card shown in Table V. Once the answers were tabulated, the qualification of very good (75.0%), good (17.6%) and regular (7.4%) was obtained, with the qualitative evaluation being very significant. What has been described is represented in Table IV and Fig. 16.

The evaluation under the pedagogical dimension was validated using the parameters of the PPP (classroom session) Table VI and Table I containing the items. The result of the application of these two instruments that formed the experimental control groups of the pre-test and post-test of the variable, strengthening of psychomotor communication skills, is illustrated in Table VII and Fig. 17.

When analyzing the results of the variable "psychomotor communication skills" - and in particular attention - it is observed that in the pre-test, the level is placed in: bad, with 51.1%, followed by regular, with 48.9%. While in the post-test the level rises to good, with 78.3% of students, followed by regular, with 21.7%.

TABLE. IV. QUALITATIVE EVLUATION

Criteria	Frequency	Percentage
Very good	81	75,0
Good	19	17,6
Regular	8	7,4
Bad	0	0,0
	108	100,0

Source: Own elaboration

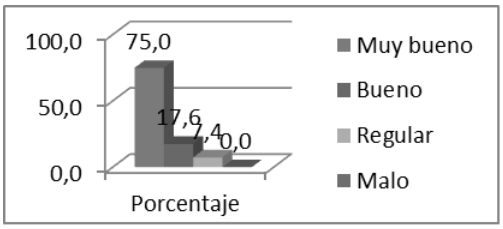


Fig. 16. Qualitative Evaluation Source Own Elaboration.

TABLE. V. EVALUATION SHEET, COMPUTER DESIGN

Criteria	Weight (1-2)
The application is easy to use.	
It is possible to identify the links clearly	
Sound quality and relevance	
Quality and relevance of graphics and images	
The interface design is appropriate	
The interface is pleasant	
It returns easily to the main menú	
Access to submenus is easy	
Promotes hardware or other equipment necessary for its use	
Color scheme	

Source: Own elaboration

TABLE. VI. PPP, PART 1. (CLASS SESSION)

Development of psychomotor activities using new technologies reviewing different topological and spatial concepts

A. Name of the proposal for pedagogical practice

Review of topological and special concepts using new technologies

B. Activity name

We dance an andean song on a rug called "Tusuna_pad 1.0"

C. Purpose

Strengthen psychomotor communication skills and abilities accordingly contribute to the concentration in V cycle students

D. (Learning that students are expected to achieve)

- Perform Physical Activities to Develop Psychomotor Capabilities Through the Application of the "Tusuna Pad 1.0" Application - Motor and Perceptual Expression - Development of Your Physical Qualities - Stimulate Motor Capabilities and Playful Skills - Rhythmic Discrimination And Reaction - Psychomotor Control - Organic Motor Expression

E. (Learning conditions)

- Organization and formation of work teams - Normal physical attitude - Respect the agreements and rules of the game - Disposition of educational materials to use (Pcs, MPC Kit Tusuna -pad 1.0) - Create a climate of trustUse the poster: What do we learn today?

F. (Sequence of actions (students and teacher))

What will we do?	What will we need?	How we will do?	When will we do them?
Teacher: - Apply a survey on LS. - Prepare and install recreational application. Student: - Have good discipline and enthusiastic attitude	- Pcs from 80586 with USB port. - MPC kit "tusuna-pad 1.0 - PPP (learning session	- Read instructions on the activity "Concentration Game" on the mat with user interface - Participating one by one in the rug Tusuna-pad 1.0. - If you have several kits groups of 4 in 4	- In the classroom - In physical education time.

Source: Own elaboration

TABLE. VII. PSYCHOMOTOR COMMUNICATION SKILLS

	Pre-test		Post-test	
	fi	%	fi	%
Bad	47	51.1%	0	0.0%
Regular	45	48.9%	20	21.7%
Good	0	0.0%	72	78.3%
Total	92	100%	92	100%

Source: Own elaboration

That is to say, the students of the I.E. 40134 "Mandil Azul" of the district of Mariano Melgar belonging to the V Cycle of primary education, present a clear improvement when raising their level from bad to good of the pre to the post test, in the variable "psychomotor communicative skills, in particular the attention".

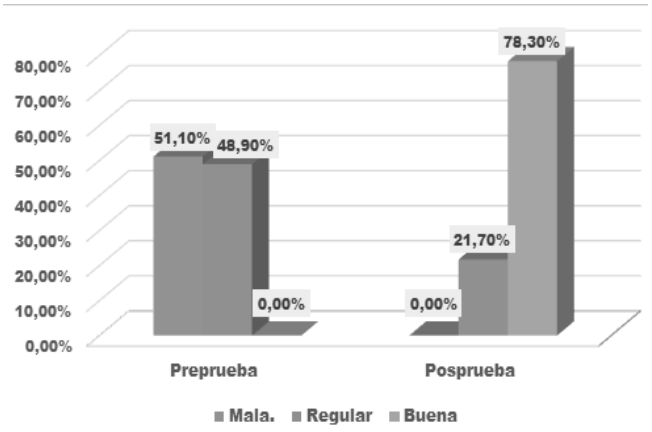


Fig. 17. Variable Representation of Psychomotor Communication Skills Source: Own Elaboration.

TABLE. VIII. LEVEL OF ATTENTION EVALUATION MATRIX

Level of attention	Conceptualization	Rhythm of the game
Focused attention	Ability to respond separately to tactile, audiovisual and auditory stimuli. Time is not measured.	Relaxed
Sustained attention	It consists of maintaining a certain response for a long period of time. It is the ability to keep focused attention over time	Sequential
Alternate attention	Ability to change the focus of attention between cue tasks require different comitive responses. That is, it allows us to move from one task to another. without sharing them because they have nothing in common and start one after finishing it without confusing them.	Slow
Selective attention	Ability to remove the effect of distractions that are unimportant to the task while keeping the focus on the stimulus truly relevant. In other words, to eliminate and dispense with what is neither useful nor necessary.	Hectic
Divided attention	Ability to attend to different stimuli at the same time without making mistakes in execution, temple driving and being able to speak with the driver.	Quick

Source: Own elaboration

Learning style test (NLP model) - Part 1

Instructions: Choose an option with which you identify most of each of the questions and mark it with an X. There are no right or wrong answers. The important thing is that you be honest with your answers.

TABLE. IX. PERCENTAGE DISTRIBUTION IN RELATION TO LEVEL OF CARE

Range of percentage	Level of attention	Rythm
0 % - 5 %	Focalized	Relaxed
6 %- 10 %	Sustained	Sequenced
11 %- 25 %	Alternate	Paused
26 %- 50 %	Selective	Frenetic
51 %-100 %	Divided	Rapid

Source: Own elaboration

TABLE. X. REAGENT SHEET FROM THE NLP TEST QUESTIONNAIRE SOURCE MODIFIED IN CONTENT AND FORM, VERACRUZ [42]

1	Which of the following activities do you enjoy most?	a	Listen to music
		b	Watch movies
		c	Dance with good music
2	What television show do you prefer?	a	World News
		b	Reports of discoveries, places
		c	Comic and entertainment
3	When you talk with another person, you:	a	You watch her
		b	You listen carefully
		c	You tend to touch her
4	If you could purchase one of the following items, which one would you choose?	a	TV
		b	Radio
		c	Bike
5	What do you prefer to do on a Saturday afternoon?	a	Go to the movies
		b	Go to a concert
		c	Stay at home
6	What type of exams are best provided for you?	a	Written exam
		b	Oral exam
		c	Multiple choice exam
7	How do you orientate yourself more easily?	a	By using a map
		b	Asking for directions
		c	Through intuition
8	What do you prefer to spend your time in a resting place?	a	Walking
		b	Thinking
		c	Rest
9	What flatters you the most?	a	You look good
		b	You have an interesting talk
		c	You have a very nice treatment
10	Which of these environments attract you most?	a	A beautiful ocean view
		b	Hear the waves of the sea
		c	A pleasant climate sits
11	How do you facilitate learning something?	a	Writing several times
		b	Repeating out loud
		c	Matching it with something fun
12	What event would you rather attend?	a	Art exhibition
		b	Confenrence
		c	A dance social gathering
13	How do you form an opinion of other people?	a	By its appearance
		b	For the sincerity in his voice
		c	By the way you shake hands
14	How do you consider yourself?	a	Athletic
		b	Intellectual
		c	Sociable
15	¿Qué tipo de películas te gustan más?	a	De acción
		b	Clásicas
		c	De amor
16	How do you prefer to keep in touch with another person?	a	E-mail
		b	Phone
		c	Having coffee together
17	Which of the following phrases identify most with you?	a	It is important that my bike is clean
		b	I perceive even the slightest noise my bike makes
		c	I like my bike to feel good when driving it
18	How do you prefer to spend time with your best friend or friend?	a	Watching something together
		b	Talking
		c	Playing
19	If you can't find the keys in a bag	a	You are looking for it
		b	You shake the bag to hear the noise
		c	You look for the touch
20	When you try to remember something. How do you do it?	a	Images
		b	Sounds
		c	Emotions

X. CONCLUSIONS

The results show that the application of the clustering analysis technique to the students of the EBR V cycle of IE 40134 made it possible to determine the predominance of the kinesthetic learning style with 49%, followed by the visual 29% and auditory 22%, of the whole population under study.

With the results obtained from educational DMs, it was possible to implement an application and hardware device baptized with the name of "Tusuna-pad 1.0", a multifunctional learning kit as a contribution to the new design of computerized pedagogical materials, with emphasis on the practice of the kinesthetic and haptic channel for EBR learners, making the teacher's work easier as facilitator and guide, supported by technological devices.

It was proven that the use of the surroundings allows the student to contribute to the improvement of attention as a result of the strengthening of psychomotor communication skills. Considering the analysis of the results of the variable "psychomotor communication skills", in particular attention, it is observed that in the pre-test the level is situated in bad at 51.1%, followed by regular at 48.9%. While in the post-test the level rises to good with 78.3% of students, followed by regular with 21.7%. That is to say that the students of the IE 40134 of the V primary cycle, present improvement when raising their level from bad to good, from the pre to post test, in the variable mentioned.

XI. FUTURE WORK

To the scientific community we present the following scheme Fig. 18, whose purpose is to design and implement a pedagogical hyper-intelligent / hyper-learning model mediated by intelligent virtual pedagogical agents (APVI) with topics based on the disciplines of Neuroscience, Neuropsychology and NLP, which allows to dose pedagogical strategies, organize contents and intelligent curricula with conceptual, attitudinal and procedural intelligent learning, as the learning by reinforcement was computationally modeled , and knowledge of the behavioral school of Skinner and Watson, which now forms part of the line of artificial intelligence research called Qlearning.

An integrated model with MAS-ALEGRO (Multi-Agent System) uses the approach of the pedagogical agents, with the capacity to provide individualized and collaborative teaching at the same time. MAS-PLANG (New Generation Platform), designed to provide adaptive support. The HABITATPRO, tool designed for content customization. FSLSM (Felder and Silverman Learning Style Model), allows students to categorize according to their ability to process, perceive, receive, organize and understand information. Agent SONIA, works cooperatively with agents, controls navigation. JAFIMA, allows to develop cognitive agents with capacities of perception, communication, mobility, reaction and behavior directed by objectives. There are enough tools to achieve the proposal. If the international scientific community takes the proposal into consideration, an intelligent technological application can be developed to support education. The objective is to raise the level of education in underdeveloped countries. In fact, it is multidisciplinary work.

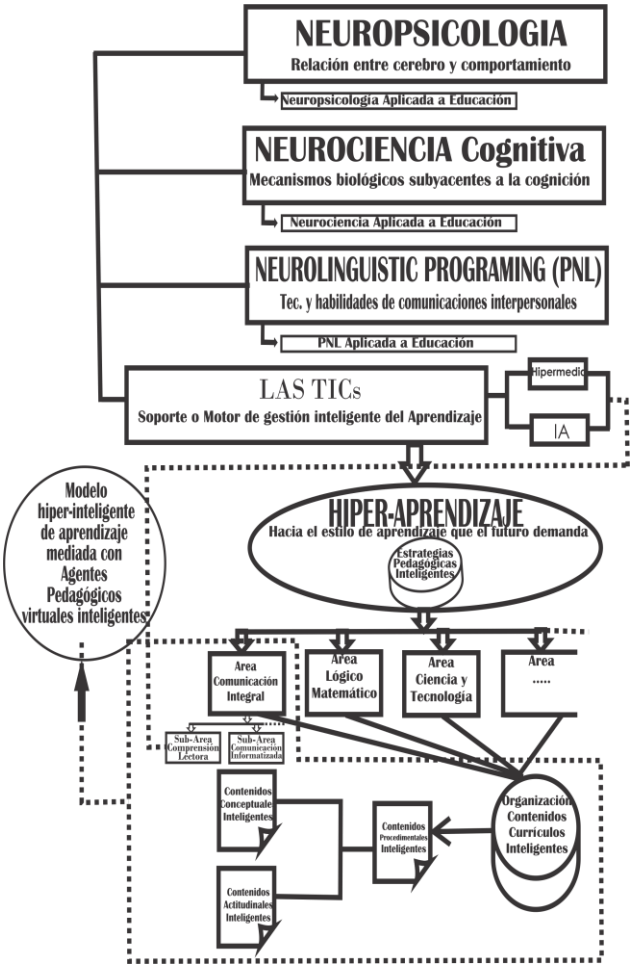


Fig. 18. Architecture of a Multidisciplinary Hyper-Intelligent Learning Model Source Own Elaboration.

REFERENCES

[1] Gramigna, A., & González, J. C. (2009). Videojugando se aprende: renovar la teoría del conocimiento y la Educación.Comunicar,7(33), 157-164. https://doi.org/10.3916/c33-2009-03-007.

[2] Sánchez, J., & Saenz, M. (2006). 3D sound interactive environments for blind children problem solving skills. Behaviour & Information Technology, 25(4), 367-378.

[3] Huizinga, Johan (2012). Homo Ludens. Translation: Eugenio Imaz Echeverría Alianza Editorial. Madrid-España.

[4] Margulis, Lucio y Martínez, Javier (2010). To play is to learn by doing. http://www.catenaria.cl/img/pdf/.

Neruda, Pablo. (1974). I confess that I have lived Valparaíso-Chile.

[5] Kotsiantis, S. Patriarcheas, K. and Xenos, M. (2010). A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students Perfomance in distance Educatión, Knowledge Base Systema, Vol. 23, no 6, pp. 529-535.

[6] Más, J. Alcover, R. Dapena, A. Valdenuten, A. Satorre R. Llopis, F. Rojo, T. Mayo, R. Bermejo, M. Gutierrez, J. Garcia, J. Tovar, E. y, Menasalvas, E. (2009). Academic Performance of Computer Studies in Some Spanish Centers. XV University Teaching Conference on Computer Science, Barcelona, Conference Report.

[7] Gómez del Valle, Manuel (2003). Identification of predominant learning styles in teachers of the Faculty of Education Sciences of the University of Cádiz. ISSN-e 1575-0965, Vol. 6, Nº. 2. Spain: Dialnet.

[8] García, José Sanizo, José y Jiménez, Mercedes. IV world congress of learning styles. Graduate College Learning Styles Magazine, Vol. 10, Núm. 10. Campus Montecillo. México.

[9] Choquehuayta, Simón. (2011). New trends in learning with digital technology. TISE 2011, International Congress of Educational Informatics. University of Chile. Faculty of Computer Science, Physical Sciences and Mathematics. Santiago, Chile.

[10] Cárdenas, Rocío y Herrera, José (2014). Proposal of a methodology for the design of an educational game interface for children from 4 to 5 years with mobile devices. Science and technology magazine ISSN 2304-8891: 2014 Tacna Perú.

[11] Mérida, David. (2002). SHAAD: Adaptive, Adaptable and Dynamic Hypermedia System for content delivery. Research Project, Information Technology PhD Program - Department of Electronics, Computer Science and Automation. Girona, 2002.

[12] Goebel, Jenny (2012). Can 'DDR' and 'ITG' Can Boost Intelligence? The Ultimate Brain Gym. http://www.smart-kit.com/s142/how-ddr-and-itg-can-boost-intelligence-the-ultimate-brain-gym/.

[13] Gómez-Martín M. Gómez-Martín, P. and González-Calero, Pedro A. (2014). Learning based on games Computer Systems and Programming Department Complutense University of Madrid. Year 2 - Vol. 2 - Journal of Communication and New Technologies ISSN: 1697 – 8293. Spain.

[14] Olague Sánchez, Juan Ramón; Torres Ovalle, Sócrates; Morales Rodríguez, Felipe; Valdez Menchaca, Alicia Guadalupe; Silva Avila, Alicia Elena. Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales. Un caso de estudio en el norte de Coahuila Revista Mexicana de Investigación Educativa, vol. 15, núm. 45, abril-julio, 2010, pp. 391-421 Consejo Mexicano de Investigación Educativa, A.C. Distrito Federal, México.

[15] Sánchez, J. y Ovalle, S. (2010). Application of data mining techniques and learning content management systems for the development of a computer learning system for computer programming. X National Congress of Educational Research. Mexico.

[16] Gutiérrez Fernández M. Romero Cuadrado, M. & Solórzano García M. (2011). Experiential learning as a teaching methodology: application of the Macbet method. Argos Magazine, 28(54),127-158.

[17] Costaguta, R y Gola, M. (2006). Identification of dominant learning styles in computer students. In XV Argentine Congress of Computer Science.

[18] Durán, E. y Costaguta (2007). Data mining to discover learning styles. Revista Iberoamericana de Edición, ISSN 1681-5653. http://www.Rieoei.org/deloselectores/1674Duran.pdf.[36] Felder, R Soloman, V. (1984). Index of Learning Styles.http://.ncsu.edu/felder-public.

[19] Richard M. Felder and Barbara A. Soloman (1993). Index of Learning Styles Questionnaire. North Carolina State University.

[20] Felder, R Silverman, L. (1988). Learning and Teaching Styles in Engineering Education Application. Engr. Education, Vol. 78(7), pp.674-681.

[21] Chumirayco, M. (2009). Neurolinguistic programming in the strategic learning of the students of the I Cycle in the professional schools of Initial and Primary Education of the University César Vallejo. Universidad Mayor San Marcos. Lima-Perú. Tesis de Maestría.

[22] Cecily Heiner, Ryan Baker y Kalina Yacef. (2006). Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS. Jhongli, Taiwan.

[23] Ballesteros, A. Sánchez, D. and García R. (2013). Data mining: A tool for researching learning patterns in an educational context. Center for Research in Applied Science and Advanced Technology, Legaria Unit of the National Polytechnic Institute. Calzada Legaria, No. 694, Col. Irrigación. Del. Miguel Hidalgo, C. P. 11500, México D. F. México.

[24] Oconnor Joseph, Seymour John. (2007). Introduction to neurolinguistic programming. Ediciones Urano S.A. Printed in Barcelona-Spain.

[25] Rassmus Kirsten (2008). User-Centered Design of Non-Visual Audio-Haptics. Certec, Rehabilitation Engineering Research. Department of Design Sciences. Faculty of Engineering Lund University. Tesis doctoral.

[26] C. Colwell, H. Petrie, A. Hardwick and S. Furner (1998). Haptic Virtual Reality for Blind Computer Users, ASSETS 98, pp 92-99.

[27] Cortés, F. y Meñacca, I. (2016). Play: a strategy for the teaching-learning in virtual environments of the non-commissioned officers of the infantry training school of Marina de Coveñas Sucre, Colombia. Scientific and technological research article. Palabra Nro. 16, August.

[28] Campo, M. y Gálvez, P. (2006). The game as a pedagogical strategy: A situation of educational interaction. Grade work. Santiago-Chile: Universidad de Chile.

[29] Torres, C. y Torres M. (2007). Play as a learning strategy in the classroom. Venezuela: Universidad de los Andes.

[30] Chacón, P. (2008). The didactic game as a teaching and learning strategy How to create it in the classroom? Revista Nueva Aula Abierta Nro. 16, año 5. Caracas Universidad Pedagógica Experimental Libertad.

[31] Padilla, N.; Collazos, C.; Gutiérrez, F.; Medina, N. (2012). Educational Video Games: Theories and Proposals for Group Learning. Ciencia e Ingeniería Neogranadina 22 (1). ISSN 0124-8170.

[32] Claire Selltiz, Lawrence S. Wrightsman, Stuart W. Cook. (2014). Research methods in social relations. New York: Holt, Rinehart and Winston.

[33] Giorgi, (1978). Participatory action research: a methodological option. Uruguay.

[34] Ruiz, F., Polo, M., Piattini, M. (2002). Use of Research-Action in the Definition of an Environment for the Management of the Software Maintenance Process 1st Workshop in: Research Methods and Philosophical Fundamentals in Software Engineering and Information Systems.

[35] Matas Toni y Dalmau. (2012). Multimedia Products: Design and Conceptual Analysis. Editor of Barcelona Multimedia. directa@ bcnmultimedia.com. España.

[36] Bandler, Richard and John Grinder (1975). The Structure of Magic I: A Book About Language and Therapy. Palo Alto, CA: Science & Behavior Books. ISBN 0831400447.

[37] Frawley, W. Piatetsky-Shapiro, G. and Matheus, C. (1992). Knowledge discovery in databases: An Overview. AI magazine, 13(3), 57.

[38] Han, J. Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan-Kaufmann Academic Press, San Francisco EE.UU.

[39] Núñez, F. Hernandez, R. Tomás, V. y Felipe, A. (2013). Identification of learning styles in university students of computation of the Huasteca Hidalguense by means of Data Mining Techniques. University of the State of Hidalgo. Escuela Superior Huejutla. https://www.uaeh.edu.mx/investigacion/productos/5566/estilos_de_aprendizaje.pdf.

[40] Chandhuri, Siva Garg, Naveen Ravi, R. (1997). The p-neighbor k-center problem. Carnegie Mellon University. Research Showcase@CMU.

[41] Secretary of Education of the State of Veracruz. (2014). http://www.orientacion andujar.es/wp-content/uploads/2014/09/test-estilo-deaprendizajes.pdf.

[42] Choquehuayta, S., Alfaro, L. y Herrera, J. (2017). Learning environment, based on the neurolinguistic programming approach (NLP): Kinese-properceptive learning system. Revista Referencia pedagógica. Año 2017. Número 1, ISSN 2308-3042. http://rrp.cujae.edu.cu/index.php/rrp/article/view/117.

Budgets Balancing Algorithms for the Projects Assignment

Mahdi Jemmali
Department of Computer Science and Information
College of Science in Zulfi, Majmaah University
Majmaah, 11952, Saudi Arabia

***Abstract*—This paper focused on the resolution of the project's assignment problem. Several heuristics have been developed and proposed in this paper to serve as lower bounds to our studied problem. In a developing country, it is interesting to make an equitable distribution of projects in different cities in order to guarantee equality and regional development. Each project is characterized by its budget. The problem is to find an appropriate schedule to assign all projects to all cities. This appropriate schedule seeking the maximization of the budget in the city that having the minimum budget. In this paper, six heuristics were proposed to carry out the objective of resolving the studied problem. The experimental results show that the algorithm given by the heuristic P_6^r outperforms all other heuristics cited in this paper.**

Keywords*—Heuristic; scheduling algorithms; project assignment

I. Introduction

The regional development is a very important thing for each country. Appropriate distribution of resources is primordial to guarantee social and economic stability. The economic cycle may directly be impacted by the development of cities. The principal main of this research is the allocation of the project. The problem is based on the fact that we have several regions and several projects to distribute to the regions. The principal goal of this work is to maximize the budget for the city that having the minimum budget. This goal will be presented as a maximization of the minimum total budget problem. This problem can be solved approximately using several heuristics. Given a set of projects to be distributed to different cities (regions). Each project has its budget and will be entirely assigned to a chosen city. Seeking the balancing of the budgets to all regions, we apply some heuristics to schedule projects on the regions to maximize the minimal total budget [1]. In our work, it is assuming that all regions are identical and have the same economic characteristic, this is meaning all regions can attract all given projects. Heuristics will be used to give a feasible dispatching solution of all projects. The unequal distribution of income across regions is studied in [2] which authors presented in the research work the theory of optimal control.

Other researchers use the operations research field applied in the allocation of investment using the dynamic programming problems with linear form. The regional allocation problem was studied to found the optimal solution proprieties [3].

In our paper we propose to make dispatching of projects on cities by using several heuristics of the follows problem. Minimizing the maximum total completion time for parallel machine $P||C_{min}$ as defined in [1]. For the latter problem, an optimal method using a branch-and-bound algorithm was given in [1]. The developed algorithm used essentially the proposed tight lower and upper bounds.

In [4], the authors developed a method for multi-mode resource with cost availability problems in the project assignment. Other approach is proposed based on differential search (DS) related to the conditions and constraints of the resource project assignment problem [5]. Recently, the author in [6] proposed three approximate Solutions to solve the problem that treated the project revenues but no experimental result is given in the latter work. Some randomized variants were developed as heuristics to solve approximately the problem of the aircraft turbines engines in [7]. Indeed, authors in the latter work show that the solution of the problem can be given by solving the maximization of the minimum completion time. This problem is the same one applied in the studied case in this paper. On other work type, authors in [8], reveal the problem of allocation resources by resolving the problem using a near-optimal solution. Recently, an equity algorithm is applied on routers to ensure the balancing assignment of data through different routers [9]. This latter research work utilizes the same approach of our studied problem. This presented paper is organized as follows. In Section 2, we present the studied problem and we give some details for the problem in general. Section 3, presents six developed lower bounds through several heuristics for the studied problem. The experimental study is done in Section 4.

II. Problem Description

The problem proposed in this paper is the maximization of the minimum total budget. This problem is described as follows. Let a set Ps that determined by the n_p projects. These projects must be distributed to a fixed number of regions n_r. We index each region R_i by the index i. The estimated budget of each project j is denoted by b_j. We can't assign the same project to different regions. Each project will be assign at only one region at a time. The studied problem concerns the maximization of the minimum total budget given by R_{min}. The cumulative budget of the project j is given by the variable Cb_j.

Example 1: Assume that the number of projects is 5 ($n_p = 5$) and the number of regions is 2 ($n_r = 2$). The budget b_j for each project j is depicted in Table I.

TABLE I. BUDGET-PROJECTS VALUES.

j	1	2	3	4	5
b_j	7	2	9	3	12

We schedule all projects on the regions applying a given algorithm. The schedule gave a result which is depicted in Fig. 1. From the latter figure, we see that region 1 has projects 3, 1 and 4. However, for region 2, projects 5 and 2 are assigned.

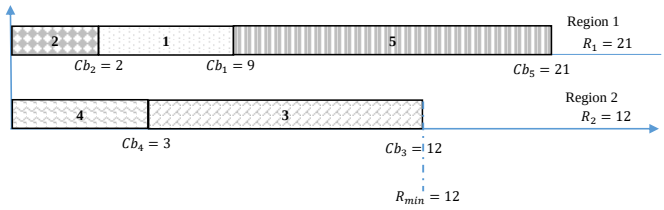


Fig. 1. Project-regions assignment.

From Fig. 1, we see that the region 1 has a total budget of 21. Nonetheless, region 2 has a total budget of 12. For this example, the maximum assigned budget is 21. However, the minimum total assigned budget is $R_{min} = 12$. The goal is to give a permutation (assignment) that maximizes the minimum of the assigned total budget of R_{min}. So, we have to search for other more performed heuristics that offers a minimum total budget better than 12.

Using the three-field notation in [10], the studied problem can be denoted as $P||C_{min}$.

III. LOWER BOUNDS

In this research work, we give a comparison of the performance of the proposed lower bounds for the studied problem with the well-known dispatching rule LPT given in [1]. Indeed, we develop six lower bounds. Two strategies are proposed in this section to develop the six algorithms. The first strategy is based on the probabilistic method applied to the choice of the project which will be assigned to the regions. The second strategy is based on the Multi-Fit algorithm to be used for the proposed problem.

A. Non-increasing Budgets Order Heuristic (NIB)

In this dispatching rule, all projects will be ordered in the non-increasing order of its budgets. After that, we assign the project having the greatest budget on the region that has the minimum total budget at this time.

B. Probabilistic based Heuristics (P_k^r)

For this type of heuristic, our study is essentially based on the selection of the project having the biggest budget to be assigned to the region that has the minimum total budget with a probabilistic approach. The probabilistic selection between the k biggest project with $k \in \{2, 3, \cdots, 6\}$ for the heuristics $\{P_2^r, \cdots, P_6^r\}$, respectively.

The chosen project is selected among the k first project having the biggest budget with probability α. We denote by $RD(a, b)$ the randomly function that return an integer between two limits (a, b). The calculation of the probabilistic selection is fixed as follows:

- $r = RD(1, k)$. The picked project will be the r^{th} biggest non-assigned project. We assign the chosen project to the region that have a minimum total budget.
- Denoted by U_p the number of non-assigned projects. If $U_p < k$, then $r = RD(1, U_p)$.

For a fixed integer k, we choose the iteration number as $limit = 1500$. The heuristic of the probabilistic project-region is the following:

Algorithm 1 Probabilistic project-region heuristic: $PPR(k)$

1: Set $itr = 1$.
2: Set $P_k = P$.
3: $r = RD(1, k)$.
4: Assign the r^{th} biggest project (L_p), to the region that have the minimum total budget.
5: $P_k = P_k \setminus L_p$, if $P_k \neq \emptyset$ goto 3.
6: Determine the R_{min}^{itr}
7: $itr = itr + +$.
8: if $itr \leq limit$ goto 2.
9: Return $PPR(k) = \max_{1 \leq itr \leq limit} R_{min}^{itr}$. , Stop.

The above algorithm is given for a fixed k.

The Algorithm 1 is used when the k value is fixed in advance. To develop the algorithm of the proposed heuristic we must iterate the above as the number between 2 and k. Indeed, the final algorithm which represents the probabilistic based heuristics P_k^r needs the iteration of $PPR(j)$ from 2 to k. The calculation of the value of the heuristic P_k^r for a predetermined k is elaborated by as following algorithm.

Algorithm 2 Probabilistic algorithm P_k^r

1: **for** $t = 2$ to k **do**.
2: Determine $PPR(t)$.
3: **end for**
4: $P_k^r = \max_{2 \leq t \leq k} PPR(t)$.

C. Multi-fit Projects based Heuristic (MFP)

The Multi-Fit heuristic is developed in several works in literature. For our proposed research work we propose an adaptive Multi-Fit heuristic. The main idea is based to found the minimum total budget when all projects will be assigned to the given regions.

The heuristic is based on the utilization of the bin-packing approach. For our case study, the region is considered as bin and the project as an item. Once the bin capacity is fixed, the First Fit Decreasing (FFD) algorithm is used to fit projects to the bin. The first step of the algorithm is to order projects according to their budget such that $b_1 \geq b_2 \geq \ldots b_{n_p}$.

Let $L = max(b_1, b_{n_r} + b_{n_r+1}, \lceil \frac{\sum_{p=1}^{n_p} b_p}{n_r} \rceil)$ and U is the value given by applying the LPT heuristic for the problem

$P||C_{max}$. The number of reputation ite is fixed in advance. This number is used for the iteration of FFD. We set $ite = 135$ and n_b denoted the number of bins used when we apply the FFD routine.

The algorithm below describe all instructions to calculate MFP.

Algorithm 3 Multi-fit Projects algorithm (MFP)

1: Put $k = 0$, $u = U$; $l = L$.
2: Calculate $mi = \lfloor \frac{u+l}{2} \rfloor$, set $k = k + 1$.
3: Apply FFD with capacity mi.
4: If we can assign all projects n_p into n_r regions, after put $u = mi$ and go to 5, else set $l = mi$ and go to 5.
5: If $k = ite$ then STOP, otherwise go to 2.
6: If $n_b > n_r$ then the permutation obtained by NIB is selected, otherwise the permutation obtained by FFD is selected. The selected permutation noted by σ.
7: Calculate the C_{min} of the permutation σ.
8: $MFP = C_{min}$.

Example 2: Let $n_p = 10$ and $n_r = 2$. The following table represent the budget of each project.

TABLE II. INSTANCE OF BUDGETS-PROJECTS FOR HEURISTIC MFP

j	1	2	3	4	5	6	7	8	9	10
b_j	81	30	33	66	53	62	66	59	96	55

Applying the NIB heuristic for the instances given in Table II, the result will as follows. For the first region, we assign the projects $\{2, 4, 5, 8, 9\}$ with a total budget of $U = 304$. For the second region, we assign the projects $\{1, 3, 6, 7, 10\}$ with total budget $L = 297$. Then, $mi = \lfloor \frac{304+297}{2} \rfloor = 302$. Applying FFD function with capacity 302 and the teen items which is corresponding to projects in Table II. The region 1 has the projects $\{1, 4, 8, 9\}$ and the second region will have the projects $\{2, 3, 5, 6, 7, 10\}$. The first region has a total budget of 302 however the second one has a total budget of 299. Thus, $C_{min} = 299$. On the other hand, the value returned by the heuristic LPT (NIB) is 297. So, the result obtained by MFP is better than NIB.

IV. EXPERIMENTAL RESULTS

We adopt several instances of classes to discuss with results and show the developed algorithms comparing with the LPT cited in previous research works.

All algorithms proposed in this paper are coded with Microsoft Visual C++ (Version 2013). The developed algorithms were coded and running on an Intel(R) Core(TM) i5-3337U CPU @ 1.8 GHz. The operating system is windows 10.
The developed algorithms were running on five types of sets of instances. These types of instances were generated as presented in [1]. We generate the budget b_j using to different distributions. Each distribution represents a class. The distribution of the classes are as following:

- Class 1: b_j is in $U[30, 100]$.
- Class 2: b_j is in $U[50, 300]$.
- Class 3: b_j is in $U[200, 500]$.
- Class 4: b_j is in $N[50, 150]$.
- Class 5: b_j is in $N[25, 500]$.

with $U[a, b]$ represent the uniform distribution between $[a, b]$ and $N[a, b]$ represent the normal distribution. The generated instances was obtained by the choice of n_p, n_r and $Class$. The pair (n_p, n_r) can have several affectation values. Indeed, we choose the affectation values given in Table III below.

TABLE III. THE n_p AND n_r DISTRIBUTION

n_p	n_r
10	2,3,5
25,50	2,3,5,10,15
100,250	3,5,10,15,25,30
300,500	10,15,30,50

Table III shows that we have in total of 1650 instances. Several indicators are given in this paper to show the performance of the proposed algorithms compared with the LPT values proposed in the previous work and presented in [1]. The indicators are given as follows.

- Lm the best heuristic value returned after running of algorithms.
- L the discussed lower bound.
- $GAP = \frac{Lm-L}{L} \times 100$.
- $Time$ the running time execution for heuristic. This time will be in seconds.
- $Perc$ the percentage among the (1650) instances that the condition $Lm = L$ is satisfied.

Table IV present the overall percentage and average time for all proposed algorithms.

TABLE IV. OVERALL PERCENTAGE AND AVERAGE TIME COMPARISON

	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
$Perc$	16.2%	16.1%	41.6%	52.2%	58.5%	68.2%	84.3%
$Time$	0.000	0.023	0.047	0.095	0.139	0.192	0.237

The above table shows that the heuristic that conducts the best value is P_6^r with $Perc = 84.3\%$ and $Time = 0.237$s, compared to LPT rule that the corresponding percentage is 16.2%.

The following table represents the behavior of Gap according to n_p.

TABLE V. VARIATION OF GAP ACCORDING TO n_p

n_p	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
10	1.67	2.17	0.47	0.08	0.03	0.00	0.00
25	1.78	4.60	0.77	0.46	0.33	0.18	0.08
50	0.84	7.19	0.24	0.11	0.07	0.04	0.03
100	0.66	30.24	0.38	0.27	0.18	0.10	0.05
250	0.15	5.12	0.10	0.07	0.05	0.03	0.02
300	0.22	1.88	0.12	0.08	0.05	0.03	0.01
500	0.08	2.35	0.05	0.04	0.03	0.02	0.01

Table V shows that when varying the number of projects (n_p), the performance of the developed algorithms varies. The

above table shows that there aren't any depends on the number of projects and the calculated GAP. For heuristics [LPT, MFP, P_2^r, P_3^r, P_4^r, P_5^r and P_6^r] the best GAP value was obtained for the following values of n_p [500, 300,500, 500, (10,500), 10, 10], respectively. Also this table shows that algorithm P_6^r having the best GAP when $n_p = 10$ and the rule LPT having the worst GAP when $n_p = 500$.

The execution time for heuristics in Table V is given in Table VI.

TABLE VI. $Time$ FOR EACH HEURISTIC AND EACH CORRESPONDING n_p

n_p	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
10	0.000	0.002	0.001	0.002	0.004	0.005	0.004
25	0.000	0.004	0.003	0.007	0.009	0.012	0.015
50	0.000	0.006	0.006	0.012	0.017	0.032	0.027
100	0.000	0.010	0.017	0.034	0.052	0.070	0.095
250	0.000	0.025	0.058	0.120	0.173	0.230	0.291
300	0.000	0.060	0.086	0.171	0.255	0.346	0.437
500	0.000	0.062	0.174	0.354	0.522	0.727	0.886

Table VII display the results of the GAP value according to the number of regions n_r. The worst GAP value is given for algorithm MFP when $n_r = 30$ and the best GAP value is given for algorithm P_6^r when $n_r = 3$ or $n_r = 50$.

TABLE VII. VARIATION OF THE GAP VALUES ACCORDING TO n_r

n_r	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
2	1.01	0.40	0.20	0.14	0.11	0.09	0.08
3	1.52	1.28	0.59	0.19	0.09	0.04	0.01
5	0.39	0.90	0.05	0.03	0.03	0.02	0.02
10	0.97	3.29	0.51	0.30	0.21	0.11	0.03
15	0.42	7.22	0.25	0.20	0.15	0.10	0.05
25	0.80	2.70	0.41	0.25	0.15	0.07	0.03
30	0.28	51.14	0.15	0.09	0.05	0.03	0.02
50	0.27	4.72	0.16	0.12	0.07	0.04	0.01

The execution time for heuristics in Table VII is given in Table VIII.

TABLE VIII. $Time$ FOR EACH HEURISTIC AND EACH CORRESPONDING n_r

n_r	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
2	0.000	0.003	0.004	0.005	0.007	0.013	0.013
3	0.000	0.001	0.013	0.025	0.037	0.054	0.070
5	0.000	0.016	0.014	0.029	0.041	0.054	0.067
10	0.000	0.040	0.050	0.101	0.149	0.202	0.251
15	0.000	0.014	0.055	0.109	0.163	0.223	0.278
25	0.000	0.033	0.045	0.093	0.135	0.183	0.230
30	0.000	0.022	0.092	0.186	0.279	0.400	0.474
50	0.000	0.088	0.154	0.322	0.465	0.619	0.790

The Table IX present the variation of the calculated GAP according to $Class$. It is clear to see that the class instances and the GAP values are not correlated.

TABLE IX. VARIATION OF THE GAP VALUES ACCORDING TO $Class$

$Class$	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
1	0.69	8.60	0.26	0.13	0.09	0.04	0.02
2	0.70	5.71	0.22	0.11	0.07	0.04	0.02
3	0.72	9.88	0.34	0.17	0.11	0.06	0.03
4	1.17	16.23	0.45	0.24	0.16	0.08	0.02
5	0.38	4.21	0.26	0.19	0.15	0.10	0.07

The execution time for heuristics and Table IX is given in Table X.

TABLE X. $Time$ FOR EACH HEURISTIC AND EACH CORRESPONDING $Class$

$Class$	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
1	0.000	0.027	0.046	0.091	0.136	0.190	0.236
2	0.000	0.021	0.048	0.094	0.140	0.205	0.242
3	0.000	0.028	0.047	0.098	0.140	0.192	0.238
4	0.000	0.003	0.047	0.096	0.141	0.188	0.234
5	0.000	0.033	0.046	0.093	0.140	0.186	0.237

For more statistics about the gap values, we present Table XI. From the Table XI, we can see that the maximum GAP is 171.59 obtained when $n_p = 100$ and $n_r = 30$ for the heuristic MFP.

TABLE XI. DETAILS FOR ALL ALGORITHMS

n_p	n_r	LPT	MFP	P_2^r	P_3^r	P_4^r	P_5^r	P_6^r
10	2	1.08	0.59	0.03	0.00	0.00	0.00	0.00
	3	3.92	3.14	1.39	0.24	0.08	0.01	0.00
	5	0.03	2.79	0.00	0.00	0.00	0.00	0.00
25	2	1.91	0.57	0.56	0.41	0.34	0.27	0.24
	3	1.64	1.94	0.73	0.26	0.11	0.04	0.00
	5	1.30	1.11	0.13	0.08	0.11	0.07	0.06
	10	4.05	16.87	2.46	1.53	1.07	0.54	0.10
	15	0.00	2.52	0.00	0.00	0.00	0.00	0.00
50	2	0.06	0.03	0.00	0.00	0.00	0.00	0.00
	3	1.57	0.73	0.49	0.21	0.10	0.05	0.01
	5	0.41	0.39	0.06	0.02	0.02	0.02	0.03
	10	1.04	1.81	0.28	0.12	0.08	0.08	0.06
	15	1.11	32.99	0.40	0.20	0.14	0.07	0.07
100	3	0.34	0.42	0.26	0.16	0.08	0.04	0.02
	5	0.16	0.18	0.06	0.03	0.02	0.01	0.01
	10	0.48	0.54	0.20	0.11	0.08	0.05	0.02
	15	1.05	4.10	0.83	0.78	0.62	0.39	0.16
	25	1.27	4.62	0.62	0.38	0.23	0.11	0.04
	30	0.68	171.59	0.32	0.16	0.07	0.03	0.03
250	3	0.11	0.17	0.10	0.08	0.05	0.03	0.01
	5	0.04	0.04	0.01	0.01	0.00	0.00	0.00
	10	0.12	0.23	0.05	0.02	0.01	0.01	0.00
	15	0.21	1.47	0.17	0.16	0.13	0.09	0.06
	25	0.32	0.78	0.20	0.12	0.08	0.04	0.01
	30	0.10	28.02	0.05	0.03	0.03	0.01	0.00
300	10	0.09	0.17	0.04	0.02	0.01	0.01	0.00
	15	0.14	0.22	0.08	0.05	0.04	0.03	0.00
	30	0.28	1.69	0.18	0.11	0.08	0.03	0.01
	50	0.36	5.43	0.20	0.15	0.08	0.05	0.02
500	10	0.05	0.10	0.03	0.02	0.01	0.01	0.00
	15	0.02	2.00	0.01	0.00	0.00	0.00	0.00
	30	0.06	3.27	0.05	0.05	0.04	0.04	0.04
	50	0.19	4.02	0.13	0.09	0.06	0.03	0.01

V. CONCLUSION

This paper treats the problem of the distribution of different projects having several budgets. The problem is NP-hard. We proposed in this paper several heuristics to give approximates solutions for the studied problem. The algorithms given in this work are tested through some classes of instances generated uniformly based on classes already proposed in the literature review. The results show that the probabilistic and iterative method outperforms the Multi-fit and dispatching rule algorithms. The results show that the algorithm P_6^r has the best GAP value through the tested instances. Future work can be

done regarding the performance of other heuristics and the exact solution of the studied problem.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project Number No. R-1441-30.

REFERENCES

[1] M. Haouari and M. Jemmali, "Maximizing the minimum completion time on parallel machines," *4OR*, vol. 6, no. 4, pp. 375–392, 2008.

[2] S. Alexiadis, "Regional allocation of investment: Implications for regional policy," *Procedia Economics and Finance*, vol. 9, pp. 176–180, 2014.

[3] Y. Ohtsuki, "Regional allocation of public investment in an n-region economy," *Journal of Regional Science*, vol. 11, no. 2, pp. 225–233, 1971.

[4] J.-J. Qi, Y.-J. Liu, H.-T. Lei, and B. Guo, "Solving the multi-mode resource availability cost problem in project scheduling based on modified particle swarm optimization," *Arabian Journal for Science and Engineering*, vol. 39, no. 6, pp. 5279–5288, 2014.

[5] N. Rahmani, V. Zeighami, and R. Akbari, "A study on the performance of differential search algorithm for single mode resource constrained project scheduling problem," *Decision Science Letters*, vol. 4, no. 4, pp. 537–550, 2015.

[6] M. Jemmali, "Approximate solutions for the projects revenues assignment problem," *Communications in Mathematics and Applications*, vol. 10, no. 3, 2019.

[7] M. Jemmali, L. K. B. Melhim, and M. Alharbi, "Randomized-variants lower bounds for gas turbines aircraft engines," in *World Congress on Global Optimization*. Springer, 2019, pp. 949–956.

[8] N. R. Devanur, K. Jain, B. Sivan, and C. A. Wilkens, "Near optimal online algorithms and fast approximation algorithms for resource allocation problems," *Journal of the ACM (JACM)*, vol. 66, no. 1, p. 7, 2019.

[9] M. Jemmali and H. Alquhayz, "Equity data distribution algorithms on identical routers," in *International Conference on Innovative Computing and Communications*. Springer, 2020, pp. 297–305.

[10] E. L. Lawler, J. K. Lenstra, A. H. R. Kan, and D. B. Shmoys, "Sequencing and scheduling: Algorithms and complexity," *Handbooks in operations research and management science*, vol. 4, pp. 445–522, 1993.

Browser Extension based Hybrid Anti-Phishing Framework using Feature Selection

Swati Maurya[1]
Guru Gobind Singh Indraprastha University
New Delhi, India

Harpreet Singh Saini[2]
New Delhi, India

Anurag Jain[3]
Guru Gobind Singh Indraprastha University
New Delhi, India

***Abstract*—Phishing is one of the socially engineered cybersecurity attacks where the attacker impersonates a genuine and legitimate website source and sends emails with the intention of stealing sensitive personal information. The phishing websites' URLs are usually spread through emails by luring the users to click on them or by embedding the link to fake website replicating any genuine e-commerce website inside the invoice or other documents. The phishing problem is very wide and no single solution exists to mitigate all the vulnerabilities properly. Thus, multiple techniques are often combined and implemented to mitigate specific attacks. The primary objective of this paper is to propose an efficient and effective anti-phishing solution that can be implemented at the client-side in the form of a browser extension and should be capable to handle real-time scenarios and zero-day attacks. The proposed approach works efficiently for any phishing link carrier mode as the execution on clicking on any link or manually entering URL in the browser doesn't proceed unless the proposed framework approves that the website associated with that URL is genuine. Also, the proposed framework is capable to handle DNS cache poisoning attacks even if the system's DNS cache is somehow infected. This paper first presents a comprehensive review that broadly discusses the phishing life cycle and available anti-phishing countermeasures. The proposed framework considers the pros and cons of existing methodologies and presents a robust solution by combining the best features to ensure that a fast and accurate response is achieved. The effectiveness of the approach is tested in a real-time dataset consisting of live phishing and legitimate website URLs and the framework is found to be 98.1% accurate in identifying websites correctly in very less time.**

***Keywords*—Anti-phishing; browser extension; machine learning; feature selection**

I. Introduction

Phishing attack is the cyber threat that has been prevailing on the internet for almost four decades despite having very exhaustive research in this area and numerous anti-phishing solutions and prevention techniques being available. Phishing is one of the prominent cyber-attack that involves sending emails with the intention of obtaining sensitive personal information by pretending to have been sent by a trustworthy or genuine sender. It is a social engineering attack that exploits the weakness found in the user's systems. It is performed over different channels ranging from sending of fake email by the attacker to fabricated fake websites, social networks or even cloud services [1]. Phishing is an automated identity theft activity, which takes the advantage of human nature and loopholes in technology to lure general users to click on fabricated fraudulent hyperlinks that lead to serious consequences for the victim by breaching the data and misusing the credentials for malicious activities. Despite having high-security implementation in internet browsers, phishing attacks succeed in breaching the security approaches.

– Phishing attack life cycle: Phishing campaign is a four-step process that starts with planning and setting up the prelims for the attack in which the target group is chosen and attack techniques are analyzed and finalized. After the attack, results gathered from the attack are processed to gather useful information out of the data collected. The steps involved in the phishing process are shown in Fig. 1.

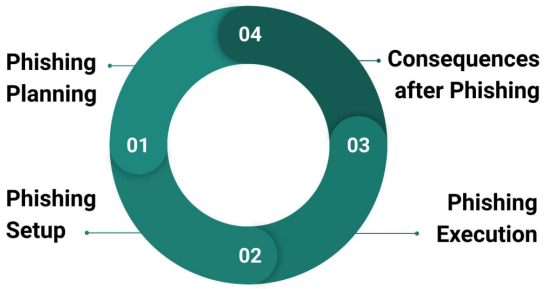


Fig. 1. Steps involved in Phishing attack life cycle

– Phishing Planning: The first step in the Phishing process is planning the attack. The adversaries choose communication media and attack targets.

– Phishing setup: In the second step, the attacking techniques are finalized. The methods for sending the email containing suspicious link and creating the fraud/phished webpage is performed in this step. This step is also known as Phishing preparation.

– Phishing Execution: The important step is performing the attack. It includes Propagation and Penetration steps. The attack material is propagated to the targets. As the target opens the attack material, either the user is prompt to enter personal information supposed to be stolen or a malware or malicious application is downloaded into the target system.

– Consequences after phishing: The phishers execute the information received through the phishing campaign. Attack results, i.e., the sensitive credentials that attack victims entered, are exploited. They misuse them to make illegal fund transactions and commit fraud.

A recent report by Anti-Phishing Working Group (APWG) shows that the number of unique phishing attack reports

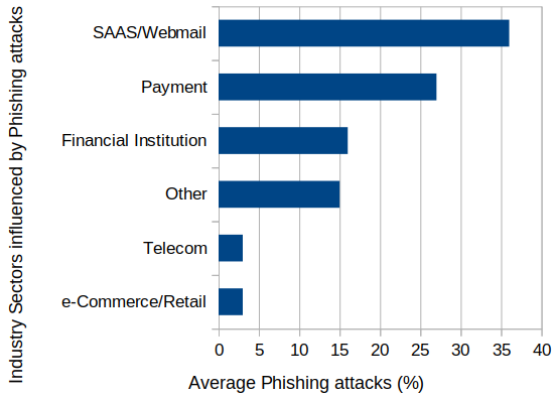


Fig. 2. Most targeted industry sectors for Phishing attacks [2]

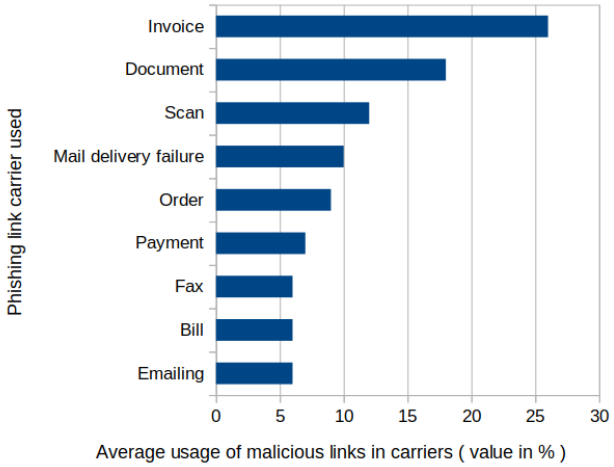


Fig. 3. Fabricated malicious link carrier tactic used for Phishing [6]

submitted during the 1st quarter of 2019 was 1,80,786 nearly 30.7% more than 1,38,328 reported in 4th Quarter 2018 [2]. According to ProofPoint's State of the Phish 2019 Report, 83% of global security professionals encountered phishing attempts [3]. Most affected industry sectors by phishing attacks are represented in Fig. 2. 255065 unique phishing attacks were reported worldwide till 2016. This represents an increase of over 10% from the 230280 attacks that were identified in 2015 due to an increase in 'Domain Name Use' attacks [4]. According to Symantec's 2018 Internet Security Threat Report [5], Sending malicious attachment with the email having fake invoice is the most popular tactic for luring the common public to open the message and click the fabricated link. The percentage of each tactic used for sending fabricated links in Phishing email is shown in Fig. 3.

Most of the anti-phishing solutions in literature claim high accuracy as 98% for phishing detection but most of these measures fail to handle real-time zero-day attacks. There is a huge gap between the high accuracy that has been reported in articles but when it comes to real-time scenario implementation, most of the existing solutions have very low effectiveness. Thus an effective and efficient anti-phishing framework needs to be designed that should provide a fast and accurate response and should be easy to implement. Client-side solutions in the form of browser extensions are easy to install and use. The computational cost associated with browser extensions should be as minimum as possible otherwise it will drastically affect the response time.

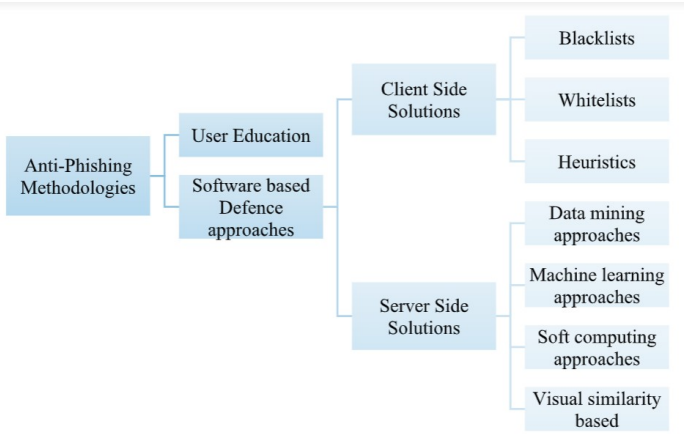


Fig. 4. Social and Software-based Anti-Phishing methodologies

The proposed framework is a hybrid client-side solution that takes the pros of existing anti-phishing methodologies and tries to minimize their associated cons. The paper is structured as follows: Section 2 discusses the literature survey on existing anti-phishing methodologies. The motivation behind the proposed framework is discussed in detail in Section 3 and Section 4 provides system architecture of the proposed framework and its effectiveness in real-time data set that consists of live phishing and legitimate websites. Section 5 represents the results and analyze the observations of the experiments. And finally in Section 6, the discussion about the outcomes of the proposed framework is done their limitations and other issues are also discussed.

II. Literature Survey on Existing Anti-Phishing Methodologies

Mitigation of phishing attacks consists of Detection, Prevention and Correction approaches. The phishing problem is very wide and no single solution exists to mitigate all the vulnerabilities properly. Thus, multiple techniques are often combined and implemented to mitigate specific attacks. Automatically detecting phishing web pages has attracted much attention from security and software providers, financial institutions, to academic researchers. Several approaches and comprehensive strategies have been suggested to tackle phishing. Anti-phishing methodologies can be grouped broadly into the categories shown in Fig. 4:

- User Education

Phishing attacks usually target the users who are not much aware of the defense mechanisms for phishing. Since phishing attacks attempt to take advantage of the inexperienced users, an obvious solution is educating the users, which would, in turn, reduce their susceptibility to falling victims of phishing attacks [7]. Several user training approaches have been proposed throughout the past years. The human factor is broad. Relying solely on educating users about the phishing attacks

and prevention mechanisms does not ensure protection from Phishing attacks.

- Software-based Defence Approaches

Software-based defense approaches can be classified according to the place where they are implemented. If the defense mechanism needs to be implemented at the client's system either in the form of a browser plug-in or toolbar or any desktop application, it is known as a Client-side solution. Server-side techniques depend upon the classification performed on the server-side.

1. Client-Side Solutions

These include user profile filters and browser-based toolbars. These tools also depend on blacklisting and whitelisting techniques where a list of detected phishing or legitimate websites is downloaded with updates at standard intervals.

– Profile Matching (Blacklists and Whitelists)

Blacklists contain suspicious URLs that might redirect to the Phishing webpage, IP addresses, and keywords. These approaches rely on previously detected attacks database and need to be updated frequently. Blacklists are generally implemented as toolbars or web browser extensions. Common examples are Mozilla Firefox's extensions, Google safe browsing API [8], PhishTank [9] and many more. Whitelists act as a filter that contains trusted and reliable URLs and IP addresses which are marked safe and legitimate websites. The rest of the content is treated suspicious and undergoes scrutiny. Blacklists are found to be less effective as compared to whitelists as they cannot deal with 'Zero-day attacks'. Whitelists suffer from a large number of false positives as the lists contain only a limited number of legitimate websites and everything else is considered to be a phishing website.

– Heuristic or rule-based methods

Heuristic techniques are directed to solve the phishing problem through a practical method that utilizes the characteristics present in a phishing attack. The heuristics are identified from general phishing attacks and used for future detection and hence are efficient for zero-hour phishing attacks.

2. Server-side solutions

Server-side defense mechanisms are based on content filtering approaches and are appropriate to fight zero-day attacks. The filters are based on the following techniques:

– Machine learning approaches

The anti-phishing approach where the data is input to complex algorithms that learn the patterns in the given data by mining the insights and takes decision is known as the Machine learning approach. It applies to structured as well as unstructured data sets. It is used for developing techniques for knowledge extraction from datasets based on Artificial Intelligence and Statistics [10]. The representation model formed with the help of this knowledge is used for predictions for new data. Mostly used machine learning classifiers are naive Bayes classifiers, support vector machines (SVM), k-nearest neighbors, Decision Tree, Random Forest, Boosting and TF-DIF.

- Decision Trees (DTs): Decision-tree learning is a symbolic induction method that produces syntactically simple, easily interpreted rules [11]. In DTs, the knowledge extracted from the given data is organized in a recursive hierarchical structure represented with the help of nodes and branches [12]. DTs aim to maximize the correct classification of all training data. The pruning technique is applied to the trained tree to avoid the problem of overfitting [13]. DTs provide the advantage of verifying the attributes that determined the final classification [14].

- Random Forests: A random tree is a tree that is formed by a stochastic process. Random forests are combinations of random tree predictors [15]. They are an ensemble learning method for classification and regression that overcome the problem of overfitting in decision trees.

- k-Nearest neighbor: k-NN is the simplest machine learning technique. It is an instance-based learning, also known as lazy learning, that stores all training data and classifies a new data point according to the class of the majority of its k-nearest neighbors in the given feature space [14]. For finding the nearest neighbors for each data, different measures are used to calculate the distance between pairs of data nodes. Euclidean distance is the most commonly applied measure.

- Naive Bayes: NB are the probabilistic classifiers based on Bayes theorem. The classifier learns the conditional probability of each attribute value from the training data given the classification of each instance. For the classification of an unknown instance, Bayes' theorem is applied. Naïve Bayes training is usually performed by using maximum likelihood algorithms.

- Support Vector Machines: SVM are linear classifiers based on Statistical Learning Theory [16]. SVMs perform structural risk minimization, for improving the complexity of the classifier. SVM constructs a hyperplane that optimally separates the data into two categories in higher-dimensional space.

– Data Mining Techniques

The techniques that come under this category consider phishing to be a classification or clustering problem and algorithms based on data mining with the help of machine learning techniques are applied to them.

– Soft Computing Techniques

Soft Computing is the fusion of methodologies that are applied to real-world scenarios to find the optimum solution to the problems that are not easily modeled mathematically. In this approach, knowledge discovery is used to simplify the evolution process. The evolving clustering method for classification and approaches based on Fuzzy neural networks is used to develop a model for phishing detection which performs the classification using some features to classify phishing and legitimate emails.

Various anti-phishing research works that are based on profile matching, heuristics approaches, machine learning algorithms, and soft computing techniques are discussed below:

Kirda and Kruegel (2005) presented "Anti-Phish" [17], a Mozilla browser extensions that captures and stores sensitive information using Paul Tero's Javascript DES implementation

with the context and domain of the website where it was submitted. Whenever a user enters any credentials on a website, Anti-Phish compares the records in the watch list and if the same values are being entered, the domain is examined thoroughly. If the domain is different from the domain in trusted records, it is assumed to be a phishing website and an alert is given to the user. This approach ensures that any sensitive information is not forwarded to any suspicious website. The approach is efficient to monitor javascript hooks embedded in HTML webpages to capture data on keystrokes and send this information to the malicious server in the background even without the user submitting the details.

Prakash et al. (2010) proposed a system that combines two components: URL prediction and Approximate URL matching for addressing the exact match of URLs limitation in case of blacklist approaches [18]. New malicious URLs are predicted from the existing entries in the blacklist and are tested whether they are malicious through 'DNS queries' and 'Content matching'. For generating new URLs, five heuristics, namely, 'Replace Top-level domains (TLD)', 'IP address equivalence', 'Directory Structure similarity', 'Query string substitution' and 'Brand name equivalence' s utilized. The new URL, after creation, is subjected to a validation process that uses the DNS Lookup mechanism to filter out URLs that do not exist or are marked as legitimate. For the approximate matching process, the input URL is broken into 4 entities- IP address, hostname, directory structure, and brand name. These entities are matched with corresponding fragments of the blacklist entries and based on the matching score; URLs are marked as Phishing or Legitimate. This technique is capable of fast detection but suffers the drawback of too high false negatives, i.e., 5%.

Belabed et al. (2012) proposed an extension for web browsers that combines a personalized whitelist and SVM classifier [19]. The whitelist is implemented in the form of an XML file which consists of user's login pages' URLs and a set of keywords that is composed of domain names of the page's URL and terms from DOM tree for the website. - "Bag of words" model [20] is used for constructing keywords' frequency vector and Cosine distance [21] to find the similarity between the visited webpage and the webpages available in the whitelist repository. A similarity check is processed as explained in Fig. 5. A feature vector that represents the webpage utilises 8 features according to its URL and content (URL with IP address, Special characters in the URL, presence of SSL certificate, whether the identity of the webpage conforms to its URL, search engine ranking, nil anchors, frequency of links, action complies with the page identity) is processed by the classifier. This approach can detect 98% phishing webpages accurately. The major drawback of this approach is the high false-positive rate, i.e, 3.5%.

Islam and Abawajy (2013) proposed a "Multi-tier classification model" for filtering phishing emails that combines multiple classification algorithms to reduce false positives and increase the overall efficiency [22]. Based on the weighting of message content and message header, the features from an email are extracted and prioritized according to their ranks. Different combinations of classifier algorithms in a multi-tier classification process are tested and the impact of rescheduling is examined. 21 features are used for classifying

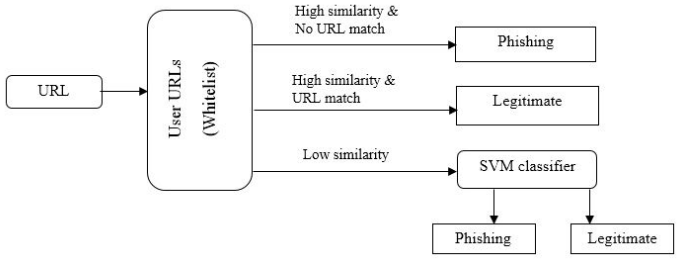


Fig. 5. Classification based on similarity with webpages in whitelist in [19]

emails and Spamassassin public corpus, 2006 and Phishing Corpus homepage 2006 dataset is used. For 1st tier- SVM, 2nd tier- AdaBoost and for 3rd tier- Naive Bayes Classification algorithms are used. This approach attained the accuracy up to 97%, at most 2% FP and at most 9% FN.

Akinyelu and Adewumi (2014) proposed a content-based phishing detection approach that works for email filtering [23]. A set of 15 features selected from the literature is used for forming a vector representation by extracting values from a dataset of 2000 emails and then the Random Forest classifier is tested and trained using 10-fold cross-validation. In this classification, before decision tree construction, information gain for all the features is calculated and the features that are found to have the best 8 information gain construct the decision tree. Prediction of emails is done by mode vote of all the trees. This algorithm assured the accuracy of 99.7% and true positives are quite low, i.e., 0.06% whereas false negatives are too high (2.5%).

Gowtham and Krishnamurthi (2014) proposed a hybrid approach that combines preliminary filters that eliminate webpages not containing login forms and 15 heuristics based on structural and behavioral properties of webpage [24]. The system architecture consists of: Module 1- Preapproved site identifier: A self-constructing private whitelist is used for prevention from pharming attacks. This module contains legitimate websites checked with remote DNS lookups. When the user accesses any website, this module checks the URL and IP address of the website and declares it legitimate if found in the whitelist or else forwarded to the next module. Module 2- Login Form Finder: In this module, 38 login keywords and a search keyword are used to verify whether a webpage has a login form or not. The webpages without login form are marked safe and stopped from entering the next step. Webpage Feature generator: The webpage's identity is extracted from its hyperlinks and contents and the heuristics clustered into six groups check the phishing characteristics of the webpage. A 15- dimensional vector is formed by domain descriptor. SVM classifier: The feature vector generated is inputted for classification. The rules developed during the training period are used for predicting the class. The class label of the classification of new test cases is the output based on the patterns found in training data. This approach is tested over a dataset of 2464 live websites from which 700 sites are legitimate and 1764 are phishing sites. TPR of 99.6% and a reduced FPR of 0.42% are achieved.

Rao and Ali (2015) developed a desktop application based

on 5 modules that input the URL of the website and declare the status of the website as a legitimate or phishing [25]. This approach processes only the websites containing the login form. Module 1- Whitelist: Comparison of the domain of URL with that of the genuine websites stored in the database is done and the legitimate status is set for the sites that match in the list. The sites that do not match are passed to the next level. Not the entire webpage but its DOM is passed to the next module. Module 2- Zero links in the body portion of HTML indicate that the body section of the webpage does not have any link embedded and text is appearing in the form of images. This heuristic if marked zero represents a phishing site and stops the process or else proceeds to the next step. Module 3- Footer links pointing to NULL: In any legitimate site, the footer link never points to null. This heuristic is used to classify the websites as phishing those have the anchor tag in the footer section pointing to null. Module 4- Copyright and Title content is extracted and compared with the sites in whitelist. A match found represents a phishing website. Module 5- Website Identity: This reveals the original targeted website that the phished sites trying to pretend. PhishShield application is based on Jsoup API and Firebug. Whitelist used is based on PhishTank's target list [9]. For evaluation of the approach, 1600 phishing sites and 250 legitimate sites are tested. The major limitation of this approach is its inability to detect phishing websites that do not have any resemblance with or try to imitate any of the legitimate websites and can successfully bypass all the filters used in this approach.

Jain and Gupta (2016) proposed an anti-phishing solution in the form of a "browser plug-in" that matches the current domain with the legitimate domains defined in the whitelist [26]. The proposed approach works through two modules: URL and DNS matching module matches the domain of the webpage and IP address from the whitelist. The phishing identification module uses 'hyperlink features' which are extracted from the DOM object properties of a webpage using Jsoup [27] to check the legitimacy of the website. Phishing detection algorithm is applied after this which takes decision based on three parameters of the hyperlinks whether the webpages contain any hyperlink, whether the webpage contains null pointer and the number of links that are pointing to its own domain and outside domain. The system can detect DNS poisoning, embedded objects and zero-hour attacks with an accuracy of 89.38%, TPR=86.07%, and FNR=1.48%.

Yang et al. (2017) analyzed and evaluated the performance and effectiveness of the C4.5 algorithm for detecting phishing websites [28]. They proposed a decision tree using C4.5 for classification of websites into phishing and legitimate. The dataset used (PWD) is obtained from the UCI machine learning repository that consisted of 11055 websites (4898 phishing websites + 6157 legitimate websites). The technique used 30 features of Phishing website dataset which are partitioned into four classes: Address bar based features, abnormal based features, HTML and JavaScript-based features, and domain-based features. To study the effect of dimension reduction of features on the performance of the decision tree, a dataset PW2 is generated from PWD using the selection attributes method. PW2 contains 9 features. After applying C4.5 on PWD and PWD2, it has been observed that PWD2 is better than PWD in terms of complexity and computational cost. However, the results show that reducing the number of features for classification decreases the accuracy slightly. Hence, it can be concluded that with more features, classification is better.

Li et al. (2019) collected real-time websites' data to form two datasets, namely, 50K-PD and 50K-IPD [29] Their approach is to extract features by using Word2Vec model to extract HTML string embedding by learning distributed representations of words and combine with features extracted from URL. In the next step, Gradient Boosting Decision Tree, LightGBM and XGBoost machine learning models are combined to form a stacking model. The predictions of training set obtained by using these 3 models are combined with the original feature set to form a new training set which is fed to the next layer of this stacking model. In the end, the GBDT model is used for the final decision. The claimed accuracy is 96.45%.

Ding et al. (2019) used Search engine based detection technology combined with heuristics to classify websites into legitimate and phishing [30]. The approach enters the title tag of the webpage as a search keyword into 'Baidu' search engine and marks it as legitimate if the website is within the top ten results from searching and skips the next steps. If the result is not in the top 10 results, URL heuristics of the webpage are matched with the defined rules and classify the webpage as phishing if it matches. If this step fails to mark the webpage as phishing, the logistic regression classifier is fed with "URL's DNS, Whois, similarity with phishing vocabulary, lexical feature and HTML" and the classifier results with the final prediction as phishing or legitimate website. The approach has been tested with datasets from Phishtank, Yahoo, URLB, and DMOZ and claims to have an accuracy of classification as 98.9%.

III. Motivation for the Proposed Framework

Various anti-phishing solutions proposed by different authors have been given in the previous section. However, no single solution is a "full proof" solution for combating phishing attack. Limitations of the existing anti-phishing solutions emphasize the need for innovative solutions. Some organizations provide guidelines and cybersecurity policies to be used by the common user as best practices for prevention from online fraud and phishing attacks [31]. The effectiveness of an anti-phishing solution depends on its capability to recognize a phishing website or email within an acceptable time period. As seen in the literature, numerous anti-phishing solutions are available, but most of them are unable to take highly accurate and precise decisions. In most of the techniques, a rise in false positives has been observed, i.e., classifying legitimate websites as a phishing website.

Every anti-phishing approach has its own associated pros and cons:

- The blacklist and whitelist approaches have been observed to have too high false positive and false negative rates. The drawback of the blacklist approach is that the blacklist cannot be updated frequently and does not provide 100% coverage of all phishing websites. Hence, they alone are not effective for zero-day phishing attacks.

- The approaches based on heuristics that use several website features for identifying the type of website are much

more successful for phishing detection as compared to list matching approaches. They are quite effective in detecting fake websites in real-time but some of them have very high FP rates.

- The machine learning and data mining approaches give the best results in phishing detection. But the selection of appropriate classifier is a challenging problem. There are pros and cons associated with every classifier. SVM classifies the webpages with high accuracy but this approach is very time-consuming and is often best used for small datasets. Naïve Bayes classifiers are easily implemented but require the features to be mutually independent. Machine learning techniques require systems to have high computation power to be implemented in real-time but are the most effective ones. It can be concluded from the above sections that Support Vector Machine (SVM), Random Forest and Logistic regression-based classifiers are the most commonly used classifiers in the literature that has been covered for this study.

Most of the anti-phishing solutions in literature claim high accuracy as 98% for phishing detection but most of these measures fail to handle real-time zero-day attacks. There is a huge gap between the high accuracy that has been reported in articles but when it comes to real-time scenario implementation, most of the existing solutions have very low effectiveness. A few major reasons for low-effectiveness of most of the existing solutions are:

1. The design and ideology behind the solution is influenced by high detection accuracy obtained by training the classifiers with dataset having limited features or spatial correlation.

2. The model used for testing the evaluation of anti-phishing solutions are not capable of representing the real-time scenarios effectively.

The major requirements for an effective anti-phishing solution is to have these characteristics:

- The detection performance should be evaluated in real-time scenarios after considering all the use cases and deployment cases.

- The temporal resilience of the dataset is a must.

- The evaluation or assessment methodology should be fast enough to provide efficient results in fractions of seconds.

As discussed in Section 3, anti-phishing solutions can be provided either on the client-side or server-side. Most of the phishing victims are tricked using emails through the webpage they open via link provided in the email. An effective implementation is considered to be the one that gives a prompt decision in the form of an active warning to the user if he is trying to access any webpage that is a phishing webpage. Client-side anti-phishing solutions seem to help in a better way by providing results immediately. The proposed solution is a client-side solution in the form of "browser extension" which gives user prompt alert if the requested webpage is classified as "Phishing" by the detection model.

Few considerations while designing a client-side solution:

Since the main motivation behind adopting a client-side solution is to maintain user privacy and super-fast decisions, a few major points should be considered side by side. The ideal solution should not demand high computational power and since it is going to be integrated with a web browser, it should not degrade the performance of the client-side. The client should be kept engaged or given proper messages in case it is taking any delay even though for seconds as user interaction is much needed. The client-side based solution should always maintain the confidentiality of the detection model.

IV. System Architecture of the Proposed Framework

The basic architecture of the proposed browser extension is shown in Fig. 6. As shown in the figure, this framework is divided into three layers:

1. Whitelist check

The top-most layer contains a list of trusted websites that have been marked as safe by reputed agencies. For now, the records have been taken from Alexa Top Sites [32]. When a URL is requested in the browser either by clicking through e-mail or by manually entering it in the address bar, the browser extension checks the whitelist whether the domain of the requested webpage matches in the records. If the domain matches, IP address is also checked and if it also matches, the webpage is declared "Legitimate", otherwise, "Phishing". If the domain doesn't match, the request is passed to the next layer. By matching domain with the IP address, the proposed solution works against the "DNS cache poisoning" attacks and identifies them.

2. Blacklisting websites

This layer acts as a "filter" which stops the webpages that are phishing webpages. For providing highly accurate results, this solution relies on experts from "PhishTank" for the webpages that they have marked as phishing after thorough examination. This process saves time that would have been consumed in evaluating the same webpages at its own end and hence gives the result very fast.

This layer queries API provided by "PhishTank". There are two types of responses from the Phishtank API. Response string contains "Valid" if the webpage has been declared as "phishing" by their experts. Response is "Unknown" if the website is under evaluation in their system or it is not found in their record. Thus, if "valid" is returned, the process stops here as the webpage is Phishing webpage and the user is given an appropriate alert stating that its a Phishing website. In the case of "unknown", the request is passed to the next layer of the extension.

3. Feature extraction and classifiers

This layer consists of three phases: 'Form evaluation', 'Hyperlinks extraction and analysis' and 'Stack of classifiers'. For extracting features from the webpage, the rules have been used to define condition or range for features as described in Table I.

Phase 1: Form evaluation

The webpage is scanned in this phase to check the presence of any form. If the form is not present, the webpage is declared

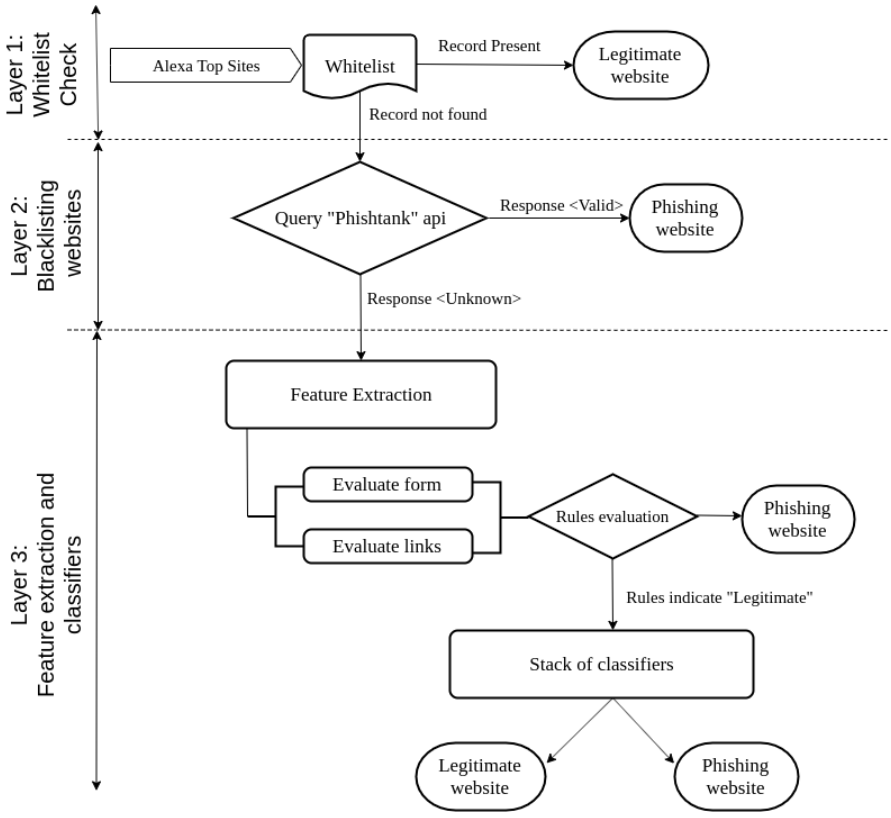


Fig. 6. System Architecture of the proposed solution

as legitimate as it doesn't attempt to capture any information. If a form is present, the request is passed to the next phase.

Phase 2: Hyperlinks extraction and analysis

In this phase, the hyperlinks are extracted from the requested webpage and analyzed according to the following rules:

- Any genuine website will have at least one hyperlink. The static informative websites also have a few hyperlinks. Thus, if any website doesn't have any hyperlink, it is declared as phishing.

- Attackers exploit the vulnerability of using "null pointer" in hyperlinks to give an illusion of having hyperlinks on the webpage. Any genuine website will never have a null pointer in hyperlinks and thus any webpage having them is declared as phishing.

- Evaluating the hyperlinks extracted from the webpage: In a genuine website, only a few hyperlinks point to any other domain and rest all hyperlinks point to the original domain of the webpage. The ratio of links pointing to the same domain vs other domain is evaluated and if the outer domain links are beyond a threshold, it is declared as phishing. From literature [30], it has been found that a 36% threshold of webpages pointing to a foreign domain is most effective in finding the phishing websites and hence this value is used.

The result of this phase if declares the webpage as phishing, the execution stop, and the user is given a prompt alert. If the result shows the website as legitimate, we don't rely on and further process the request to the next layer to be very sure of the result.

Phase 3: Stack of classifiers

In the experiments performed by various literature, it has been found that machine learning classifiers tend to give accurate results but consume too much time and computational cost. This is the reason, the request is passed to this layer only when all other possibilities to decide the status of the webpage have been tried that could give a fast response. In other words, the requests coming to this layer are new websites that have not been encountered yet. Three best classifiers based upon the high accuracy performance provided by each of them when applied to the phishing dataset have been chosen. Random Forest, SVM, and Logistic Regression have been chosen in this approach. If a single classifier is used, chances of true positives and true negatives are high whereas the negative detection results are exempted when a combination of three classifiers is used.

In this approach, every classifier is trained with a phishing website dataset available from UCI machine learning repository [33] which contains 11055 records having 30 features. For optimum results, only the best parameters are chosen which are evaluated by Random Forest Classifier by removing 'Gini impurity' from the features. This involves pruning trees to that extent for a node where a subset of most important features is created. Evaluation is further discussed in the next section.

The prediction from all three classifiers by using the best features is combined to give the final prediction as if a

TABLE I. RULES FOR CLASSIFICATION APPLIED ON THE FEATURES OF WEBSITES

Rules	Feature	Condition
	1. Address Bar based features	
1	URL Having IP	Usually not present in legitimate websites
2	URL Length	Usually <54 observed for legitimate websites
3	Tinu URL	Check URL shortening services
4	Having @ symbol	Not allowed in legitimate websites
5	Double slash	Allowed only once (after http/https ://)
6	Prefix/Suffix	Domain Name shouldnt include (-) symbol for legitimate sites
7	Sub Domain/ Multi Sub Domain	Phishing if number of dots greater than 2 in domain part
8	SSL final state	Checks for https and compares certificate issuer with trusted issuer list. Age of certificate should be greater than 1 year
9	Domain Registration	Checks updated date and expiration date of domain to find out vailidity. If expires <1year – Phishing
10	Favicon	Favicon loaded from foreign domain means Phishing site
11	Non-standard ports	Check status of common ports
12	HTTPS token	Checks if attacker is tricking the user by putting https in domain part
	2. Abnormal based features	
1	Request URL	Checks the requested URLs outside the webpage. Phishing if >61%
2	URL of Anchor	Average of hyperlinks pointing to foreign domain (<31% for legitimate website)
3	Links in tags	Average numer of meta, link and script tags should be <17% for legitimate
4	Server Form Handler	If SFH is “about: blank” or empty – phishing
5	Email submit	Checks usage of ‘mailto:’
6	Abnormal URL	If hostname is not included in URL – Phishing
	3. HTML and JavaScript based features	
1	Website forwarding	website redirection >4 – Phishing
2	Status bar customisation	if “onMouseOver” event changes status bar – Phishing
3	Right Click	If disabled – Phishing
4	Pop-up window	If pop-up window asks for user credentials input – Phishing
5	IFrame Redirection	If using IFrame without “frame orders” – Phishing
	4. Domain based features	
1	Age of Domain	Finds out how old the URL is. If >6 months – Legitimate
2	DNS Record	If claimed identity not recognised by WHOIS or no records found for hostname or DNS record is empty – Phishing
3	Website Traffic	Website Rank >10000 in Alexa database – Phishing
4	PageRank	If <0.2 – Phishing
5	Google Index	Webpage indexed by Google – Legitimate
6	Links pointing to page	If 0 – Phishing

particular classifier fails to identify the phishing website, the rest of the classifiers can help.

V. RESULTS AND EVALUATION

The proposed solution in the previous section is available in the form of a browser extension that has been tested in Google Chrome and Firefox web browser. The extension is developed in Python and is browser-independent. To test the performance and check the accuracy of this browser extension, a dataset containing random 1000 legitimate URLs and 1000 phishing URLs is being used and the achieved accuracy for classifying these web URLs accurately is found to be 98.1%.

The proposed browser extension has two execution layers:

- After the requested URL is not found in the whitelist store, then execution passes to the layer where PhishTank API is queried for the URL verification. This process saves time in processing the URLs that are already known to be phishing and hence the fast response to the user.

- The next layer is based on feature extraction where the website data is pre-processed and features are extracted. It checks for the presence of any form in the website and evaluates hyperlinks in the website based upon the rules defined in the previous section. If a website is declared phishing here at this phase, the execution stops. If it is declared suspicious, we further process the next phase.

- The final phase consists of a stack of three pre-trained classifiers. For choosing the best important features to achieve optimum performance, the best parameters need to be selected out of 30 features related to website data. For this purpose, Random Forest Classifier is used by choosing ‘Gini’ criteria. Fig. 7 shows the graph of the importance of each feature.

For testing the effectiveness of the proposed solution, the same dataset has been evaluated on individual classifiers, namely, Logistic Regression, SVM, Random Forest and a stack which combines prediction of these three. Observations of the experiments are described below and summarized in Table II:

- Fitting logistic regression and creating confusion matrix of predicted values and real values on the Phishing website dataset, 92.3% accuracy is observed.

- Using support vector machine with an rbf kernel and using “gridsearchcv” to predict best parameters for SVM turns out to be a really good choice, and fitting the model with predicted best parameters, 96.47% accuracy is attained.

- Using “gridsearchcv” in Random Forest for feature importance to get the best parameters and fitting best parameters, accuracy 97.26% is achieved.

- Using combined prediction of stack of 3 classifiers (Logistic Regression, SVM and Random Forest classifiers) turns out to be the most effective solution achieving accuracy of 98.1%.

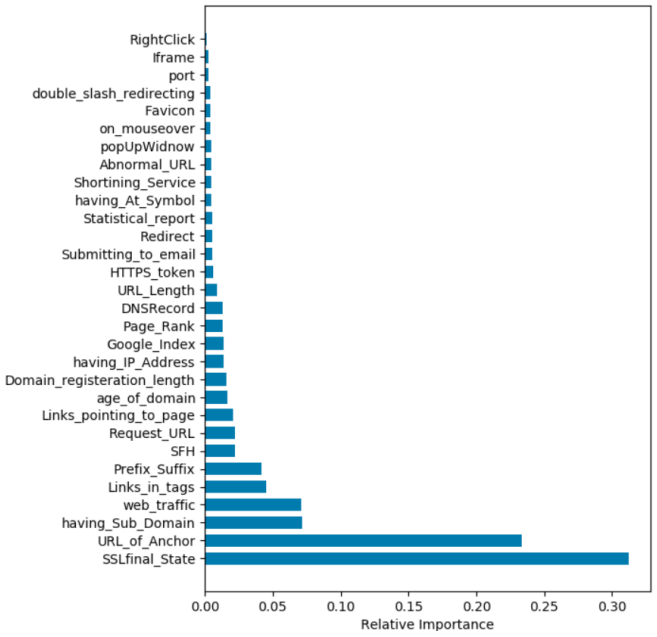


Fig. 7. Relative Importance of features evaluated through Random Forest classifier using GridSearch for Websites features

TABLE II. PERFORMANCE ANALYSIS OF CLASSIFIERS COMPARED TO PROPOSED FRAMEWORK FOR CLASSIFYING WEBSITES AS LEGITIMATE AND PHISHING CORRECTLY

	Accuracy	Mean Squared error	Average execution time(ms)
Logistic Regression	92.3%	0.864	0.98
SVM	96.47%	0.041	0.87
Random Forest	97.26%	0.051	1.75
Proposed framework	98.1%	0.047	0.62

VI. CONCLUSION

Phishing attack is a cyber-security issue that has been prevailing the internet for almost 3 decades now and still, people are getting phished daily. Attackers befool internet users with new phishing techniques and steal their personal secure information. Mitigation of phishing attacks consists of Detection, Prevention and Correction approaches. The phishing problem is very wide and no single solution exists to mitigate the entire vulnerabilities properly. Phishers find out vulnerabilities in the existing solutions and come up with new attacks. Prevention from getting Phished is better safety measure and user's awareness may help in protecting themselves from becoming phishing victims. User awareness helps them in analyzing the website or email in first look based on its prominent features. Software-based anti-phishing solutions to defend the user from email and website phishing are discussed in detail in this paper. Multiple techniques are often combined and implemented to mitigate specific attacks. Automatically detecting phishing web pages has attracted much attention from researchers. Several approaches and comprehensive strategies have been suggested to tackle phishing. This research paper provides a basic understanding of phishing attacks, its life cycle, and popular attack techniques. The advantages and disadvantages of every solution are discussed which helps in analyzing and choosing the appropriate mechanism for implementation. This helps in finding the vulnerabilities in each solution and a direction for future research through modification in these solutions.

After a thorough analysis of existing anti-phishing methodologies, a new browser extension is proposed in this paper which combines the advantages of profile matching techniques and machine learning classifiers. The main motive is to provide the result to the user as fast as possible. Thus, the solution should be efficient to give accurate results in less time. The proposed framework utilizes the resources available from reliable sources to speed up the process. The trusted agencies like 'Alexa' are used for creating whitelist having legitimate websites and 'PhishTank' is used for blocking phishing websites already researched by them. This way, most of the website requests get responses in a very short time. In the case of new websites, that may cause zero-day attacks, the stack of classifiers utilizes the best features and evaluates the website's extracted features to classify it into phishing or legitimate. This approach is capable to classify websites correctly with 98.1% accuracy.

REFERENCES

[1] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *Journal of Information Security and applications*, vol. 22, pp. 113–122, 2015.

[2] "Phishing activity trends report 1 quarter 2019." https://docs.apwg.org/reports/apwg_trends_report_q1_2019.pdf. Accessed: 11-10-2019.

[3] "State of the phish 2019 report." https://info.wombatsecurity.com/hubfs/Wombat_Proofpoint_2019%20State%20of%20the%20Phish%20Report_Final.pdf. Accessed: 14-10-2019.

[4] "Global phishing survey: Domain name use and trends in 2016." https://www.antiphishing.org/resources/apwg-reports/domain-use-and-trends. Accessed: 06-07-2019.

[5] "Internet security threat report, symantec volume 22." https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf. Accessed: 11-09-2019.

[6] "Must-know phishing statistics 2017." https://blog.barkly.com/phishing-statistics-2017. Accessed: 14-08-2019.

[7] P. Khurana, A. Sharma, and P. K. Singh, "A systematic analysis on mobile application software vulnerabilities: Issues and challenges," *Indian Journal of Science & Technology*, vol. 9, no. 32, pp. 1–6, 2016.

[8] "Google safe browsing." https://developers.google.com/safe-browsing/. Accessed: 06-07-2019.

[9] "Phishtank." https://www.phishtank.com/. Accessed: 06-07-2019.

[10] T. M. Mitchell *et al.*, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[11] R. S. Ludwig and M. J. Piovoso, "A comparison of machine-learning classifiers for selecting money managers," *Intelligent Systems in Accounting, Finance & Management: International Journal*, vol. 13, no. 3, pp. 151–164, 2005.

[12] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[13] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[14] A. C. Lorena, L. F. Jacintho, M. F. Siqueira, R. De Giovanni, L. G. Lohmann, A. C. De Carvalho, and M. Yamamoto, "Comparing machine learning classifiers in potential distribution modelling," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5268–5275, 2011.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[17] E. Kirda and C. Kruegel, "Protecting users against phishing attacks with antiphish," in *29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, vol. 1, pp. 517–524, IEEE, 2005.

[18] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in *2010 Proceedings IEEE INFOCOM*, pp. 1–5, IEEE, 2010.

[19] A. Belabed, E. Aïmeur, and A. Chikh, "A personalized whitelist approach for phishing webpage detection," in *2012 Seventh International Conference on Availability, Reliability and Security*, pp. 249–254, IEEE, 2012.

[20] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[21] "Cosine distance." http://reference.wolfram.com/language/ref/CosineDistance.html. Accessed: 22-11-2017.

[22] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 324–335, 2013.

[23] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, 2014.

[24] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," *Computers & Security*, vol. 40, pp. 23–37, 2014.

[25] R. S. Rao and S. T. Ali, "Phishshield: a desktop application to detect phishing webpages through heuristic approach," *Procedia Computer Science*, vol. 54, pp. 147–156, 2015.

[26] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, vol. 2016, no. 1, p. 9, 2016.

[27] "jsoup: Java html parser." https://jsoup.org/. Accessed: 25-12-2018.

[28] Y. Xiang, Y. Li, Y. Bo, and Y.-f. LI, "Phishing website detection using c4. 5 decision tree," *DEStech Transactions on Computer Science and Engineering*, no. itme, 2017.

[29] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using url and html features for phishing webpage detection," *Future Generation Computer Systems*, vol. 94, pp. 27–39, 2019.

[30] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, "A keyword-based combination approach for detecting phishing webpages," *computers & security*, vol. 84, pp. 256–275, 2019.

[31] S. Maurya, S. Sharma, A. Singh, and A. Jain, "Investigation of cyber security practices in academic institutions," in *5th International Conference on Cybercrime and Computer Forensics (ICCCF)*, IEEE, 2017.

[32] "Alexa." https://www.alexa.com/. Accessed: 05-10-2019.

[33] "Phishing websites data set.." https://archive.ics.uci.edu/ml/datasets/phishing+websites. Accessed: 14-02-2019.

Autonomous Navigation of Unmanned Aerial Vehicles based on Android Smartphone

Talal Bonny[1], Mohamed B. Abdelsalam[2]
Department of Electrical and Computer Engineering
College of Engineering, University of Sharjah, UAE

***Abstract*—In the past few years, the adoption of drone technology across industries has increased dramatically as more businesses started to recognize its potential, uses, and scale of global reach. This paper proposes a design solution of a smart GPS quadcopter aircraft navigation system, discusses its hardware and software implementation process and eventually, analyses and reports the final test results. The flight path of the quadrotor is remotely manipulated via an Android based Graphical User Interface. This outdoor handheld application allows the operator to select a point of interest through Google map satellite view; consequently, the quadrotor takes off then hovers and ultimately lands on the destination location. Instructions in conjunction with coordinates are sent and received throughout a web server which serves the communication operation between the smartphone and the quadrotor. Experimental results yield fruitful data communication and successful autonomous flight control with smooth and stable maneuvering.**

***Keywords*—UAV; drone; quadrotor; android; Arduino; GPS; GPRS; GSM**

I. Introduction

Unmanned Aerial Vehicles (UAVs) are deemed one of the utmost pioneering aviation innovations in the history of humankind. In fact, the first attempts traced back to the early 1990s. The French aircraft designer Louis Charles Breguet was the first to build a rotary wing aircraft, in 1907, which was able to lift itself off the ground [1]. Few years later, in 1920, Oehmichen tried six other rotorcrafts designs [2]. The second design he built, which uses four rotors and eight propellers, demonstrated considerable degree of stability for its time and remained airborne for several minutes. More robust prototypes emerged in 1922, 1956 and 1958 [3], [4]. Ever since, endeavors continued to improve and advance, thus, sophisticated yet simple designs and refined prototypes proliferated and; consequently, scholars are, presently, racing to develop competitive applications of quadrotors in various sectors and multiple domains. From surveillance, reconnaissance and weapon delivery, in military and law agencies, through Internet of Vehicles [5], to aerial imagery and television news coverage [6], [7], [8]. Besides, it is widely utilized in commercial transportation of goods such as food, pharmaceutical products and parcels delivery [8]. To process the data, a controller usually is used such as CPU [9], [10], FPGA [11], [12], or Microcontroller as explained in this work. Drones and control systems play an important role in our daily life, from hobbies like photography to more important uses like surveillance and military purposes. All these applications focus on specific usage, all of which lack a navigation system to move the drones autonomously from one location to another. Any application that will use drones will have to call them or send them to different locations. Our objective is to provide a friendly prototype that provides the utilities which will execute these operations. Lots of contemporary research studies attempted at designing and developing different approaches and various methodologies towards the goal of achieving autonomous aerial navigation maneuvers [13]. In this context, [14] proposed a framework, operational in unknown indoor and outdoor environments, of a Micro Aerial Vehicle (MAV) with autonomous capabilities using one monocular camera, as exteroceptive sensor, along with an onboard Inertial Measurement Unit (IMU). Likewise, the work presented by [15] extended the earlier research study to implement an innovative approach of automatic navigation systems using two cameras with non-overlap in their respective field of views (FOVs). Similarly, and according to [16], the navigation of aerial vehicles can be accomplished using multi-sensor data fusion method which determines precise pose parameters of the vehicle in real time. In another work, [17] used information obtained from an RGB-D camera to successfully develop an autonomous flight control system. Besides, [18] proposed a navigation framework that enables small-scale UAVs to navigate autonomously in foliage environment using a 2D laser range finder. The framework contains real-time onboard motion estimation and trajectory smoothing. Despite their accuracy and precision, these solutions use complex algorithms in order to operate and result in heavyweight platforms which make them inconvenient for daily usage. Consequently, other lightweight solutions were explored; hence, autonomous navigation of aerial vehicles throughout onboard Android smartphones were recently employed and developed [13], [19], [20], [21]. These communication devices are equipped with sensors and IMUs essential for estimating flight condition data and orientation of the quadrotor. Moreover, they have computational power (processor) useful in implementing flight control algorithms. Some approaches used supplementary computer besides the smartphone, as in [19], where it used wireless networks to communicate with a PC. Using this connection, data could be transmitted from the smartphone to the PC, for instance live images or log data. It is also used to send commands from the PC to the quadcopter, for example angle references. Reference [13], unlike other works, proposed a different approach in which the smartphone has the control, guidance and navigation systems implemented inside, depending on the information provided by the onboard sensors. Twitter account was used to enable the drone to share information and receive commands in multi-agent system. The solution in this paper proposes a novel methodology of achieving autonomous aerial navigation operations of a quadrotor with very simple, lightweight, customizable and inexpensive design framework using a combination of onboard GPS/IMU

Fig. 1: System Overview

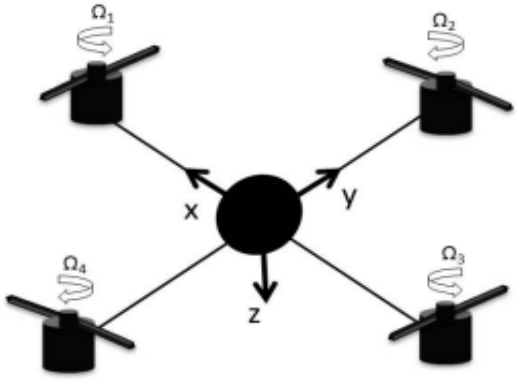


Fig. 2: Quadrotor configuration

and GSM/GPRS technology. Furthermore, the proposed design allows the autonomous control of a quadcopter using the developed application. Fig. 1 illustrates the design of the prototype proposed. It consists of two main components; the quadcopter and the smartphone. The latter enables the operator to select the waypoint he wishes the drone to hover to. This is accomplished via a developed android-based application which uses Google Maps APIs. Commands are sent to a web server via Wi-Fi. The received data will be stored in a database, for future usage. The web server ensures the bidirectional communication between the two ends. As a second phase, the server forwards the data to the Arduino microcontroller mounted on the drone. The latter responds back with the requested information using a GPRS, and the server, in its turn, directs it to the android device. The design criteria considered during the implementation of the proposed application are: flight durability and stability, ease of implementation and maintenance, usability, customizability and cost effectiveness with long wireless communication range. The paper is organized as follows. Section II illustrates the technical background. Section III discusses the building methodology of the quadrotor along with a step-by-step process of achieving a successful navigation operation. Section IV presents the implementation and testing of the software module. Section V shows the Communication Operation. Section VI introduces the results and the analysis. Constraints are listed in Section VII. Section VIII conducts the conclusion.

II. Technical Background

This section discusses some of the important technical concepts in which this project was based upon.

A. Drones

Unmanned Aerial Vehicles (UAV), commonly known as drones are aircrafts without a human pilot aboard. Controlling UAVs flight can be achieved in different ways such as using autonomous control, using onboard computers or by using a remote control. Remotely controlled UAV are the most common type of drones, but in the last few years autonomous UAV applications and functionalities increased especially after the introduction of flight controllers. A Flight Controller (FC) is a small circuit board of varying complexity. Its function is to direct the Round per Minute (RPM) of each motor in response to an input. A command from the pilot of the multi-rotor is fed into the flight controller, which determines how to manipulate the motors accordingly [1]. Most of the flight controllers employ sensors to supplement their calculations. These sensors could be a Gyroscope for orientation, a Barometer to hold altitudes, or a GPS for auto-pilot and fail-safe purposes. The first drone ever created was an autonomous torpedo that drops and explodes at a preset time, made by Dayton-Wright Airplane Company during the World War 1. As of late 1980's, usage of drones expanded after they have been proven efficient in surveillance and monitoring purposes. Nowadays, drones' popularity is increasing rapidly. In civil applications, drones can be used in many fields such as policing, surveillance and monitoring, rescuing, and inspection. Cinematography is one of the most common applications of drones nowadays. Companies like DJI and Parrot manufacture special types of multi-rotors drones that can take very high-quality videos and pictures. Furthermore, drones nowadays are being used in delivery fields. Multi-rotors are compact aircraft that contains more than two rotors. This type of drones is the most common between civilians due to its low price, customizability, ease of use and small size. Some of the multi-rotor aircraft are the Tricopter which has 3 motors, a Quadcopter with 4 motors and Hexacopter with 6 motors. Their main advantage compared to other drones is that it can be easily assembled and used in different fields like cinematography, surveillance, and search operations. In the past years, many countries have adapted to this technology and made some rules and regulation regarding flying multi-rotors.

Quadrotor flight dynamics: Many literature surveys addressed the topic of quadrotor flight dynamics [22]. This nonlinear dynamic model is necessary for flight stability and for accurate control. The quadrotor has four motors, each two diagonal rotors spin in the same direction and against the direction of the adjacent rotors [23], as shown in Fig. 2.

The model has six degrees of freedom, thus six distinct variables express its position in the space. The variables [x, y

[1] https://en.wikipedia.org/wiki/Unmanned_aerial_vehicle

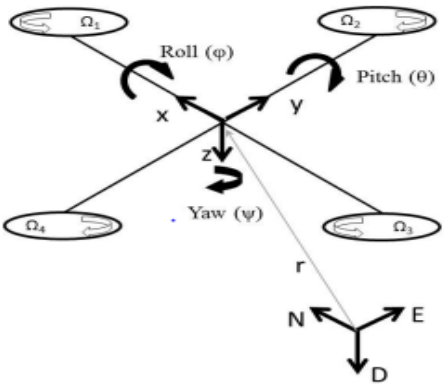


Fig. 3: Euler angles in quadrotors

, z] represent the distances of the quadrotor's center of mass along the x, y, and z axes, respectively. Whereas, the variables $[\Phi, \Theta, \Psi]$ denote the orientation of the quadrotor. They are known as Roll, Pitch, and Yaw, respectively [23], [24]. Fig. 3 clearly explains the Euler angles.

The most common type of multi-rotors is the Quadcopter or "Quad". The Quad has four motors, each two diagonal rotors spin in the same direction and against the direction of the rotors next to them. When all the rotors spin in equal magnitudes in addition to the airfoil blades lift is generated, and due to the symmetrical and unique shape of the Quad thrust is achieved by the Roll and Pitch. The Roll and Pitch can be achieved by increasing the rotational speed of two neighbor rotors. As they spin in deferent directions they negate the reaction of the rotating blades, and because their lift is higher than the other two rotors they cause the Quad to tilt applying some force sideways and creating the thrust. When increasing the speed of two diagonal rotors the reaction from rotating the blades in a certain direction increases, leading to the increase in the reaction of the blades on the Quad, making the Quad rotate to the opposite direction thus creating the Yaw. Fig. 4 demonstrates how to apply the flight mechanisms on a Quadcopter.

B. GSM/GPRS

Global System for Mobile communications (GSM) is an open, digital cellular technology used for transmitting mobile voice and data services[2]. It uses the frequencies 900 MHz and 1.8 GHz in Europe, whereas in the US it uses 850 MHz and 1.9 GHz. GSM supports voice calls, transmission of Short Message Service (SMS), and data transfer speed up to 9.6 kbps. GSM was designed to be a secure wireless system which relied on user authentication using a pre-shared key, challenge-response, and over-the-air encryption. General Packet Radio Service (GPRS), sometimes referred to as 2.5G, is used by GSM for data transmission, it offers a throughput rate up to 40 kbps, enabling mobiles to access online services in a reasonable speed at a rate similar to a dial-up modem. GPRS is a best-effort service, which means that the throughput and latency that depend on the number of other users sharing the service concurrently. GPRS supports the Internet protocol (IP), and the Point-to-point protocol (PPP).

[2]http://www.gsma.com/aboutus/gsm-technology/gsm

III. Building Methodology and Navigation Operation

A. Building Methodology

1) Hardware components: Now-a-days, plenty of companies are producing affordable easy to fly and ready to use drone models. Nevertheless, a DIY approach of building a quadrotor is more favorable as it guarantees the hardware customizability; it facilitates regular upgrades if desired and further enhancements when needed. Fig. 5 demonstrates the general structure of the quadcopter constructed. The main mechanical components necessary for building the drone are:

Body frame: The DJI F450 body frame kit was used. It consists of four arms and an average sized center plate made of carbon fiber, as shown in Fig. 6. Lately, due to its structural rigidity and light weight, the demand for carbon fiber composites have increased. The frame holds all the components together through the arms and the center plate. The latter is usually reserved for both flight controller and battery. Its robustness and resistance to crashes are the main reasons behind its selection.

Propellers: Two types of propellers exist; carbon fiber propellers, and plastic propellers. Due to their stiffness, carbon fiber propellers produce less vibrations and a lower propeller spinning noise; however, they are twice as expensive as the regular plastic propeller. Quadrotors use two clockwise (CW) and two counter-clockwise (CCW) propellers. While designing a quadrotor, various factors should be considered; flight usage, payload of the quadrotor and the KV. KV is the number of revolutions per minute that the motor will turn when 1 V is applied. Large propellers are mainly used in two fields, the cinematography and weight carrying. On the other hand, smaller propellers are mainly used in drone racing. In the proposed design, the 10x45 large plastic propellers were used due to the stability they provide to the drone in addition to their low price and proven efficiency.

Electric motors: Thrust is the main aspect when it comes to choosing motors. Thrust needed per motor can be calculated theoretically by the following equation:

$$thrust = \frac{\text{total weight} \times 2}{\#\ \text{motors}}$$

The estimated total weight of this mechanical design is calculated as follows:

$$\begin{aligned}\text{battery (450 g)} + \text{APM2.8 flight controller (20 g)} \ + \\ \text{body frame (280 g)} \ + \ \text{motor (50} \times \text{4g)}\end{aligned} \quad (1)$$

Hence, the minimum thrust required per motor is equivalent to (11302)/ 4 = 565 g thrust.

As a result, the DJI 2212 - 920KV brushless motor was used. Placing the motors and the propellers at equal distances, provides both, best performance and simplest control algorithms.

The electrical components required in quadrotors are:

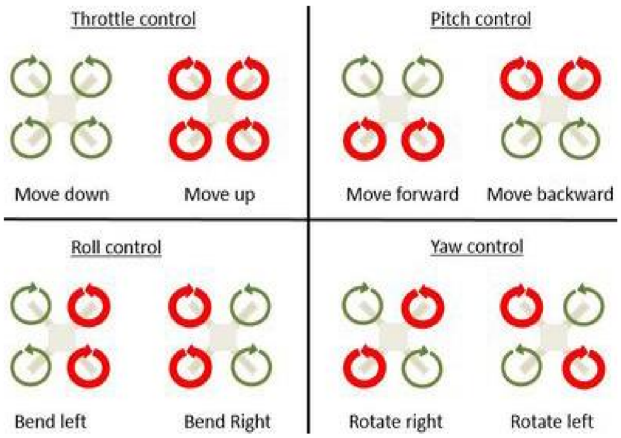


Fig. 4: Flight mechanisms of a Quadcopter

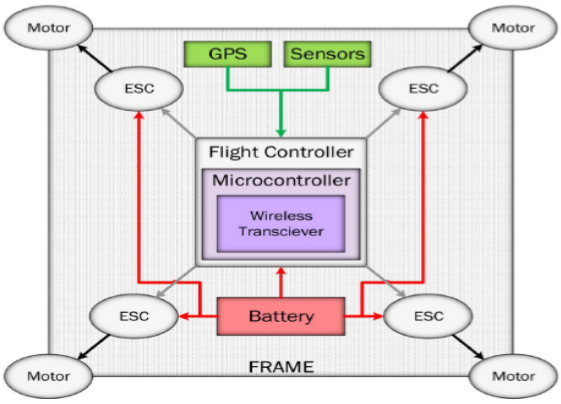


Fig. 5: Quadrotor mechanical design

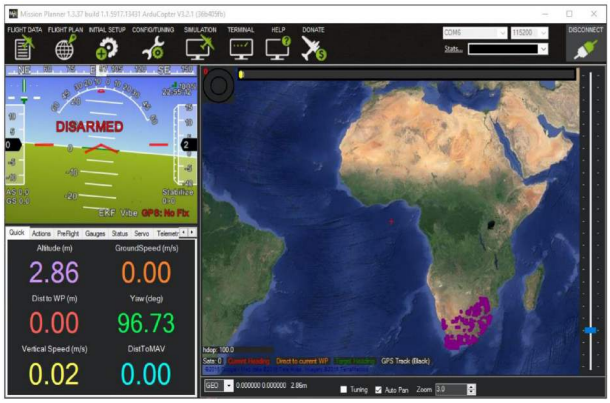


Fig. 7: Live data of APM through Mission Planner

Fig. 6: DJI F450 Body Frame

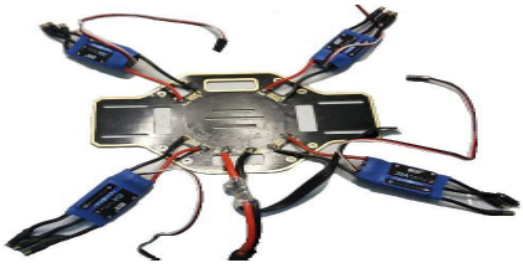


Fig. 8: Lower plate of the drone

Electronic Speed Control (ESC): The role of an ESC is to control and regulate the speed of a motor. In this design, it must be able to provide a maximum current of 30 A. The Simonk 30A ESC was used thanks to its reliability and low cost. Quadcopters use four motors each is connected to one ESC. Flight controllers send signals to the ESC, which drive the brushless motor. Battery Elimination Circuit (BEC) is usually built within the ESC, which is a voltage regulator to convert the Li-Po battery voltage (2S-7.4V, 3S-11.1V, 4S-14.8, etc.) to a smaller voltage like 5V.

Controller board: It contains different controllers

Flight controller: ArduPilot Mega (APM) 2.8 flight controller was used for a number of reasons. In fact, it is programmable, easy to use, fair price and has all the sensors required to attain a stable flight. In order to program it, a software called Mission Planner (AP) was used. Mission planner provides the ability to get live data midflight, as shown in Fig. 7. Flight controller continuously reads inputs from sensors, and changes the speed of the motors based on these inputs.

Microcontroller: The Arduino Uno microcontroller was used due to its low price, low power consumption, and simplic-

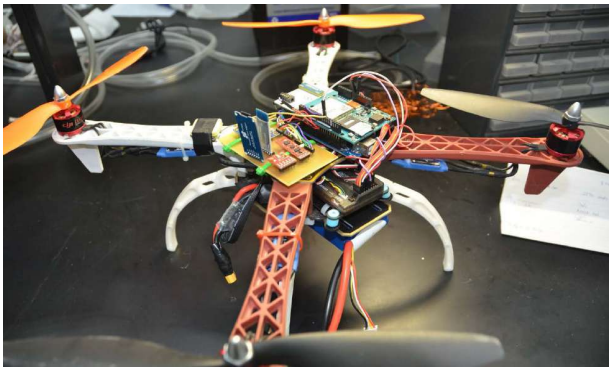


Fig. 9: The Arduino connected to the HC05 on the drone with APM on top

ity with regard to programming. Furthermore, its clock speed fits the developed prototype. Moreover, it supports different modules of sensors, actuators, and transceivers, in addition to shields. Shields are circuits which can go on top of the Arduino Uno. They can provide different utilities from FM transceivers to Wi-Fi modems.

Battery: It is the only power supply for the designed quadrotor unless an ESC without built-in BEC was used. In such systems, another 5V power supply is needed to supply the flight board and the motors. LiPo batteries consist of cells, each cell has a nominal voltage value of 3.7V. For example, a 4S LiPo battery indicates that we have four cells in series providing a voltage of 4x3.7=14.8 V. The HRB 3S LiPo battery, with a capacity of 6000 mAh, was used because it satisfies the design requirements in terms of weight and flight duration in addition to its reasonable price.

2) Components assembly and testing: Component testing approach was performed. Upon positive functionality of each and every unit, integration phase was next; hardware parts were combined together incrementally. Fig. 8 shows the ultimate structure of the lower plate of the quadrotor after hardware assembly. The power module (voltage regulator) is connected to the battery which powers the flight controller and the ESCs. The actual battery is placed between the upper and lower body plates. The flight controller is connected to the rest of the components in the upper plate, as Fig. 9 illustrates. Afterwards, the flight of the quadrotor built was tested with a remote control. The flights were very stable with smooth maneuvering.

3) Testing a functional quadrotor: A ready to fly drone was used in order to get hands on knowledge on flight mechanisms and a thorough understanding on the behavior of the remote control in conjunction with the sort of responses it sends to the flight controller.

Initially, the Remote Control (RC) receiver was connected to the oscilloscope. The responses gained from the receiver designated the start and the end of each channel used. As more and more tests were performed, more channels were connected, simultaneously, as shown in Fig. 10. Results demonstrated that two consecutive connected channels do not interfere with one another and each channel had the same period. Channels provided 1.5 ms width pulses at normal state, 1.8 ms width

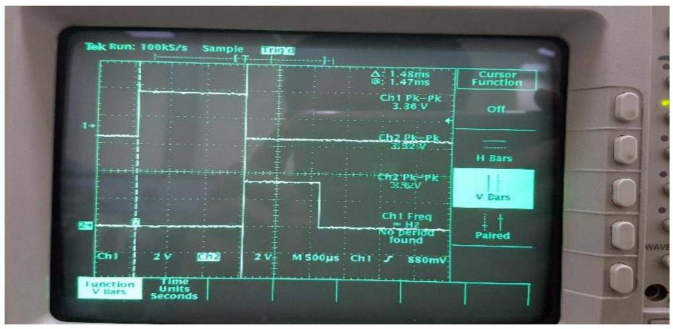


Fig. 10: RC signals

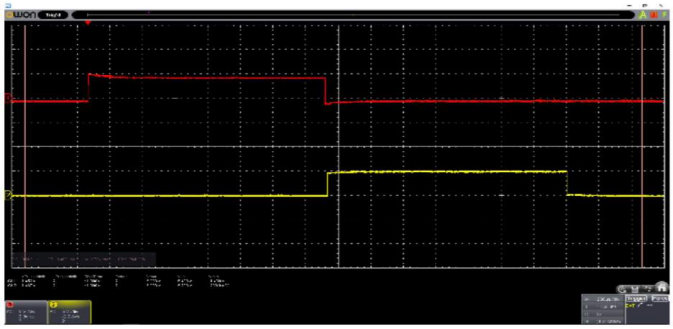


Fig. 11: Signals produced by Arduino Mega

pulses at high state and 1.2 ms width at low state. Channels could be ordered ascendingly (in terms of pulses), as follows: Channel 1, Channel 3, Channel 2, then Channels 4 - 8. Remote controls differ from one another by the number of channels they support. A channel corresponds to one command only. The RC used, in the previous test, had eight channels; four for the main movement directions and four extra channels designated for operator's usage. The other channels could be used to control mounted cameras or gimbal systems.

4) Imitating RC receiver's behaviour: Following the integration of all hardware parts, the constructed quadrotor was tested via a remote control. Results showed stable flights with smooth maneuvering. Next, the automation process started by imitating the behaviour obtained from the RC receiver. This was accomplished using Arduino Mega Pulse Width Modulation (PWM) pins. The signals obtained showed similar behaviour as the signals produced using the RC receiver, as can be seen in Fig. 11 below.

Testing the automation of the key operations, take-off and landing, came afterwards. This was accomplished by connecting channel #3 to the Mega and leaving all other channels connected to the RC receiver, for safety reasons. This test allowed to achieve the first semi-autonomous flight.

5) GSM/GPRS Transceiver: Arduino GSM Shield module enables the Arduino board to make phone calls, send/receive SMS messages and to connect to the internet over the GPRS wireless network. This is done thanks to the GSM library included in Arduino IDE. The GSM shield is used thanks to its compatibility with Arduino board and its ease of use and implementation. In order to successfully operate, the shield has

to be, simply, plugged onto the board. Similarly, a SIM card, offering a GPRS coverage, has to be plugged in.

6) GPS installation: In order to allocate the drone, a Global Positioning System module (GPS) has to be connected to the Arduino Mega microcontroller. For this purpose, the NEO-6M GPS module is used; a module with high accuracy and low power consumption, in chorus. Moreover, it is a basic module that is compatible with most microcontrollers. In this proposed design, the margin of error should not exceed 3 meters, this range of error is acceptable because of the size of streets and turns in the roads.

7) IMU installation: Knowing the direction and the orientation of the drone is essential while navigating. As a result, the Integrated Measurement Unit (IMU), manufactured by SparkFun, was one of the solutions adopted. It contains the following [25]:

- 3-axis Magnetometer sensor which provides information about its location in space and the direction of the drone.
- 3-axis Accelerometer sensor to measure the speed of the drone.
- 3-axis Gyroscope sensor which measures the tilting angle. Also, it can be used to add more stability to the quadrotor.

The use of the 3-axis magnetometer sensor is crucial as the navigation operation requires sensing angle difference across points. In order to conduct accurate navigation, it is required that the starting angle is considered the reference angle. Left and right rotations depend on the readings of the sensor. They alter the heading of the quadrotor through the Yaw angle. For instance, if the net angle is less than 180 degrees then the drone rotates to the left otherwise, it rotates to the right [25].

As an alternative for the ultrasonic sensor, the pressure sensor "barometer" was used. The barometer takes the initial location pressure as a reference and continuously measure the pressure along the journey to estimate the attitude from pressure reading. The pressure of the altitude is measured using the following equations [26]

$$h_{alt} = (1 - (\frac{P_{sia}}{1013.25})^{(0.190284)}) \times 145366.45$$

$$h_m = 0.3048 \times h_{alt}$$

where,

h_{alt} is the altitude measured in [ft]

P is pressure measured in [mb]

h_m in the altitude measured in [m]

IV. Mobile application design

A. Test Application 1

This application was developed in order to make sure that Google services such as Google Maps can be successfully imported and used. The application renders a map view, gets the phone's current location and zooms in on it. In addition, few preset markers are present which can be placed and removed by clicking the corresponding buttons. Besides, there is a slider to change the mode og the map view.

B. Test Application 2

This application was developed to test the Google Directions API. The application starts by rendering a map where the operator is prompted to set two markers. Afterwards, the shortest navigational path between the two points is computed and, consequently, a list of latitude and longitude coordinates is returned. Then, a polyline, representing the navigational route, is displayed as well as the list of coordinates and the total number of points alongside the path.

C. Final Application

The prototype is designed to work on Android device that has Android 4.1.0 (Jelly Bean) or higher, this is due the lack of some features in earlier versions of Android. The core functionality of the main application is as follows: once the mobile application is launched, the user is prompted to enter his login credentials; username and password. The input data is validated against the database and accordingly, five distinct cases could possibly emerge:

- Case 1 - "Login Successful": the database grants the username and password combination. Hence, the user is directed immediately to a google map page.
- Case 2 - "Missing Credentials": this occurs when the user forgets to enter either the username or the password.
- Case 3 - "Incorrect Combination": this happens when the user enters a wrong username or password.
- Case 4 - "Connection Error": this denotes that the database is currently inaccessible. Reasons could be an internet connection failure or a momentary server failure.
- Case 5 - "Drone Busy": this means that the requested drone is currently in use by another user. Thus, the user is informed that he is able to select the drone only when its status changes to "available".

Fig. 12 is a flow chart showing the sequence of the android application. Upon granted access, the operator can easily perceive the label on the map. It denotes the actual location of the quadrotor. The user picks a point of interest by clicking on the map. It is the drone's flight destination. Consequently, a method, which is responsible of creating requests to Google's servers, is invoked. The method considers the current location of the drone as the starting point; whereas, the selected point as the ending point.

Consequently, the server returns an array of points representing the latitude and longitude coordinates of the route which leads to the user's point of interest. Besides, a polyline of the navigation path is drawn on the map.

The quadcopter would not fly until the user clicks on the "Start" button. Once done, the android application sends a

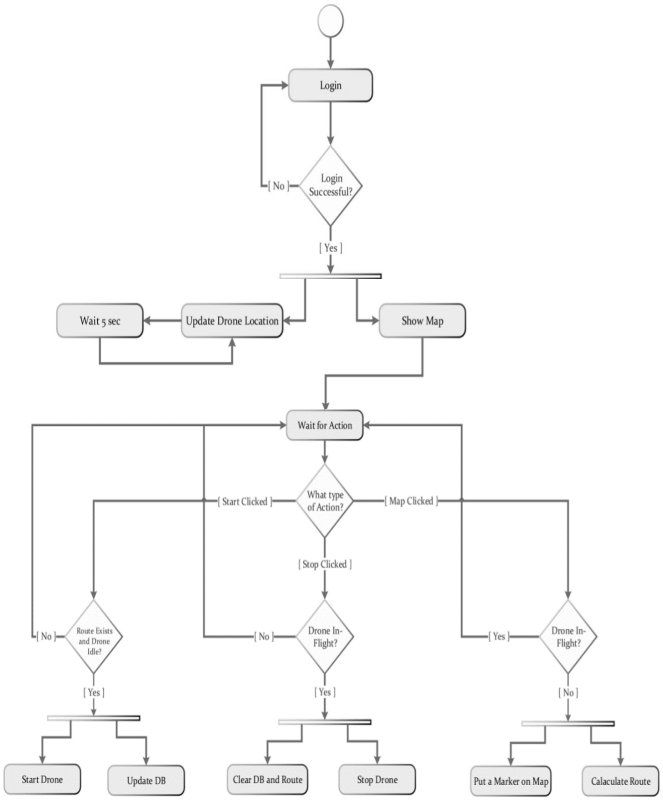


Fig. 12: Android application activity diagram

request to the database with all the required points in order for them to be saved and read later on for navigation.

The user is not allowed to request a new navigation path until the actual flight comes to an end or a stop has been issued. The stop button terminates all sorts of activities the drone is actively performing. As a result, the latter's status is set to "available", hence it can conduct new flights.

In the case where the drone is idle, for a period of time, its status alters to "available", thus other users would be able to connect to it and maneuver it. In the case where the user issues commands to a drone already in use, the database would issue a time out and the user would be directed to the home page (login page).

V. Communication Operation

Establishing a successful connectivity between the software and the hardware modules was the next milestone to reach. As mentioned earlier, the web server is the communication tool that ensures data exhange between the two parties. Participants in this operation are the:

- database
- web server
- quadrotor
- operator

Subsections below demonstrate, in details, the steps followed to design, test and implement the communication operation.

A. Database

Fisrt, an IP address along with a domain name were reserved to the database. In addition, a server, linked to the database IP address, was reserved. Hosting24 services were used the help accelerate the process. Second, a demo database was created to test the connection between the database and the web server. Eventually, the server was able to successfully execute SQL operations like SELECT, UPDATE, and INSERT. The third step was the database design. Fig. 13 below illustrates the entity relation diagram of the database. It is a one-to-one relationship between both the drone and the smartphone entities. In fact, the smartpone navigates the drone via commands and a list of coordinates (longitudes, latitudes). Moreover, the smartphone identifies the drone through a unique ID and it is authenticated through password. The drone responds by the GPS coordinates of its actual location.

B. Web Server

The web server receives HTTP requests, from both the smartphone and the drone. Firstly, a reserved IP address and a domain name were allocated to the server. Secondly, simple codes were developed to test the connectivity with the database and to ensure that simple SQL queries are operational. Thirdly, the different scenarios that could happen, during communication, were established and identified along with the data to be exchanged. In fact, both the drone and the Android-based application send the required data values through URL Special Characters encoding which, can be embedded in the HTTP request of the page.

The data exchanged, at the web server level, can be seen from two perspectives; smartphone perspective and quadrotor perspective.

C. Smartphone Perspective

User log in:

In order to establish connection with the quadrotor, the user needs to be authenticated. The operator transmits, to the server, the following: drone ID, password, and a unique user identifier (i.e. MAC Address). The response will be one of the following options:

1) Access granted, followed by the drone current location
2) Invalid data / Missing credentials
3) Connection error has occurred
4) Incorrect ID/Password combination
5) Drone is busy
6) Update Error

Fig. 14 shows the activity diagram for user login.

D. Quadrotor Perspective

The drone will transmit its current location. The response will be one of the following options:

- 1: To stop the drone
- 2 - 4: Reserved for different user commands, followed by a list of waypoints the drone has to follow)

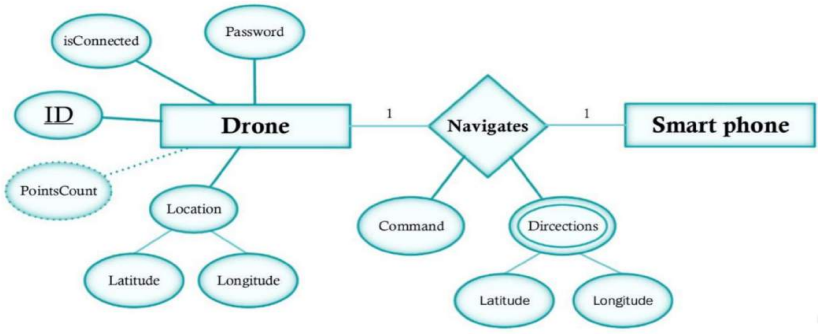


Fig. 13: ER Diagram of the Database

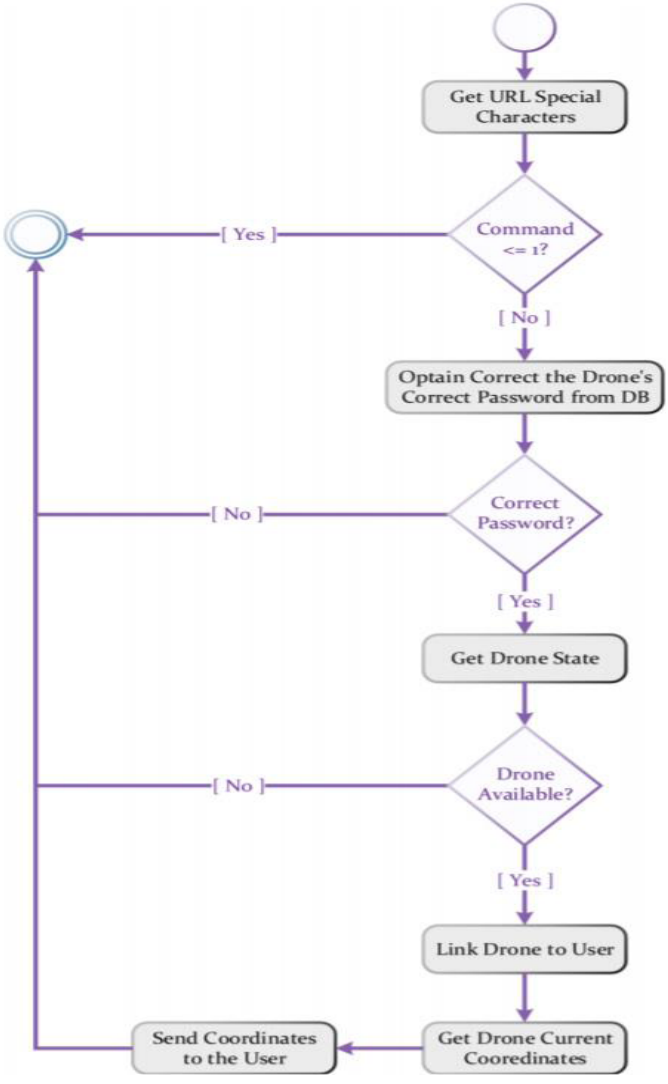


Fig. 14: Activity diagram for user login

- 5: Connection error has occurred
- 6: Invalid transmitted data / Missing credentials

Fig. 15 presents the activity diagram of the web server.

E. Controlling the Drone

Once the connection with the quadrotor is established, the monitoring of the drone, by the user, is now enabled. The response will be as follows:

- 1: Update successful
- -1: Invalid data / Missing credentials
- -2: Connection error has occurred
- -3: Timeout (the user has been leaving the drone idle for a while)
- -4: Duplicate entries (already exist)
- -5: INSERT Error
- -6: UPDATE Error
- -7: Inconsistent data

The final step was the creation of web pages content. PHP language was used as it is HTML-embedded and it supports database operations. Fig. 16, below, shows the activity diagram.

F. Drone (Arduino)

In order to develop the communication system, at Arduino level, the first step was to make sure that the Arduino is connected to the service provider. The simplest way to test the GSM connectivity is by sending and receiving an SMS message. The second step was to get connected to GPRS services. In order to do this, three main attributes were required:

- service provider Access Point Name (APN)
- login
- password

The last step was to communicate with the web server by sending the current location of the quadrotor, and receiving the commands and the direction.

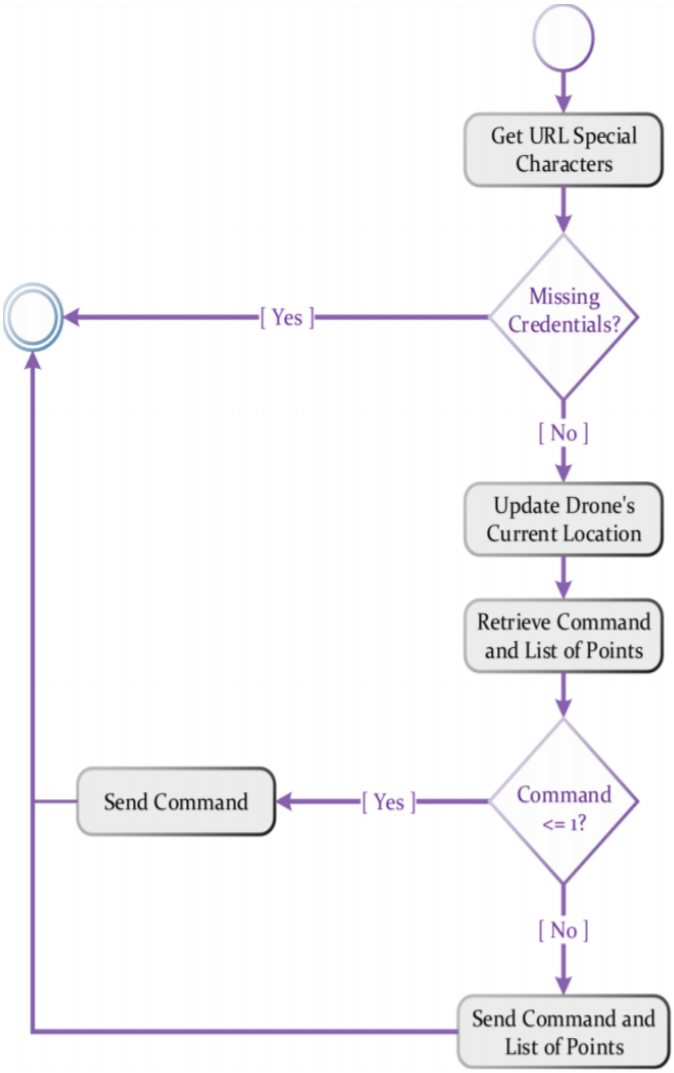


Fig. 15: Drone.php activity diagram

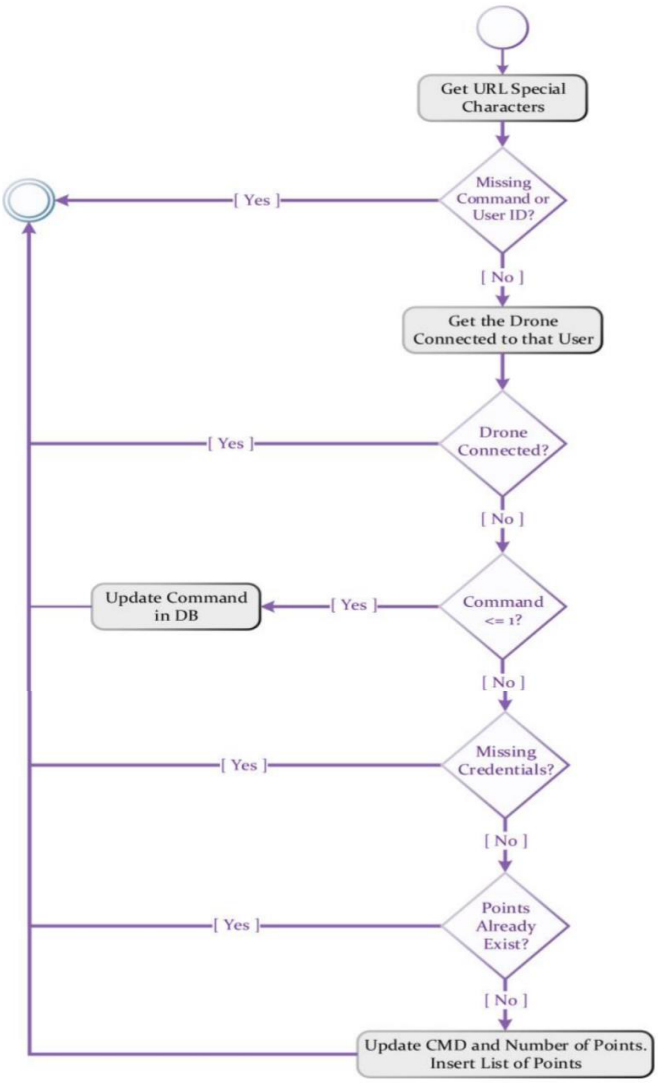


Fig. 16: CMD.php activity diagram

G. User (Android)

In order to develop and establish the communication between the android and the server, the first step was to obtain permissions to access internet from the android device. Second, was testing the HTTP libraries followed by testing the connection with the server.

H. Overall System

At a final stage, the prototype was assembled together and tested as a whole.

VI. Results and Analysis

Experiments and analysis demonstrated that the proposed prototype is fruitfully accomplished. In fact, data transmission between android application and Arduino board is successful and consequently, the quadcopter is effectively monitored via the smartphone app. While the prototype is fully functional and meets the design criteria set, it has some limitations and shortcomings which can be considered in further developments. In fact, the quadrotor uses a battery as power supply and will not be operational once the battery is dead. Hence, users will need to recharge the battery before each use of the drone. Besides, the flight time of the quadcopter is limited by the payload lifted and the battery size. Furthermore, any sort of failure in the web server will cut off the communication between the drone and the smart phone. Moreover, the GPS reading is not accurate in tight streets which affects the efficiency of the navigation. In addition, the design is fragile against specific weather conditions such as rain, wind, snow, etc. And it does not avoid obstacles. Also, the drone has to be directed to the north upon starting. Likewise, quality of communication is proportional to the quality of the service provider. Similarly, the coverage area of the communications depends on the coverage range allowed by the service provider.

VII. List of Constraints

Some of the limitations of this design would be that the Quadcopter runs on a battery and will not work once the charge of the battery runs out. Users will have to recharge the

battery before they can use it again. Also, the Quadcopter's flight duration is restricted by the battery size and the weight it carries. The quality of the communication depends on the quality of the service provider and the coverage area of the communications depends on the service provider. The Quadcopter must not fly over a No-Fly Zones like airports or military facilities, which constrains the Quad to fly over streets only. The communication between the Quadcopter and the smart phone is done through a Web server, and any fault in the server will cut off the communication between them. For the GPS, its accuracy could make some problems in tight streets, affecting the efficiency of the navigation. Bad weather conditions affect the Quadcopter usage.

VIII. Conclusion

In summary, autonomous drones are gaining more and more attention by highbrows and researchers. As a result, their usage will only grow with time. This article attempted at presenting the design and implementation of a simple, customizable and cost effective smart solution of a remotely monitored GPS quadrotor through an Android-based mobile platform. The user needs to be successfully logged-in in order to establish connection with the quadrotor and start maneuvering operations. Consequently, controlling the drone is enabled for the operator. In addition, the quadrotor sends information and receives instructions through the GSM/GPRS module. This work can be employed eminently in multiple and diverse quadrotor applications as it should open the door to improved flight control and autonomous hover in the near future.

References

[1] J. G. Leishman, "The breguet-richet quad-rotor helicopter of 1907," *National Academies of Sciences, Engineering, and Medicine*, vol. 47, pp. 58–60, 2001.

[2] S. Spooner, "A successful french helicopter," *Flight*, vol. XVI, p. 47, 1924.

[3] "A successful french helicopter," *Flight*, p. 247, 1960.

[4] I. H. Culver, T. F. Hanson, and L. G. Look, "Helicopter rotor system," Patent US3 261 407A, 08 05, 1964.

[5] M. Bassem and T. Bonny, "Iov road safety: Vehicle speed limiting system," *IEEE International Conference On Communications, Signal Processing, and Their Applications*, 2019.

[6] M. Gao, X. Xu, Y. Klinger, J. van der Woerd, and P. Tapponnier, "High-resolution mapping based on an unmanned aerial vehicle (uav) to capture paleoseismic offsets along the altyn-tagh fault, china," 2016.

[7] M. Fuhrmann and M. C. Horowitz, "Droning on: Explaining the proliferation of unmanned aerial vehicles," 2017.

[8] R. Luppicini and A. So, "A technoethical review of commercial drone use in the context of governance, ethics, and privacy," 2016.

[9] M. A. Z. Talal Bonny and K. N. Salama, "An adaptive hybrid multiprocessor technique for bioinformatics sequence alignment," *The 5th Cairo International Conference on Biomedical Engineering Conference, (CIBEC'10)*, 2010.

[10] A. Allagui, A. E. Rojas, T. Bonny, A. Elwakil, and M. A. A. Kareem, "Nonlinear time-series analysis of current signal in cathodic contact glow discharge electrolysis," *Journal of Applied Physics,*, 2016.

[11] T. Bonny and A. S. Elwakil, "Fpga realizations of high speed switching-type chaotic oscillators using compact vhdl codes," *Journal of Nonlinear Dynamics*, 2018.

[12] T. Bonny and S. Henno, "Image edge detectors under different noise levels with fpga implementations," *Journal of Circuits, Systems and Computers*, 2018.

[13] M. Hayajneh, M. Melega, and L. Marconi, "Design of autonomous smartphone based quadrotor and implementation of navigation and guidance," *Elsevier*, 2016.

[14] M. Achtelik, M. Achtelik, S. Weiss, and R. Siegwart, "Onboard imu and monocular vision based control for mavs in unknown in- and outdoor environments," *IEEE*, 2011.

[15] haowu Yang, S. A. Scherer, and A. Zell, "Visual slam for autonomous mavs with dual cameras," 2014.

[16] F. Samadzadegan and G. Abdi, "Autonomous navigation of unmanned aerial vehicles based on multi-sensor data fusion," *IEEE*, 2012.

[17] R. G. Valenti, I. Dryanovski, C. Jaramillo, D. P. Ström, and J. Xiaoi, "Autonomous quadrotor flight using onboard rgb-d visual odometry," *IEEE*, 2014.

[18] J. Q. Cui, S. Lai, X. Dong, P. Liu, B. M. Chen, and T. H. Leei, "Autonomous navigation of uav in forest," *IEEE*, 2014.

[19] A. Bjalemark and H. Bergkvist, "Quadcopter control using android based sensing," *Advances in Electrical and Computer Engineering*, 2014.

[20] G. Loianno, G. Cross, C. Qu, Y. Mulgaonkar, J. A. Hesch, and V. Kumar, "Flying smartphones: Automated flight enabled by consumer electronics," *IEEE*, 2015.

[21] G. Loianno, Y. Mulgaonkar, C. Brunner, D. Ahuja, A. Ramanandan, M. Chari, and S. Diaz, "Smartphones power flying robots," *IEEE*, 2015.

[22] N.-S. Pai, W.-C. Li, M.-H. Chou, and P.-Y. Chen, "Flight control for a quadrotor of attitude control based on android system and using optimal-tuning design," *elsevier*, vol. 54, pp. 170–184, 2016.

[23] G. M. Hoffmann, H. Huang, S. L. Waslander, and C. J. Tomlin, "Quadrotor helicopter flight dynamics and control: Theory and experiment," 2007.

[24] S. Swarnkar, H. Parwana, M. Kothar, and A. Abhishek, "Development of flight dynamics model and control of biplane-quadrotor uav," 2018.

[25] J. E. Kain and C. Yates, "Airborne imaging system using global positioning system (gps) and inertial measurement unit (imu) data," Patent US 005 894 323A, 03 22, 1996. [Online]. Available: https://patents.google.com/patent/US5894323A/en

[26] D. Burch, *The Barometer Handbook: A Modern Look at Barometers and Applications of Barometric Pressure*, T. Burch, Ed. Starpath Publications, 2009.

SQL to SPARQL Conversion for Direct RDF Querying

Ahmed ABATAL[1], Khadija Alaoui[2], Mohamed Bahaj[4]
Mathematics and Computer Science Department
Hassan I University, Faculty of Sciences and Techniques
Settat, Morocco

Larbi Alaoui[3]
Mathematics and Computer Science Department
International University of Rabat
Sala Al Jadida Morocco

***Abstract*—With the advances in native storage means of RDF data and associated querying capabilities using SPARQL, there is a need to let SQL users benefit from such capabilities for interoperability objectives and without any conversion of the RDF data into relational data. In this sense, this work present SQL2SPARQL4RDF an automatic conversion algorithm of SQL queries into SPARQL queries for querying RDF data, which extends the previously established algorithm with relevant SQL elements such as queries with INSERT, DELETE, GROUP BY and HAVING clauses. SQL users are provided with a relational schema of their RDF data against which they can formulate their SQL queries that are then converted into SPARQL equivalent ones with respect to the provided schema. This avoids the birding of translating instances and data replication and thus saving loading times and guaranteeing fast execution especially in the case of massive amounts of data. In addition, the automatic mapping framework developed by the java programming language, and implement many new mapping functionalities. Furthermore, to test and validate the efficiency of the mapping approach and adding a module for automatic execution and evaluation of the various obtained SPARQL queries on Allegrograph.**

***Keywords*—Resource Description Framework (RDF); Structured Query Language (SQL); Simple Protocol and RDF Query Language (SPARQL); schema mapping; query conversion; Allegrograph**

I. Introduction

The relational database (RDB) systems have been used as a standard for data management for many years and involve the development of various tools. However, in recent years, data management solutions based on the ontology language RDF have proven to be a well-suited alternative to relational databases for the storage and querying RDF data. In the last decade, the use of RDFs has indeed evolved considerably and the amount of RDF data has increased enormously. Since the standardization of RDF and its query language SPARQL ([6], [7], [8]), RDF has found a growing interest for its use in many application domains such as biology [3], health [11], geology [15], smart cities [12], etc. This interest has also been accompanied with the development of various tools for handling RDF data. One of the motivations behind this development is the need for database solutions that are independent of existing relational databases (RDB) technologies in order to handle massive amount of data produced and to tackle the problems related to the limitations of the RDB technologies. To be also noticed, is that due to the simplicity of RDF and its power of presenting data in a machine-readable format many attempts have been done to convert huge amounts of relational data into RDF (e.g., [1], [4]). This was done with the aim to benefit from the opportunities RDF provides for integration purposes and for linking data to make it accessible for the semantic web. The RDF data model allows data to be structured in graphical form, which ensures flexible navigation through the use of well-established graphical algorithms. These characteristics of RDF have encouraged the development of various technologies for managing and querying RDF data.
For all these reasons, the problem that raises itself is how to let RDB users interact with RDF data without any conversion of data into relational data in order to facilitate sharing of RDF data across applications. Solutions to such a problem will also make it possible to take advantage of the multiple RDF dedicated stores and their associated SPARQL capabilities (e.g. Sesame [13], CliqueSquare [16], 4store [17], SOR [18], RDF-3X [19], SHARD [21], ...), avoiding thus the use of relational stores and their associated problems such as lack of adequacy to support dynamic RDF schemas or to support large amounts of RDF data.

This work provide a solution to this interoperability problem that consists of an extension of the SQL2SPARQL framework established in [10] for RDF querying using SQL without any conversion of RDF data into relational data. The framework provides RDB users with a relational schema that let them query RDF data using SQL. The SQL queries are translated to SPARQL queries that will be executed directly on RDF dataset.The extension give aims at presenting a complete conversion solution with algorithms for the automatic exchange of data with an execution of queries directly on RDF data stored in native systems. Such a solution avoids thus an extra data store and associated problems such as synchronization of data changes between two data stores and loading times of RDF data into relational stores.
The rest of the paper is organized as follows: Section II, he gives an overview of the existing mapping algorithms from SQL to SPARQL. Section III is devoted to the extension SQL2SPARQL4RDF of the SQL2SPARQL framework where various mapping algorithms are presented. Section IV deals with the implementation and tests of the framework for the validation of mapping algorithms. Section V concludes this work.

II. Related Works - SQL2SPARQL Conversion Framework

This section describes the SQL2SPARQL in the work [10] conversion framework and other related works. For the advantages of such a framework over other conversion techniques (e. g. [14],[20]), this work refer to the previous work in [10]. The SQL2SPARQL framework consists in converting SQL queries to SPARQL equivalents, which can be executed on real RDF datasets. SQL queries are formulated against a relational schema that the framework extracts from the RDF data and provides to SQL users.

The steps involved in the relational schema extraction are discussed in the following subsection. Those related to the query conversion according to this schema are detailed in the subsequent subsection.

A. Proposed System Architecture

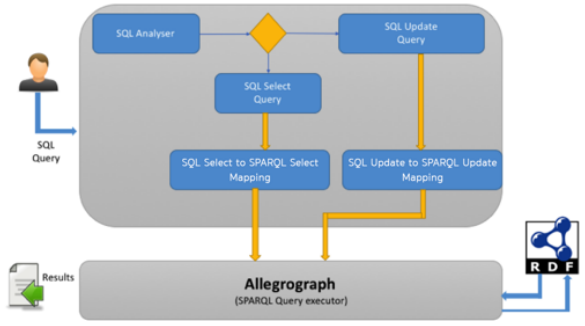


Fig. 1. System architecture

Fig. 1 present the proposed architecture that allows the user to give a complete SQL query and translates to equivalent SPARQL Query, in the first time SQL Analyzer that takes the SQL query as input, then it checks whether it is a SQL Select query or SQL update, after it passes the query to the corresponding component "SQL Select Query" or "SQL Update Query", to break down and extract the necessary elements like the Triple pattern or the attributes and the sending's to the component responsible for the translation of the requests via dedicating algorithms for the conversion then generates the final SAPRQL query to run it on "SPARQL Query Executor" the component that uses Allegrograph [2] and returns the result to User.

B. Relational Schema Extraction

The extraction of the relational schema in the SPARQL2SQL framework is provided by the schemaMapping()-algorithm given in Tables I and II, is based on the vertical partitioning work in[9]. It traverses all triplets and extracts unique predicates that are stored in variables to be used in the other algorithms of the framework. First, the property table method is used to group subjects with common predicates. In each table contains a subject as an attribute to identify the table and a collection of predicates attributes.

In addition, this work does not use the property table method, which allows us to think of a response with many fewer tables. The proposed solution reduces joint conditions and provides a more logical and efficient relational schema that makes it easier for users to query the schema.

TABLE I. Relational Schema Extraction / Part 1

Algorithm: Part 1 - schemaMapping()	
Input:	Bx : list of Abox
Output:	N: hashMap key is "relation name" and value is "predicate name"
1	N = "" {initialize N by null }
2	for i = 1 to Bx.size do
3	T= Bx[i] {extract a triple T from the Bx}
4	r = T.type extract type r from the triple T
5	p = T.predicate
6	extract predicate p from T
7	notExist = true
8	k = 1
9	while ((notExist) AND (j¡=N.size)) do
10	if ((r==N(k).getRelation() AND v==N(k).getPredicate())
11	notExist =false;
12	end if
13	end while
14	if (notExist)
15	N.put (r,v) { predivate p, and relation r to N }
16	end if
17	end for
18	Return N

TABLE II. Relational Schema Extraction / Part 2

Algorithm: Part 2 - schemaMapping()	
Input:	N: hashMap of Part-1
Output:	M : Relational schema
1	For each key v of N
2	Create a table M with an attribute SUBJECT
3	For each value k associated with v in N
4	add an associated attribute k to M
5	return M

C. Query Conversion

The conversion of SQL queries that are issued on the extracted relational schema into equivalent SPARQL queries is the task of the convertSqlQuery()-algorithm given in Table III. This algorithm accepts an SQL query as an input in string format and Map P as the output of the schemaMapping() algorithm, and produces an output in the form of a SPARQL query.

The algorithm analyzes the SQL query to extract the SELECT, SQwhereA and SQwhereB clauses, the corresponding SQwhereA has a clause where with join conditions and SQwhereB has a WHERE clause contains boolean conditions, the algorithm affects the NULL value in WHERE clause in the case this clause contains neither joins nor boolean conditions.

TABLE III. ALGORITHM FOR CONVERSION OF SQL QUERY

Algorithm: convertSqlQuery()	
Input	S :SQL query; N: hashMap
Output	SP :SPARQL Query
1	SP =""
2	QR=analyze(S) {analyze SQL query to obtain clauses }
3	Sselect= QR.getSelectClause()
4	SwhereA= QR.getWhereJC()
5	SwhereB= QR.getWhereBE()
6	SPselect="SELECT "
7	SPwhere="WHERE { "
8	T = "" { initialize T (Triple pattern) by NULL }
9	T = ConvSqlSelect(Sselect).getTP()
10	SPselect=ConvSqlSelect(Sselect)
11	SPwhere +=
12	ConvWhereSql(SwhereA, SwhereB,T,N)
13	SP= SPselect+ SPwhere+"}"
14	if QR.type!= null then
15	q1=QR.GetLeftSubSL()
16	q2=QR.getRightSubSL()
17	SP1=queryConvert(q1)
18	SP2=queryConvert(q2)
19	SP=combine(SP1, SP2, QR.type)
20	end if
21	Return SP

The convertSqlQuery() algorithm uses the subalgorithms ConvSqlSelect(), ConvSqlWhere() given in Tables VII and VIII. if the query contains UNION or INTERSECT, this query will be considered as two SQL queries. First of all, each one is converted by the convertSqlQuery() algorithm. The results of these queries are then grouped by the combinatorial algorithm () given in Table IV to give the final SPARQL query.

TABLE IV. ALGORITHM FOR CONVERSION OF A QUERY WITH COMBINATION OF TWO WHERE CLAUSES

Algorithm: Combine()	
Input:	query1 :SPARQL Query ; query2 :SPARQL Query ; type : {type is either INTERSECT, EXCEPT or UNION }
Output:	SP : SPARQL query
1	SP ="" initialize SP by empty
2	sprSelect= query1.getSelectClause()
3	sprWhere=" { ";
4	sprWhere1=" { "+query1. getSparqlWhere()+" }";
5	sprWhere2 = " { "+query2. getSparqlWhere ()+" }";
6	sprWhere +=sprWhere1+type+sprWhere2+" }"
7	SP += sprSelect+spWhere
8	return SP

Tables V and VI example shows a SQL query converting to SPARQL:

TABLE V. EXAMPLE FOR INTERSECT CONVERSION

SQL query:
(SELECT client.name from client) INTERSECT (SELECT supplier.name from supplier)
SPARQL query:

SELECT ?o0
WHERE {
{
?s0 ¡http://uhp.ac.ma/ontology/name¿ ?o0
}
INTERSECT{
?s0 ¡http://uhp.ac.ma/ontology/name¿ ?o0
}
}

TABLE VI. EXAMPLE FOR ORDER BY CONVERSION

SQL query:
SELECT client.name, order.date FROM client, order WHERE client.subject=order.client ORDER BY order.date
SPARQL query:

SELECT ?o0 ?o1
WHERE {
?s1 ¡http://uhp.ac.ma/ontology/client¿ ?s0.
?s0 ¡http://uhp.ac.ma/ontology/name¿ ?o0.
?s1 ¡http://uhp.ac.ma/ontology/date¿ ?o1.
}
ORDER BY ASC(?o1)

Conversion of the select clause

In the first version of the framework [10], the authors developed a "ConvSelectSql ()" algorithm that allows you to convert simple Select clauses into SPARQL Select. Now this framework has improved this algorithm in order to convert a complete Select clause into SPARQL.

The algorithm for the conversion of the select clause of the SQL query is the ConvSqlSelect() algorithm given in Table VII. It takes as input a SELECT clause, and gives as output a SPARQL SELECT clause with a triple pattern list TP, as mentioned in Table VII.

TABLE VII. ALGORITHM FOR CONVERSION OF THE SELECT CLAUSE

Algorithm: ConvSQLSelect()	
Input:	B: list of attributes of an SQL-Select query
Output:	SpSelect : SPARQL Select query ; T : list TP of triple patterns
1	SpSelect =""
2	T =""
3	for j = 1 to B.size do
4	r = B{j}.relation { Retrieve the relation r from attributes A}
5	p = B{j}.attribute { Retrieve the attribute p from A }
6	if p!='subject' then
7	SpSelect += "O" + j
8	tp={ ?Sj P Oj } { A triple pattern is constructed }
9	T.put (r, tp)
10	Else
11	SpSelect += "S " + j
12	tp={ ?Sj rdf:type r } { A triple pattern is constructed }
13	T.put (r, tp)
14	Endif
15	End for
16	Return SpSelect, T

Conversion of the where clause

The ConvSqlWhere() algorithm takes an SQL Where clause and the outputs of schema-mapping() and conSelect-Sql() algorithms to give an equivalent SPARQL clause. The algorithm is given in Table VIII and the addJCtoWhere() Method used by The ConvSqlWhere() algorithm is given in Table IX.

TABLE VIII. ALGORITHM FOR CONVERSION OF THE WHERE CLAUSE

	Algorithm: ConvSqlWhere()
Input:	LJC: List of Join conditions, LBE: List of boolean expressions,T : triple patterns
Output:	SpWhere :SPARQL WHERE clause
1	SpWhere = ""
2	if (LJC.isEmpty() && LBE.isEmpty()) then
3	for each t from T do
4	SpWhere += "?" + t.subject + " " + t.predicate +" ?" + t.object
5	end for
6	else if (! LJC.isEmpty) then
7	for each p from LJC do
8	lOp=p.LeftOperand;
9	rOp=p.RightOperand ;
10	SpWhere = addJCtoWhere(SpWhere,lOp,rOp);
11	end for
12	{if we have a Boolean conditions }
13	else if (! LBE.isEmpty) then
14	for each e from LBE do
15	lOp=e.LeftOperand;
16	t =T.get(lOp.relation)
17	SpWhere += "?" + t.subject + " " + t.predicat + "?" + t.object
18	End for
19	{ add FILTER in SpWhere }
20	SpWhere += "FILTER("
21	for each k from LBE do
22	lOp=k.LeftOperand;
23	rOp=k.RightOperand;
24	t =T.get(lOp.relation)
25	SpWhere += t.object+"" + e.operator + "" + rOp
26	End for
27	SpWhere+=+")"
28	Endif
29	Return SpWhere

TABLE IX. METHOD FOR ADDING JOIN CONDITIONS TO A CONVERTED WHERE-CLAUSE

	Algorithm: addJCtoWhere()
Input:	TP : Triple patterns, lOp: left operand ;rOp: right Operand
Output:	SpWhere : SPARQL WHERE clause
1	tr1 =TP.get(lOp.relation)
2	tr2 =TP.get(rOp.relation)
3	SpWhere += "?"+tr1.subject+""+ tr1.predicat + "?"+tr1.object
4	if lOp.relation = rOp.attribut then
5	SpWhere+="?"+tr1.subject+""+rOp.attribut+" ?"+tr2.subject + "."
5	+ " ?"+tr2.subject +""+rOp.attribut+" ?"+tr1.object+". "
6	elseif (lOp.attribut = rOp.relation) then
7	SpWhere+="?"+tr2.subject+" "+lOp.attribut+" ?" + tr2.subject + "."+
7	" ?"+tr1.subject+""+lOp.attribut+" ?"+tr2.object+". "
8	else
9	SpWhere+="?"+tr1.subject+" "+lOp.attribut+" ?"+tr1.object +
10	"."+ " ?"+tr2.subject +" "+ rOp.attribut+" ?"+tr1.object+". "
11	end if
12	return SpWhere

III. SQL2SPARQL4RDF: EXTENSION OF SQL2SPARQL

This work has improved the SQL2SPARQL framework to convert a complete SQL query of selection including all the clauses (group, classify by ...) and the functions of aggregate (MIN, MAX, AVG, COUNT), / in the second part this extension is able to convert INSERT and DELETE SQL queries into SPARQL INSERT and UPDATE queries, this part introduces pseudo-codes of algorithms

A. Conversion of Select Clauses for Aggregate Functions

The exConvSelectSql () algorithm is the improvement of the convSelectSql () algorithm so that it can detect and extract aggregate functions:

TABLE X. EXTENDED VERSION OF THE SELECT-CLAUSE CONVERSION ALGORITHM

	Algorithm: exConvSqlSelect()
Input:	N :The list of attributes of an SQL Select
Output:	SpSelect: SPARQL Select, TP: triple patterns
1	SpSelect =""
2	TP =""
4	for j = 1 to N.size do
5	r =N{j}.relation extract r from attributes N
6	p=Nj.attribute { extract p from N }
7	if p!='subject' then
8	if isAggregate(N{j})
9	SpSelect += getAggregate(N{j})+"(o" + j+})+")"
10	end if
11	else SpSelect += "o" + j
12	tp={ ?si p oi } { triple pattern generated }
13	TP.put (r, tp)
14	else
15	if isAggregate(N{j})
16	SpSelect += getAggregate(N{j})+"(s" + j+})+")"
17	else SpSelect += "s " + j
18	end if
19	tp={ ?si rdf:type r } {triple pattern generated }
20	TP.put (r, tp)
21	endif
22	endfor
23	return SpSelect, TP

In the exConvSqlSelect() algorithm given in Table X, the isAggregate() function checks each attribute in the select clause contains an aggregate function (MIN, MAX, AVG, ...) if it exists he calls the getAggregate () function before adding the attribute to SPARQL Select to convert to the proper aggregate function.

B. Conversion of GROUP BY and HAVING clauses

The algorithm for the conversion of GROUP BY clause is given in Table XI.

TABLE XI. ALGORITHM FOR CONVERSION OF SQL GROUP BY CLAUSE

	Algorithm: ConvSqlGroupBy()
Input:	G an attribute of SQL-GroupBy query ,TP triple patterns
Output:	SPARQL Group By,TP triple patterns
1	groupBy = ""
2	tp=""
3	if(TP.contains(G.relation) then
4	tp = TP.get(G.relation,G.attrubit)
5	else
6	{ Extract p from A }
7	if G.attributeÊ!='subject' then
8	s = TP.getSubject(G.relation)
9	p = G.attribute
10	o = "o"+TP.getLastObject()+1
11	tp = { ?si p oi } { triple pattern generated }
12	TP.put (r, tp)
13	groupBy = "Group By "+tp.getObject();
14	else
15	s = TP.getSubject(G.relation)
16	o = "o"+TP.getLastObject()+1
17	tp = { ?s rdf:type ?o} { triple pattern generated }
18	TP.put (r, tp)
19	groupBy = "Group By "+tp.getSubject();
20	endif
21	end if
22	return groupBy, TP

.

.

.

Table XII presents the algorithm for converting HAVING clauses.

TABLE XII. ALGORITHM FOR CONVERSION OF SQL HAVING CLAUSE

Algorithm: ConvSqlHaving()	
Input:	H having condition of SQL-Having query ,TP triple patterns
Output:	SPARQL Having
1	having = ""
2	tp=""
3	{extract left and right operand from H}
4	p1 = H.getLeftOperand()
5	p2 = H.getRightOperand()
6	att = p1.getAttrubitFromAgregat();
7	tp = TP.get(att.relation,att.attrubit)
8	agg = H.getAgregat ();
9	if att.atribut="subject"then
10	having = "HAVING(+agg+"("+tp.subject+")" +H.operator+p2
11	else
12	having = " HAVING(+agg+"("+tp.object+")"+H.operator+p2
13	end if
14	return having

C. Conversion of INSERT Queries

Both getTP () and generateInsert () algorithms given in Tables XIII and XIV are needed to convert an INSERT SQL query. The first algorithm traverses an INSERT query to extract all the attributes and their values, to store them in a S hashmap, then the GenrateInsert () algorithm that takes the result of getTP () and generates an equivalent SPARQL insert.

TABLE XIII. EXTRACTION OF PREDICATES AND VALUES FROM SQL QUERIES

Algorithm: getTP()	
Input	Input: sql : SQL Insert Query
Output	Output: S : Map of predicates and values
1	A= sql.getAttributes() // return all attributes from INSERT
2	V= sql.getValues() //return values from INSERT
3	S={ };
4	for j = 1 to A.size do
5	S.put(A.value(j), V.value(j));
6	end for
7	return S

TABLE XIV. INSERT CONVERSION

Algorithm: generateInsert()	
ínput	A : Map of predicates and values
Output	SPARQL Insert Query
1	sparql= "INSERT DATA \n" ;
2	for i =1 to A.size do
3	sparql += V.value(0) +" "+V.getValue(i) +" "+A.getValue(i)+". \n";
4	end for
5	return sparql

D. Conversion of DELETE Queries

The conDeleteSql () algorithm given in Table XV was called to convert a SQL DELETE to SPARQL DELETE, which analyzes the SQL deletion request to extract the BE Boolean conditions, and generate the triple patterns TP.

TABLE XV. CONVERSION ALGORITHM FOR DELETE QUERIES

Algorithm: ConvDeleteSql()	
Input:	SQL : an sql delete query
Output:	SPARQL DELETE query
1	sparql = "DELETE { /n "where = " WHERE { /n"
2	BL= sql.getBE() // get Boolean conditions
3	TP = sql.getTP() // generate TP from SQL query
4	if (BL.isEmpty()) then
5	for tp in TP do
6	where += "?" + tp.subject + " " + tp.predicate +" ?" + tp.object
7	end for
8	{verification of booleans conditions }
9	else if (! BL.isEmpty) then
10	for e in BL do
11	p=e.getLeftOperand();
12	tp =TP.get(p.relation)
13	sparql += "?" + tp.subject + " " + tp.predicat +"?" + tp.object
14	where += "?" + tp.subject + " " + tp.predicat +"?" + tp.object
15	end for
16	{ add FILTER in query }
17	where += "FILTER("
18	for e in BL do
19	p1=e.getLeftOperand();
20	p2=e.getRightOperand();
21	tp =TP.get(p1.relation)
22	where += tp.object+"" + e.operator + "" + p2
23	endfor
24	sparql+="}"
25	where+=")"
26	endif
27	where+="}"
28	Return sparql + where

IV. IMPLEMENTATION AND TESTS

To verify the efficiency of the conversion framework, this work was implemented using java programming language and jena [5] with connection to Allegrograph[2] for data storage and query processing.

AllegroGraph [7] is a database and application framework for building semantic applications. It can store data in the form of triples, query these triples via SPARQL.

Fig. 2 shows a screenshot of the API execution platform for the case of the conversion of previously given SQL query with ORDER BY into an equivalent SPARQL query.

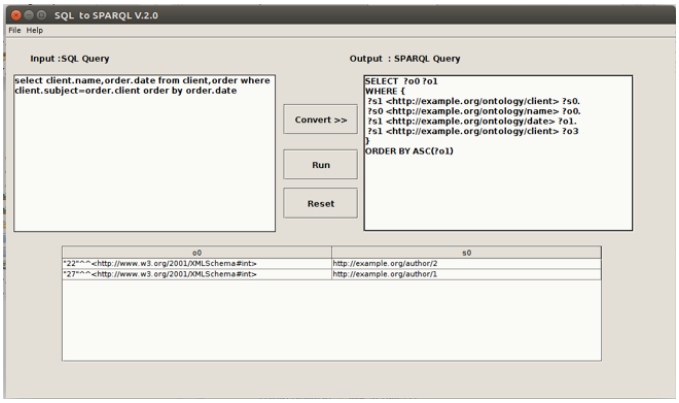


Fig. 2. Implementation of SQL2SPARQL4RDF Framework

The API also allows the execution on Allegrograph of the SPARQL query that is obtained through the conversion algorithm and the presentation of the execution result.

V. Conclusion

This paper present an extension of SQL2SPARQL framework already developed in the previous work for efficiently querying RDF documents using SQL queries expressed on the basis of a relational schema model without any conversion of extensions. With this framework, users who use SQL language no longer need to store data in a relational database, just store them in RDF data according to a schema mapping and SQL processors to execute their queries. They can also query large amounts of RDF data that are not possible to store in relational databases. They can also query a mass of RDF data that cannot be stored in relational databases. The considered extension dealt with the problems of converting SQL constructs related to INSERT, DELETE, GROUP BY and HAVING clauses. An associated java platform is also developed with a module for tests on the Allegrograph triplestore to prove the efficiency of this established conversion algorithms. In the future work, intend to optimize the conversion algorithms to apply them in web services to provide an API to make the mapping of SQL queries to equivalent SPARQL queries very easy to use in other platforms.

Acknowledgment

The authors would like to express thiere very great appreciation to there supervisors for thier valuable and constructive suggestions during the planning and development of this research work. Thier willingness to give thier time so generously has been very much appreciated. The authors would also like to extend thiere thanks to the colleagues of the laboratory of the mathematics and computer science department of Faculty of Sciences and Techniques, Hassan I University of settat, for their help and support.

References

[1] R2RML: RDB to RDF Mapping Language: https://www.w3.org/TR/r2rml/.

[2] Allegrograph: http://www.franz.com.

[3] Bioportal, http://sparql.bioontology.org/.

[4] Chhaya, P. et al.: Using D2RQ and Ontop to publish relational database as Linked Data. In: ICUFN. pp. 694–698 IEEE (2016).

[5] Kruti Jani, Dr. V.M. Chavda A Study on Semantic Web Framework: JENA and Protégé Indian Journal of Applied Research, Vol.IV, Issue. I - Jan 2014

[6] F. Alam, S. Ali, M. A. Khan, S. Khusro, A. Rauf, "A Comparative Study of RDF and Topic Maps Development Tools and APIs", BUJICT Journal, Volume 7, Issue 1, December 2014, pp. 1-12.

[7] W3C: SPARQL Protocol for RDF. http://www.w3.org/TR/rdf-sparql-protocol/.

[8] W3C: SPARQL Query Language for RDF. http://www.w3.org//TR/rdf-sparql-query/.

[9] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach. Scalable semantic web data management using vertical partitioning. In VLDB, pages 411–422, 2007.

[10] L. Alaoui, A. Abatal, K. Alaoui, M. Bahaj and I. Cherti, (2015). SQL to SPARQL mapping for RDF querying based on a new Efficient Schema Conversion Technique. International Journal of Engineering Research and Technology, 4(10). IJERT.

[11] S. Anand and A. Verma, (2010). Development of Ontology for Smart Hospital and Implementation using UML and RDF. IJCSI Int. J. of Computer Science Issues, Vol. 7, Issue 5.

[12] P. Bellini and P. Nesi (2018). Performance assessment of RDF graph databases for smart city services. Journal of Visual Languages and Computing 45, 24–38.

[13] J. Broekstra, A. Kampman and F. van Harmelen (2002). Sesame: a generic architecture for storing and querying RDF and RDF schema. In: ISWC, pp. 54–68.

[14] E. I. Chong, S. Das, G. Eadon and J. Srinivasan (2005). An efficient SQL based RDF querying scheme. In: VLDB, pp. 1216–1227.

[15] G. Garbis, K. Kyzirakos and M. Koubarakis (2013). Geographica: a benchmark for geospatial RDF stores. In Proceedings of the 12th International Semantic Web Conference, 343–359.

[16] F. Goasdoué, Z. Kaoudi, I. Manolescu, J. A. Quiané-Ruiz and S. Zampetakis (apr 2015). CliqueSquare: Flat plans for massively parallel RDF queries. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 771–782.

[17] S. Harris, N. Lamb and N. Shadbolt (2009). 4store: The design and implementation of a clustered RDF store. In Proceedings of the 5th Int. Workshop on Scalable Semantic Web Knowledge Base Systems, 16.

[18] J. Lu, L. Ma, L. Zhang, J.-S. Brunner, C. Wang, Y. Pan and Y. Yu (2007). SOR: A practical system for ontology storage, reasoning and search. In: Proceedings of VLDB, pp. 1402–1405.

[19] T. Neumann and G. Weikum (2010). The RDF-3X engine for scalable management of RDF data. The VLDB Journal, 19(1):91–113.

[20] J. Rachapalli, V. Khadilkar, M. Kantarcioglu and B. Thuraisingham (2011). RETRO: a framework for semantics preserving SQL-to-SPARQL translation. In: EvoDyn Workshop.

[21] [21] K. Rohloff and R. E. Schantz 2010). High-performance, massively scalable distributed systems using the mapreduce software framework: the shard triple-store. In ACM Programming Support Innovations for Emerging Distributed Applications, 2010

Instagram Shopping in Saudi Arabia: What Influences Consumer Trust and Purchase Decisions?

Taghreed Shaher Alotaibi[1], Afnan Abdulrahman Alkhathlan[2], Shaden Saad Alzeer[3]
College of Computing and Information Technology
Shaqra University, Riyadh, Saudi Arabia

***Abstract*—The recent developments of social networking sites (SNSs), along with the increasing usage of online shopping, has led to the emergence of social commerce platforms. Social commerce (s-commerce) is the use of Web 2.0 technologies and social media to deliver e-commerce services for consumers. The Kingdom of Saudi Arabia (KSA) has been witnessing a rapid growth in s-commerce usage, with Instagram being the most popular networks in the region. This paper is one of the few that investigates the factors affecting consumers' trust and purchase intentions in Instagram as a s-commerce platform in Saudi Arabia. The proposed model explores a number of factors, such as Social Media Influencers (SMIs), Key Opinion Leaders (KOLs) and consumer feedback, in terms of their influence on consumers' trust and purchase decisions. In addition to the effect of Maroof, which is an e-service provided by the Saudi Ministry of Commerce and Investment to evaluate the reliability of online stores. Following a quantitative approach and using Partial Least Squares Modeling (PLS-SEM), findings of this study revealed a positive relationship between consumers' trust and their purchase intentions. Additionally, the impact of SMIs and consumer feedback was shown to increase consumers' trust, in turn affecting intent to buy from Instagram stores, while the effect of Maroof and KOLs was shown to directly influence consumers' purchase intentions.**

Keywords*—Social commerce; instagram; trust; purchase intentio; maroof; key opinion leader; social media influencers

I. Introduction

The emergence of social media and Web 2.0 technology has enormously transformed people's lives, creating new ways to connect, collaborate and socialize. and with the increasing usage of smart devices, there has been a notable shift toward online shopping, which allows the enjoyment and convenience of home delivery and time-saving benefits. Recently, an emerging subset of e-commerce called social commerce (s-commerce) has come to the fore, wherein SNSs are used as online markets. The combination of SNSs and e-commerce into s-commerce has changed the game for both the seller and the buyers; the shopping experience is no longer just a click away. Consumers are now actively engaging in the full shopping journey, from exploring products and comparing prices to making purchases and affecting each other's decisions with opinions and recommendations, all while enjoying the social intimacy and bonding these networks provide [1].

Conversely, it is also becoming very common to see businesses and commercial brands launching their pages on SNSs such as Facebook, Twitter and Instagram, which create opportunities to promote their products and services as well as immediate bridges to reach and connect with their customers. A recent statistic by [2] revealed that over 25 million brands have an account on Instagram.

Instagram is one of most popular social networking platforms for visual sharing. With one billion active users monthly in 2019 [3], it has enormously changed the way people share their photos and videos. Instagram users in general are led mostly by millennial and post-millennial generations [2]. The growing popularity of Instagram has led to a rise in using it as a s-commerce platform, especially in Arab countries. According to the Northwestern University in Qatar, Instagram usage has overtaken Twitter usage in the regions of Saudi Arabia and different Arab countries such as Egypt, Lebanon, Tunisia and the United Arab Emirates. Moreover, a recent survey showed that Instagram is the social network most used by the e-commerce industry in the Middle East region at a rate of 77% [4]. Additionally, over a third of Instagram users have made purchases through their accounts. Approximately a third of them are interested in shopping and fashion [4], yet there is only a limited number of studies [5][6][7] in the current literature that focus on Instagram as a s-commerce platform.

In recent years, Saudi Arabia has witnessed rapid growth in s-commerce usage. About 42% of online shoppers in Saudi Arabia have purchased from social media platforms [8], finding a great sense of assurance in that they can easily find other consumers' recommendations and experiences. They also perceive a higher sense of transparency when the store owner is active online [8]. Although Saudi people are among the most active users on Instagram [4], little research has been conducted on the use of s-commerce generally in Arab regions and more specifically in Saudi Arabia [9][10]. Therefore, this research will shed light on the adoption of Instagram in Saudi Arabia as a s-commerce platform. The study aims to gain a deeper understanding of how Saudi consumers make their purchasing decisions and what factors are affecting these decisions. These factors are SMIs, KOLs, consumer feedback and Maroof as a third party.

SMIs are those who have attained popularity through their participation and activities on social media. Their followers usually have a great sense of attachment and admiration toward their lifestyle and behaviors [11]. Even though business companies usually sponsor and pay for influencers' posts to advertise their products, SMIs are more credible and have greater impact on their followers compared to traditional advertisements [12]. Nearly 35% of people who use Instagram

on a daily basis have made a decision to purchase a product recommended by an influencer [13]. KOLs, on the other hand, are those such as doctors, educators and business leaders who are perceived by the public as experts in a specific area of knowledge and are valued by their followers for their opinions and expertise [14][15]. KOLs are not necessarily active online, yet they are found to have a significant effect on consumers' trust and purchase intentions with respect to s-commerce platforms [7]. In addition, consumers tend to check comments and feedback shared by others on the services or products they buy online, so many s-commerce platforms provide features that allow consumers to write and share their experiences and opinions, which usually influence other shoppers' purchase decisions [16].

Furthermore, due to the lack of direct interaction in s-commerce, trust is one of the fundamental issues in e-commerce in general and more recently in s-commerce [6]. Many research studies have been conducted on the relationship between consumers' trust in s-commerce and their purchase intentions [17][18][9][19]. The proposed model in this study aims to further investigate how trust and purchase intention are related among Saudi consumers. Moreover, this work takes into consideration the impact of the Maroof e-service. Created and supervised by the Ministry of Commerce and Investment in Saudi Arabia, Maroof is a free interactive platform where consumers rate and evaluate online stores based on their past experiences, which will affect consumer trust in registered stores. The service also benefits store owners, who can reach more customers. To our knowledge, this is the first study that considers Maroof as an impact factor on the trust and purchase intentions of Saudi consumers in s-commerce. This work is proposed in the hope that it will provide a useful foundation for future s-commerce research in the region.

The reminder of the paper is organized as follows: first, s-commerce and trust are defined as concepts through the current literature. Second, the proposed model is described along with each of the proposed hypotheses, followed by the research methodology and findings. Finally, the limitations and future directions for the research are discussed to conclude the paper.

II. Conceptual Framework

A. Social Commerce

Social commerce is an important topic that emerged in 2007; researchers' focus on it has risen from 2009 to the present [20]. Yet one of the fundamental issues in s-commerce is the lack of a standard definition. Prior studies tried to make a clear definition from many perspectives, but there is still no agreed-upon definition. Esmaeili et al. [21] defined it as "an Internet-based commercial application that makes use of Web 2.0 technologies and social media, and it supports user-created content and social interactions." Turban et al. [22] made a broader definition of s-commerce which is "e-commerce transactions delivered via social media." Moreover, Han et al. [23] conducted a review of 22 definitions of s-commerce and constructed their own definition, stating, "social commerce is a new business model of e-commerce, which makes use of Web 2.0 technologies and social media to support social-related exchange activities."

There are several perspectives that define s-commerce. E-commerce applications are the basis s-commerce is built upon. In addition to Web 2.0 technologies, which led to the creation of social media and SNSs that facilitated the interaction between consumers and sellers, smartphones are considered one of the factors that have helped the growth of s-commerce [22]. It is estimated that 89% of the Saudi population uses the Internet, with 68% of active users on social media [24]. In terms of e-commerce, Saudi Arabia is considered the first country in the Middle East [4]. Seventy-nine percent of Saudi users search online to purchase services or products, and 64% of them made an online purchase [24]. In the context of Instagram, Saudi users are still the largest population in the Middle East, accounting for 13 million active users in 2018 [24]. Nevertheless, there have been a few studies that investigated the adoption of s-commerce in this country.

B. Trust

Trust is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [25]. According to [26], trust in s-commerce in general includes two dimensions: information-based trust, which denotes consumers' trust in the information found in the s-commerce sites; and identification-related trust, which is the trust s-commerce consumers place in each other. As for the first dimension, many scholars have found that consumers' trust toward s-commerce sites significantly affect their purchase intention[27][28][29][18][19]. For the second dimension, [29] found that consumers' trust toward other members of the s-commerce platform increases their trust toward the platform itself and consequently, their purchase intention. On another note, the number of Instagram store followers was found to be a great indicator of legitimacy and therefore, trustworthiness of the seller. People tend to believe that the more followers a seller has, the more trustworthy he or she is [30]. Consumers' prior experience with a certain company does not necessarily contribute to their trust in the company's s-commerce site [26]. In addition, [18] examined the importance of trust in determining consumers' engagement in s-commerce. Based on the trust transfer theory [31], they studied multiple technical factors such as ratings, recommendations, forums and communities that affect consumers' perceived trust of s-commerce sites, and found a significant effect on consumers' trust and engagement. Similarly, the work by [18] found that social presence, familiarity and perceived sense of security were found to have the same positive effect on trust.

Prior research on trust in the s-commerce context mainly focused on either the psychological or the social aspects of customers that build their trustfulness, such as familiarity [18][17][19][5], enjoyment [10][17][5], and habit [10][5]; or the technical factors of an s-commerce site, such as information quality [10][32][9], perceived ease of use and perceived usefulness [9][27][5]. Yet little work has been done to investigate the effects of external factors on consumers' trust or purchase intention. For example, [7] studied the effect

of KOLs and peer customer endorsements on trust and found that KOLs have a positive impact while customer endorsements have an insignificant one. However, their research needs more generalizability across different cultures.

III. Research Model and Hypothesis Development

This study is will investigate the effect of a group of external factors such as Maroof e-service, SMIs, KOLs and consumer feedback on Instagram shopping in Saudi Arabia, in order to deeply understand the formation of consumers' trust and intention to purchase. The proposed research model is depicted in Fig. 1.

A. Maroof

Maroof is an e-service that aims to enhance the e-commerce sector in Saudi Arabia. It was created by the Ministry of Commerce and Investment in conjunction with Thiqah Business Services, the operator and developer of the platform. Maroof works as a platform that helps to evaluate online stores in Saudi Arabia, which will benefit both buyers and sellers at the same time. Online sellers registered with Maroof can reach more customers easily. Through the Maroof network, they can share their social networking accounts and contact information for their stores. Buyers who look for services or products can search for them using Maroof in an easy and trustworthy way. Moreover, buyers can browse ratings, reviews and other consumers' feedback in a store's page, to decide whether to buy from a given store. Also, consumers can rate stores depending on their experiences [33]. Consumer ratings and comments are checked for validity by Maroof team. All information shared on Maroof is highly secure, and all services and products provided by these stores are under the regulation of the Ministry of Commerce and Investment. The Maroof initiative started in 2016 and, a recent report from the ministry showed that in 2018, the number of stores registered on Maroof has reached 20,000. Registering on Maroof is free to encourage merchants to use it. Moreover, when a merchant creates a commercial registry, then he or she is awarded a special certificate from Maroof called a Golden Certificate. Most online stores in Saudi Arabia have an Instagram account, so they register on Maroof to become accredited. Since this service is under the ministry's control, it is mainly used as a way to enhance trust between buyers and sellers.

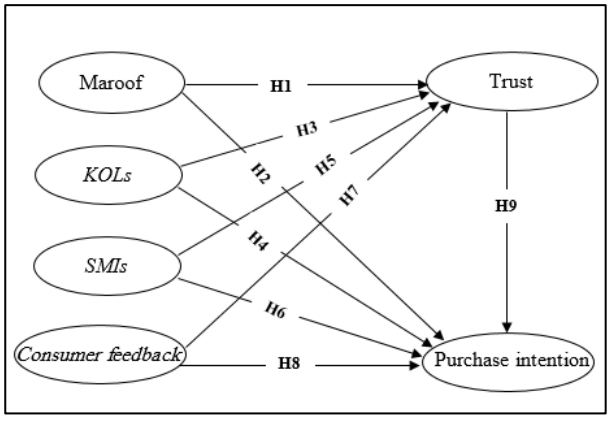


Fig. 1. Research Model.

This research considers Maroof as a special and unique factor that impacts Saudi consumers' trust and intentions to purchase from Instagram. To the best of our knowledge, this study is the first one that investigates the effect of Maroof on s-commerce in Saudi Arabia. Thus, it hypothesizes the following:

H1: Maroof e-service has a positive impact on consumers' trust.

H2: Maroof e-service has a positive impact on consumers' purchase intentions.

B. Key Opinion Leaders (KOLs)

KOLs represent a small portion of people who are deeply knowledgeable and skillful in a certain field [34]. Their superior levels of education, qualifications and social status empower them to influence their audience [35]. An opinion leader is an individual who greatly influences people's decisions, behaviors and attitudes [14]. They are either experts in a specific area of knowledge, or have a wide range of social connections [14]. Previous research has identified several characteristics of KOLs, namely that they are professionals, innovative, involved, and sociable [36]. Moreover, [37] demonstrated several characteristics in determining KOLs: they must be knowledgeable in their field, respected by their audience, have wider access to information sources than their followers, and lastly, have the ability to maintain their leadership. According to [36], KOLs can be generally categorized into two types: the ones who are active offline such as public figures and celebrities, and thus can easily become online opinion leaders. And those who share their opinions online exclusively.

However, there is some degree of overlap between the terms "social media influencer" and "key opinion leader." Both terms describe individuals who have the ability to impact people within their areas of interest. The main difference is that the KOL term extends to experts in particular fields such as technology, science, medicine and more. Also, they usually have profession parallel to their subjects of expertise, such as doctors, entrepreneurs or writers [38]. Their audience members are not necessarily fans of their personalities, but rather of their knowledge and expertise. On the other hand, SMIs are popular personalities on social media platforms such as Instagram, Snapchat, Facebook and YouTube. They have a large community of followers who greatly admire their lifestyles and personal choices [39].

KOLs are considered experts in their fields, and are usually asked for advice and recommendations [7]. In their work, [34] found a positive effect on marketing strategies caused by the strong relationship KOLs have with their followers. Instagram is the social platform most frequently used by opinion leaders [40]. Therefore, Instagram users heavily rely on KOLs' opinions and recommendations due to the lack of explicit measures of store ratings and reputations [7]. KOLs and consumer comments were found to be strong influencers of consumers' trust and thus their purchase intention in Indonesia [41]. Furthermore, [7] found that KOLs have a strong effect on consumers' trust toward an Instagram store.

This work considers KOLs an important factor in determining consumers' trust in Instagram stores and distinguishes them from SMIs, who do not possess deep knowledge in a certain field. Thus, the following hypotheses are formed:

H3: KOLs have a positive impact on consumer trust.

H4: KOLs have a positive impact on consumer purchase intentions.

C. Social Media Influencers (SMIs)

In recent years, with the continually growing use of social media, a global marketing phenomenon called "influencer marketing," in which brands and business use social media figures to reach their target audience, has emerged to the forefront. SMIs are people who attained their popularity by generating and sharing content via social media platforms such as Instagram, Facebook, YouTube and Snapchat. SMIs usually focus on a certain field, such as healthy lifestyle, food, travel, beauty or fashion. Unlike KOLs, SMIs' audiences are nonspecific and usually interested in their online personas and lifestyles rather than their discipline of knowledge. SMIs usually advertise for products or services related to their own interests and lifestyles. Consumers tend to make their purchase decisions under the impact of the influencers' experience and recommendations [42]. Contrary to public celebrities who are well-known through traditional media channels such as TV and newspapers prior to the social media era [42], a study by [11] found that SMIs or the "Instafamous" are more admirable than traditional celebrities, and that consumers perceive them as more relatable and credible. Social media marketing is getting more attention and appreciation from consumers compared to traditional advertisements [12]. Companies pay influencers as part of their marketing strategies. Financial benefits are usually the main reason for influencers to adopt corporate marketing and vice versa. According to [43], the returns on investment (ROI) from influencer marketing is 11 times greater than what traditional advertisements make in a year.

Influencers have likable personalities and are more relatable than celebrities and public figures. They indulge their audiences in their personal lives and share their opinions and experiences with them, influencing their attitudes, ideas and decisions [12]. The author in [44] found that the more attached a consumer is to an influencer, the more driven he or she is to buy the products promoted by the influencer. A study by Twitter in 2016 revealed that consumers trust SMIs similarly as they trust their own friends or relatives, and 40% of them say that they have purchased an item online after seeing it used by an influencer on Twitter.

Influencer marketing is related to the halo effect theory described by [45], a cognitive bias wherein the perception of an individual is based upon a single trait or characteristic. Accordingly, brands are using influencers with positive halos to promote their products. This creates a positive association between the influencer and the product [11]. Hence, [46] revealed a positive correlation between influencers' content informative value and consumers' trust as well as their purchase intention. Moreover, the trustworthiness of influencers and their attractiveness and similarity to their followers have significant effects on the followers' trust in their sponsored content. Another study by [44] investigated the persuasion cues of Instagram and YouTube fashion and beauty influencers in France. They found that the credibility of a given influencer positively impacts the purchase intention of his or her target audience. In addition, [47] investigated how influencers' credibility, physical attractiveness, promoted-products congruency and meaning transfer affect the purchase intentions of their followers. Their findings revealed that credibility and physical attractiveness failed to affect consumers' purchase intentions, while there were strong relationships between product match-up and meaning transfer with purchase intentions.

However, research on influencer marketing is still insufficient. The work by [15] contributes to a deeper understanding of how SMIs affect consumer behavior and decisions. The focus of this study is limited to Instagram influencers due to the popularity and ongoing growth of Instagram as an s-commerce platform. Hence, this study hypothesizes:

H5: SMIs have a positive impact on consumers' trust.

H6: SMIs have a positive impact on consumers' purchase intentions.

D. Consumer Feedback

Despite the increasing usage of online shopping, consumers still experience some uncertainty when making purchase decisions. This is mainly because all products provided by online stores are displayed merely as pictures with a little description. These pictures do not necessarily reflect the real quality or condition of these products. In order to solve this dilemma, most e-commerce platforms have added a variety of technical features that allow consumers to share their experiences with their purchased products. These features can vary from ratings, reviews and recommendations to forums, communities [1] and comments[48]. Consumers usually search for peer reviews and ratings before making purchase decisions because they tend to believe more in consumer experiences rather than the information posted by the business itself [19].

Moreover, several studies showed that consumer feedback, whether a review or a comment on a website or in a community, has a positive impact on consumers' trust. For instance, [16] studied the features provided by Web 2.0 technologies that encourage user participation in online interactions. The study aimed to investigate how the social interaction between consumers increases trust toward a specific vendor. They built a model for the social constructs of forums and communities, recommendation and referrals and ratings and reviews. Their findings showed that s-commerce constructs are essential to influencing customer trust. Furthermore, [48] found that consumers' trust toward a s-commerce site is influenced by the information generated by that site's consumers, which in turn increases their purchase intentions.

In the context of Instagram, consumers have the ability to write their honest opinions on a store's page, so other

customers can see the comments and decide whether to buy from the store. In addition, some Instagram store owners can share testimonials from their previous customers as posts or by using the highlight feature. The study [41] demonstrated that trust and purchase intention are influenced by consumer feedback. Nevertheless, a few studies have been conducted to investigate the effect of consumer feedback. This study considers consumer feedback as an important factor that affects trust and purchase intention and hypothesizes the following:

H7: Consumer feedback on an Instagram store positively impacts trust.

H8: Consumer feedback on an Instagram store positively impacts purchase intentions.

E. Trust and Purchase Intention

Due to the lack of physical contact between vendors and consumers, uncertainty is the dominant sense in s-commerce [16]. Consumers tend to rely on each other's reviews, ratings and recommendations in their decision-making processes to establish their trust in a specific vendor or site. Therefore, trust plays a crucial role in the s-commerce context.

Purchase intention is the willingness of an individual to make purchases from a s-commerce platform [19]. Previous studies demonstrated that customers' intent to purchase from a specific store is highly dependent on their trust in that store [18][7][19]. Furthermore, [5] identified trust as a significant factor in influencing consumer purchase intentions on s-commerce sites. Prior studies [27][28][29] consistently revealed a strong positive correlation between a consumer's trust and his or her intent to purchase. Aligned with these studies, the following hypothesis is proposed.

H9: Consumers' trust toward an Instagram store positively impacts their purchase intentions.

IV. Research Methodology

A. Sample and Data Collection

This study was designed to investigate the impact of Maroof e-service, SMIs, KOLs and consumer feedback on the trust and purchase intentions of consumers. An online survey was built to collect responses from Instagram users. To maximize the number of participants, a link for the survey was distributed over different social media platforms such as Instagram, Telegram, WhatsApp and Twitter as well as by email. Because the study targeted only Instagram, this was highlighted at the beginning of the survey to make sure participants understand the target sample. Furthermore, since the study targeted Saudi users, the survey was translated into Arabic before distribution. As a result, 225 complete responses were received. Responses were not missing for any of the survey items.

Based on the critical Mahalanobis distance [49], nine outliers were identified and removed from the respondents' sample, so a total of 216 valid responses were used in the analysis. The study included more females than males (75.5% vs. 24.5%). Various age groups were represented in the study sample (Table I). Regarding education, most participants (71.3%) had a bachelor's degree, 12% completed high school, 4.63% had a diploma, and 12% had a postgraduate degree.

B. Measures

The survey was built using a five-point Likert scale from 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree to 5 = strongly agree. Two academic faculty members participated in this study to check and correct the survey questions before they were distributed. These faculty members were familiar with s-commerce and statistical techniques (Table II).

TABLE. I. Descriptive Statistics of the Study Sample

	N=216
Sex:	
Female	163 (75.5%)
Male	53 (24.5%)
Age:	
18 or less	13 (6.02%)
19 - 25	82 (38.0%)
26 - 30	53 (24.5%)
>30	68 (31.5%)
Education:	
High school or lower	26 (12.0%)
Bachelor's degree	154 (71.3%)
Diploma	10 (4.63%)
Postgraduate	26 (%12.0)

V. Analysis

PLS-SEM analysis was performed using the SEMinR package, which is used to create and estimate structural equation models using partial least squares path modeling (PLS-PM or PLS-SEM). Advantages of PLS-SEM over conventional covariance-based structural equation modeling (CB-SEM) include that it can handle small sample sizes and is suitable for exploratory and confirmatory research [50]. It also uses the bootstrap (resampling) approach to compute standard errors, p values and confidence intervals. The present study used 200 resamples to assess path significance and calculate the 95% confidence intervals. Hypothesis testing was performed at a 5% level of significance. Goodness of fit was assessed using R^2 for the dependent variables included in the analysis (trust and purchase intention).

A. Reliability

Reliability is a measure of internal consistency in a questionnaire (i.e., how coherent items of the same scale are, or how closely they are correlated) [51]. Cronbach's Alpha (α) and composite reliability (rhoC or CR) were two methods used to assess the reliability of the six factors in this study [52]. Reliability assessment using Cronbach's α and composite reliability showed that the factors included in the analysis were reliable as demonstrated by values > 0.7 for all scales (Table II).

TABLE. II. SOURCES OF CONSTRUCTS, RELIABILITY AND VALIDITY

Code	Scales	Factor loading	CR	AVE	Cronbach's α
	Maroof *e-service (M)* *(New)*		0.817	0.536	0.77
M1	Maroof e-service is a trusted platform that protects shopper interests.	0.755			
M2	I trust Instagram store that has a Maroof certificate.	0.922			
M3	I check Instagram store's page on Maroof e-service to see customers' comments about the store.	0.547			
M4	I search for Instagram stores that have a high rate on Maroof e-service page.	0.651			
	Social media influencer (SMI) (*Adopted and adjusted from [7] [56])*		0.91	0.718	0.86
SMI1	I feel that social media influencers are generally trustworthy	0.867			
SMI2	I feel that recommendations by social media influencers are generally reliable.	0.905			
SMI3	I'm more likely to try a new product if my favorite influencer recommends it.	0.787			
SMI4	I believe that social media influencers have my best interests in mind.	0.826			
	Key opinion leaders (KOL) *(Adopted and adjusted from [7] [56] [44])*		0.928	0.719	0.9
KOL1	I find that KOLs are experts in their domain.	0.788			
KOL2	I feel that KOLs are generally trustworthy.	0.84			
KOL3	I feel that recommendations by KOLs are generally reliable.	0.902			
KOL4	I'm more likely to try a new product if a KOL recommends it.	0.847			
KOL5	I believe that KOLs have my best interests in mind.	0.861			
	Consumer Feedback (CF) *(Adopted and adjusted from [7] [56])*		0.851	0.54	0.78
CF1	Instagram stores display testimonials from satisfied customers.	0.642			
CF2	I can see from the comments in an Instagram store that existing customers are satisfied with the store.	0.736			
CF3	I feel Instagram users' comments are generally honest.	0.855			
CF4	I feel Instagram users' comments are reliable.	0.853			
CF5	I usually check customers' comments in Instagram stores before making a purchase.	0.534			
	Trust (T) (*Adopted and adjusted from [7])*		0.884	0.657	0.83
T1	Instagram shopping is generally trustworthy.	0.824			
T2	I trust information on Instagram to be true.	0.839			
T3	I think that Instagram stores will not do anything to take advantage of their customers.	0.751			
T4	I feel comfortable making purchases from Instagram stores.	0.824			
	Purchase Intention (PI) (*Adopted and adjusted from [7] [57])*		0.907	0.764	0.85
PI1	I would seriously contemplate buying from Instagram.	0.904			
PI2	I am likely to make future purchases from Instagram.	0.847			
PI3	I intend to purchase products or services from Instagram whenever I need to shop.	0.87			

Notes: CR, composite reliability; AVE, average variance extracted; α Alpha

B. Validity

To check the validity of this research, construct validity was tested. Construct validity can be assessed by discriminant and convergent validity [53].

Convergent validity was examined through loadings (correlations) of manifest variables on the corresponding latent variables. Loading should be equal to or greater than 0.5. The average variance extracted (AVE) was also used to assess convergent validity. Values greater than 0.5 were considered satisfactory [54]. All scales scored > 0.5 (Table II), indicating that convergent validity can be assumed.

Discriminant validity was assessed by examining the correlation between factors. The correlation between any two factors should not exceed 0.7 and should not exceed $\sqrt{AVE}$. For each factor, $\sqrt{AVE}$ was higher than its correlation with all remaining factors (Table III). This indicates that discriminant validity was met.

A further assessment to test the discriminant validity is checking the factor loading of each indicator. Cross-loadings were also examined to ensure that none of the items were loaded on more than one factor (Table IV). Factor loadings were greater than 0.7 for all items (which is considered ideal) except for M3 and CF5. However, removing these items from

the model did not alter the final model results. Thus, it was decided to keep them in the model as they did not cross-load on any other factors and had a loading > 0.5, which is considered acceptable [55]. The AVE was > 0.5 for all factors, which was another reason to keep these items in the model (Table II).

TABLE. III. CORRELATION BETWEEN FACTORS

	SMI	CF	M	KOL	T	PI
SMI	0.847					
CF	0.212	0.734				
M	0.011	0.305	0.732			
KOL	0.178	0.16	0.143	0.848		
T	0.517	0.391	0.12	0.19	0.81	
PI	0.352	0.288	0.235	0.243	0.649	0.87

Notes: Numbers on the diagonal (highlighted) are the √AVE. Other numbers represent the correlations between factors

TABLE. IV. CROSS LOADINGS

	M	SMI	KOL	CF	T	PI
M1	0.755	-0.043	0.117	0.187	0.02	0.132
M2	0.922	-0.022	0.128	0.277	0.13	0.262
M3	0.547	0.052	0.12	0.153	-0.016	0.01
M4	0.651	0.125	0.094	0.248	0.104	0.1
SMI1	0.023	0.867	0.079	0.201	0.48	0.286
SMI2	-0.003	0.905	0.113	0.212	0.481	0.28
SMI3	-0.016	0.787	0.224	0.094	0.34	0.356
SMI4	0.033	0.826	0.2	0.203	0.441	0.279
KOL1	0.114	0.053	0.788	0.078	0.05	0.128
KOL2	0.111	0.153	0.84	0.121	0.155	0.168
KOL3	0.132	0.163	0.902	0.108	0.198	0.244
KOL4	0.117	0.164	0.847	0.218	0.15	0.241
KOL5	0.132	0.174	0.861	0.127	0.195	0.207
CF1	0.31	0.113	0.166	0.642	0.209	0.273
CF2	0.266	0.204	-0.015	0.736	0.267	0.141
CF3	0.196	0.189	0.099	0.855	0.348	0.215
CF4	0.174	0.214	0.165	0.853	0.404	0.242
CF5	0.242	-0.009	0.185	0.534	0.119	0.185
T1	0.172	0.363	0.122	0.302	0.824	0.553
T2	0.082	0.384	0.141	0.38	0.839	0.502
T3	-0.072	0.455	0.189	0.252	0.751	0.371
T4	0.169	0.474	0.169	0.326	0.824	0.64
PI1	0.218	0.362	0.2	0.313	0.621	0.904
PI2	0.195	0.245	0.247	0.153	0.465	0.847
PI3	0.203	0.304	0.199	0.271	0.599	0.87

Notes: Numbers on the diagonal (highlights) are the factor loading of each item

VI. FINDINGS

PLS-SEM provided enough evidence to support H2, H4, H5, H7 and H9 but not enough evidence to support H1, H3, H6 and H8. (Table V) shows that Maroof certification was directly associated with higher purchase intention ($\beta = 0.152$, $P < 0.05$). This indicates that intent to purchase from an Instagram score increases by 0.152 points for each 1-point increase in trust in Maroof. Also, KOLs were directly associated with higher purchase intentions ($\beta = 0.105$, $P < 0.05$), indicating that purchase intention score increases by 0.105 points for each 1-point increase in trust in KOLs. In contrast, the SMI factor was associated with higher trust in Instagram ($\beta = 0.446$, $P < 0.001$). This indicates that consumers' trust scores increase by 0.446 points for each 1-point increase in their trust in SMIs. Similarly, consumer feedback was associated with higher trust ($\beta = 0.28$, $P < 0.001$), indicating that trust in Instagram shopping score increases by 0.28 points for each 1-point increase in trust in consumer feedback. Lastly, trust in Instagram was associated with higher purchase intentions ($\beta = 0.607$, $P < 0.001$). This indicates that the purchase intention increases by 0.607 points for each 1-point increase in trust.

Although there were no statistically significant direct effects of SMIs and consumer feedback on purchase intention (H6 and H8), mediation analysis showed that there was a statistically significant indirect effect (through trust) for SMIs on purchase intention ($\beta = 0.267$, $P< 0.001$), and there was a statistically significant indirect effect (through trust) for consumer feedback on purchase intention ($\beta = 0.17$, $P < 0.001$).

The model's goodness of fit was assessed by examining the proportion of variance in trust and purchase intention from Instagram stores that was explained by the independent variables (IVs). Examining the adjusted R^2 for trust and purchase intention revealed that the four IVs (Maroof, KOLs, SMIs and consumer feedback) explained 34.3% of the variability in trust. The four IVs and trust combined explained 44.5% of the variability in the purchase intention. Furthermore, the research tested three control variables (sex, education and gender), but none of them had significant effects on purchase intention or trust.

TABLE. V. PLS-SEM PATH COEFFICIENTS

H	Path	β	LLCI	ULCI	P
H1	M -> T	0.02	-0.093	0.157	> 0.05
H2	**M -> PI**	**0.152**	**0.031**	**0.266**	**< 0.05**
H3	KLO -> T	0.063	-0.055	0.168	> 0.05
H4	**KLO -> PI**	**0.105**	**0.014**	**0.215**	**< 0.05**
H5	**SMI -> T**	**0.446**	**0.357**	**0.544**	**< 0.001**
H6	SMI -> PI	0.021	-0.089	0.145	> 0.05
H7	**CF -> T**	**0.28**	**0.182**	**0.381**	**< 0.001**
H8	CF -> PI	-0.017	-0.121	0.096	> 0.05
H9	T -> PI	0.607	0.457	0.701	< 0.001

LLCI: Bootstrapped lower 95% confidence interval, ULCI: Bootstrapped upper 95% confidence interval, H: Hypothesis, β: Estimate

VII. Discussion

This research sought to investigate the factors influencing consumers' purchase intentions from Instagram in Saudi Arabia. PLS-SEM was used to test the proposed model. Our findings demonstrated that the proposed model had a good validity and reliability as well as a good predictive power. All four independent variables (Maroof, KOLs, SMIs and consumer feedback) affected consumers' purchase intentions. However, these variables exerted their effects through different pathways. Maroof and KOLs directly affected consumers purchase intentions. Their indirect effects on purchase intention (through the mediator trust) were not statistically significant. Thus, the effects of Maroof and KOLs on purchase intention are entirely direct. However, this result is slightly different than the studies [41][7] which found that KOLs directly influences consumers trust.

The exact opposite was observed with SMIs on purchase intention, whose direct effect was not statistically significant. This is in line with [48] but differs from [44] which found that SMIs credibility significantly affect their followers' purchase intentions. Similarly, consumer comments direct effect was on trust but not on purchase intentions. This result is different than the one by [42] which found that comments influences purchase intentions directly, while the one by [7] revealed that consumers comments do not directly affect their trust. However, the indirect effects of both SMIs and consumers feedback on purchase intention (through the mediator trust) were statistically significant. These results show that the effect of SMIs and consumer feedback on purchase intention is solely indirect.

Overall, the research indicates that the above factors statistically influence individuals' trust and purchase intention. Consistent with and in addition to previous studies [19][7][5][28][29], our findings revealed that trust is associated with higher purchase intentions.

VIII. Conclusion

This study proposed a new model to study the factors influencing consumers shopping on Instagram in Saudi Arabia. The model consists of the Maroof e-service, SMIs, KOLs, and consumer feedback and how they influence consumer trust and purchase intention. The study's contribution to the s-commerce field is by studying these factors combined. To our knowledge, this study is the first study to investigate the effect of the Maroof e-service and jointly analyze the influence of SMIs and KOLs on consumer trust and purchase intention.

Moreover, these research findings can help vendors, whether they are individuals or companies, to understand the important factors that influence their customers' purchase intentions to maximize their profits. We found that an Instagram store with a Maroof certificate will enjoy higher purchase intentions from consumers. Likewise, choosing the appropriate KOLs to advertise or recommend products will positively influence consumers' intent to buy. Meanwhile, SMIs were found to be positively related to consumers' trust in Instagram stores. Previous consumers' comments and feedback were also found to have the same positive effect on building consumer trust.

IX. Limitations and Future Directions

The aim of this study was to investigate the effects of different factors on consumer trust and purchase intention in the context of Saudi Arabia, and it is not without limitations. First, the proposed model was tested on the Instagram network only due to its high popularity in Saudi Arabia. Future studies could experiment on other social networks such as Twitter and Facebook as well. Second, the number of respondents was relatively small and the majority of them were females, so these results might be biased. It is hoped that future studies will enlarge their sample sizes to reflect the actual numbers of Saudi Instagram users. Furthermore, the targeted sample of this study was limited to the Saudi population. Hence, these findings might not be generalizable across different cultures. Third, this study investigated a limited number of factors. More factors, such as previous purchase experience and store reputation, could be examined in the future to measure their influence on consumer trust. Finally, although this study followed a quantitative approach and the proposed model had a good validity and reliability, we recommend that future studies use qualitative measures such as content analysis and participant observations, which might offer a deeper understanding of how consumers actually build their trust and make purchase decisions.

References

[1] N. Hajli and J. Sims, "Social commerce: The transfer of power from sellers to buyers," Technol. Forecast. Soc. Change, vol. 94, pp. 350–358, 2015.

[2] K. Smith, "49 Incredible Instagram Statistics," 2019.

[3] Statista, "Social Media Usage Worldwide," 2019.

[4] Crowd Analyzer, "State of Social Media 2018," 2018.

[5] I. Ben Yahia, N. Al-Neama, and L. Kerbache, "Investigating the drivers for social commerce in social media platforms: Importance of trust, social support and the platform perceived usage," J. Retail. Consum. Serv., vol. 41, no. September 2017, pp. 11–19, 2018.

[6] S. Abed, "An empirical examination of Instagram as an s-commerce channel," J. Adv. Manag. Res., vol. 15, no. 2, pp. 146–160, 2018.

[7] J. W. S. Che and C. M. K. Cheung, "Consumer Purchase Decision in Instagram Stores : The Role of Consumer Trust," in Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 24–33, 2017.

[8] C. and I. T. Commission, "ICT Report: E-Commerce in Saudi Arabia," 2017.

[9] M. Aldhahery, M. R. Wahiddin, M. A. Khuhro, and Z. A. Maher, "Investigation of Adoption Behaviour for Social Commerce in the Kindom of Saudi Arabia," in Proceedings 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences, ICETAS 2018, pp. 1–4. 2018.

[10] S. Abed, Y. Dwivedi, and M. Williams, "Consumers' Perceptions of Social Commerce Adoption in Saudi Arabia," in Proceedings 14th Conference on e-Business, e-Services and e-Society (I3E), 2015.

[11] E. Djafarova and C. Rushworth, "Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users," Comput. Human Behav., vol. 68, pp. 1–7, 2017.

[12] M. De Veirman, V. Cauberghe, and L. Hudders, "Marketing through instagram influencers: The impact of number of followers and product divergence on brand attitude," Int. J. Advert., vol. 36, no. 5, pp. 798–828, 2017.

[13] Kaitlin Augustine (Civic Science), "Influencer Marketing," 2019.

[14] L. V. Casaló, C. Flavián, and S. Ibáñez-Sánchez, "Influencers on Instagram: Antecedents and consequences of opinion leadership," J. Bus. Res., no. July, pp. 0–1, 2018.

[15] B. Godey et al., "Social media marketing efforts of luxury brands: Influence on brand equity and consumer behavior," J. Bus. Res., vol. 69, no. 12, pp. 5833–5841, 2016.

[16] N. Hajli, "Social commerce constructs and consumer's intention to buy," Int. J. Inf. Manage., vol. 35, no. 2, pp. 183–191, 2015.

[17] M. J. M. Razi, M. Sarabdeen, M. I. M. Tamrin, and A. C. M. Kijas, "Influencing Factors of Social Commerce Behavior in Saudi Arabia," in Proceedings 2019 International Conference on Computer and Information Sciences (ICCIS), pp. 1–4, 2019.

[18] S. Sharma, P. Menard, and L. A. Mutchler, "Who to Trust? Applying Trust to Social Commerce," Journal of Computer Information Systems, vol. 59, no. 1. pp. 32–42, 2019.

[19] N. Hajli, J. Sims, A. H. Zadeh, and M. O. Richard, "A social commerce investigation of the role of trust in a social networking site on purchase intentions," J. Bus. Res., vol. 71, pp. 133–141, 2017.

[20] L. Esmaeili and S. A. Hashemi G, "A systematic review on social commerce," J. Strateg. Mark., vol. 27, no. 4, pp. 317–355, 2019.

[21] L. Esmaeili, S. Mardani, M. Mutallebi, and S. Golpayegani, "Studying the Affecting Factors on Trust in Social Commerce," Int. J. Adv. Stud. Comput. Sci. Eng. IJASCSE, vol. 4, no. 6, 2015.

[22] E. Turban, J. Whiteside, D. King, and J. Outland, Introduction to Electronic Commerce and Social Commerce, Fourth edi. Springer, 2017.

[23] H. Han, H. Xu, and H. Chen, "Social commerce: A systematic review and data synthesis," Electronic Commerce Research and Applications, vol. 30. pp. 38–50, 2018.

[24] Socialize, "Social media and digital in Saudi Arabia: 2019 report," 2019.

[25] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust." Acad. Manag. Rev., vol. 20, no. 3, pp. 709–734, 1995.

[26] S. Shi and W. S. Chow, "Trust development and transfer in social commerce: Prior experience as moderator," Industrial Management and Data Systems, vol. 115, no. 7. pp. 1182–1203, 2015.

[27] N. Hajli, Y. Wang, M. Tajvidi, and M. S. Hajli, "People, Technologies, and Organizations Interactions in a Social Commerce Era," IEEE Trans. Eng. Manag., vol. 64, no. 4, pp. 594–604, 2017.

[28] L. Chen and R. Wang, "Trust Development and Transfer from Electronic Commerce to Social Commerce : An Empirical Investigation," no. May, pp. 568–576, 2016.

[29] S. Farivar, Y. Yuan, and O. Turel, "Understanding Social Commerce Acceptance: The Role of Trust, Perceived Risk, and Benefit," Twenty-second Am. Conf. Inf. Syst., pp. 1–10, 2016.

[30] H. Hairudin, H. Mohamed Dahlan, M. H. Selamat, and A. R. Che Hussin, "Follower's quality factor in social commerce," J. Phys. Conf. Ser., vol. 1196, no. 1, 2019.

[31] D. Sirdeshmukh, J. Singh, and B. Sabol, "Consumer Trust, Value, and Loyalty in Relational Exchanges," J. Mark., vol. 66, no. 1, pp. 15–37, 2002.

[32] H. Alhulail, M. Dick, and A. Abareshi, "Factors that Impact Customers' Loyalty to Social Commerce Websites," in Proceedings International Conference on Information Resources Management (CONF-IRM), 2018.

[33] Ministry of Commerce and Investment, "'Maroof: an Initiative to Support e-Commerce in the Kingdom," 2018.

[34] C. Egger, "Identifying Key Opinion Leaders in Social Networks," TH Ko˙ln - University of Applied Sciences Institute of Informatics, 2016.

[35] S. M. Aghdam and N. J. Navimipour, "Opinion leaders selection in the social networks based on trust relationships propagation," Karbala Int. J. Mod. Sci., vol. 2, no. 2, pp. 88–97, 2016.

[36] Y. WANG, "The Influence of Opinion Leaders towards Consumer Information Adoption in the Virtual Communities of Consumption," DEStech Trans. Soc. Sci. Educ. Hum. Sci., no. apme, pp. 444–451, 2017.

[37] A. Algi and I. Irwansyah, "KOL (Key Opinion Leader) as Consumer Trust Factor at Instagram Store," 2nd Indo IGCC, pp. 459–462, 2018.

[38] "What are KOLs? key opinion leader definition, examples & more," mediakix. [Online]. Available: https://mediakix.com/blog/kols-key-opinion-leaders-definition-influencers/.

[39] Djmira and C. Rushwortafarova, Elh, "Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users," Comput. Human Behav., vol. 68, pp. 1–7, 2017.

[40] Hashoff, "Influencer marketer: A #Hashoff state of the union report.," 2017.

[41] A. Algi and Irwansyah, "Consumer trust and intention to buy in Indonesia instagram stores," Proc. - 2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2018, pp. 199–203, 2019.

[42] V. A. Einarsdóttir, "From celebrities to the girl next door: Influencer marketing with a special focus on the social media platform, Instagram," University of Iceland, 2017.

[43] N. C. S. (NCS), "Sales Effect Study:Nielsen Catalina Solutions (NCS)," 2016.

[44] K. Sokolovaa and H. Kefia, "Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction in uence purchase intentions," J. Retail. Consum. Serv., 2019.

[45] E. L. Thorndike, "A constant error in psychological ratings," J. Appl. Psychol., vol. 4, no. 1, pp. 25–29, 1920.

[46] C. Lou and S. Yuan, "Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media," Journal of Interactive Advertising, vol. 19, no. 1. pp. 58–73, 2019.

[47] X. J. Lim, A. R. bt Mohd Radzol, J.-H. (Jacky) Cheah, and M. W. Wong, "The Impact of Social Media Influencers on Purchase Intention and the Mediation Effect of Customer Attitude," Asian J. Bus. Res., vol. 7, no. 2, pp. 19–36, 2017.

[48] C. Liu, Z. Bao, and C. Zheng, "Exploring consumers' purchase intention in social commerce: An empirical study based on trust, argument quality, and social presence," Asia Pacific J. Mark. Logist., vol. 31, no. 2, pp. 378–397, 2019.

[49] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," Chemom. Intell. Lab. Syst., vol. 50, no. 1, pp. 1–18, 2000.

[50] J. F. Hair, M. Sarstedt, T. M. Pieper, and C. M. Ringle, "The Use of Partial Least Squares Structural Equation Modeling in Strategic Management Research: A Review of Past Practices and Recommendations for Future Applications," Long Range Plann., vol. 45, no. 5–6, pp. 320–340, Oct. 2012.

[51] R. Sekaran and R. Bougie, Research methods for business: A skill building approach, 7th ed. John Wiley & Sons., 2016.

[52] G. W. Cheung and C. Wang, "Current approaches for assessing convergent and discriminant validity with SEM: issues and solutions," in Academy of Management Proceedings, vol. 2017, no. 1, p. 12706, 2017.

[53] W. W. Chin, A. Gopal, and W. D. Salisbury, "Advancing the Theory of Adaptive Structuration: The Development of a Scale to Measure Faithfulness of Appropriation," Inf. Syst. Res., vol. 8, no. 4, pp. 342–367, Dec. 1997.

[54] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," J. Mark. Res., vol. 18, no. 1, p. 39, Feb. 1981.

[55] J. Hulland, "Use of partial least squares (PLS) in strategic management research: a review of four recent studies," Strateg. Manag. J., vol. 20, no. 2, pp. 195–204, Feb. 1999.

[56] B. Han and J. Windsor, "User's willingness to pay on social network sites," J. Comput. Inf. Syst., vol. 51, no. 4, pp. 31–40, 2011.

[57] H. Liu, H. Chu, Q. Huang, and X. Chen, "Enhancing the flow experience of consumers in China through interpersonal interaction in social commerce," Comput. Human Behav., vol. 58, pp. 306–314, 2016.

Evaluate Metadata of Sparse Matrix for SpMV on Shared Memory Architecture

Nazmul Ahasan Maruf[1], Waseem Ahmed[2]
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Sparse Matrix operations are frequently used operations in scientific, engineering and high-performance computing (HPC) applications. Among them, sparse matrix-vector multiplication (SpMV) is a popular kernel and considered an important numerical method for science, engineering and in scientific computing. However, SpMV is a computationally expensive operation. To obtain better performance, SpMV depends on certain factors; choosing the right storage format for the sparse matrix is one of them. Other things like data access pattern, the sparsity of the matrix data set, load balancing, sharing of the memory hierarchy, etc. are other factors that affect performance. Metadata, that describes the substructure of the sparse matrix, like shape, density, sparsity, etc. of the sparse matrix also affects performance efficiency for any sparse matrix operation. Various approaches presented in literature over the last few decades given good results for certain types of matrix structures and don't perform as well with others. Developers thus are faced with a difficulty in choosing the most appropriate format. In this research, an approach is presented that evaluates metadata of a given sparse matrix and suggest to the developers the most suitable storage format to use for SpMV.**

Keywords*—Sparse matrix vector multiplication; sparse matrix metadata; sparse matrix vector multiplication parallelization; shared memory architecture; sparse matrix storage formats; high performance computing

I. Introduction

Sparse matrix-vector multiplication (SpMV) is an essential and frequently used kernel in high-performance computing (HPC), scientific and engineering applications. The operation $y = Ax$ is performed by the SpMV kernel, where A is a sparse matrix of size $M \times N$, and y and x are dense vectors of size M. Although SpMV is one of the most popular and essential kernels, it usually performs poorly for large sparse matrices. As has been shown by Goumas and others, SpMV achieves less than 10% of the peak performance of microprocessors [1]. Higher performance for SpMV depends on various factors - the choice of the right sparse matrix storage format is one of them. Other factors like data access pattern, the sparsity of the matrix data set, load balancing, sharing of the memory hierarchy, etc. are important when we work on shared memory architecture. In shared memory architectures, performance degradation might happen when all the processors try to access the memory simultaneously. Applications on shared memory systems that have no data dependencies and have good temporal locality tend to perform well. On the other hand, contention on memory subsystems in other shared memory applications with streaming access patterns results in poor performance [2]. In literature, we found that the SpMV kernel performs poorly on a shared memory system because of its streaming access patter [2]. In shared memory architecture, memory bandwidth also is a performance bottleneck for SpMV operations when operating on large matrices [3]. Each element of the matrix is only used once in the SpMV operation. Irregular data access is another performance bottleneck for SpMV operation when we use sparse matrix storage formats like COO and CSR. For example, when using Coordinate format (COO) for large scale SpMV, high cache miss rate and poor performance were noticed as a result of indirect addressing. Irregular data accessing in SpMV also results in reduction of performance.

To address these kinds of problems, researchers have proposed different sparse matrix storage formats [2], [4]–[6]. For example, Compressed Sparse Row (CSR) has been proposed to address space overhead but CSR does not reduce irregular data accesses completely if non-zero elements (*nnz*) elements are mostly in a certain row. Also, when parallelizing CSR on shared-memory architectures for operations on such matrices, load balancing becomes an issue. Additionally, matrix substructures also affect performance efficiency for all sparse matrix storage format; storage formats that perform well with certain substructures do not perform well with others. Take for example, ELLPACK [7]. The ELLPACK storage format is very suitable for diagonal substructures but not for horizontal substructures. SpMV performance also depends on other metadata like diagonal density, row or column-major order, max non-zero values (row thickness) per row, etc. Another example is that if the sparse matrix is not pattern symmetric and the row thickness is very large, CSR performs better than ELLPACK. Thus, it's important for a developer to know beforehand the metadata of a sparse matrix to help him decide which storage format to use. In this research, we propose new metrics to describe sparse matrices to add to existing metadata description which will help developers to take decisions easily. Our work was motivated by two key observations

1) Different storage formats give higher performance for particular substructures of matrices and lower on others.
2) Most storage formats are not based on metadata or substructures of the sparse matrices.

A tool was developed to generate the metrics to help developers find the optimal storage format to use for any given sparse matrix. The rest of the paper is organized as follows.

Section 2 presents some popular and relevant storage formats.Related work discussed in Section 3. Section 4 describes the motivation and approach for our metrics, tool,and also described the sparse matrix benchmarks used in this research.

Section 5 presents the empirical analysis using metrics and experimental evaluation. Section 6 discusses future work and concludes the paper.

II. Storage Formats

There are many storage formats for sparse matrices described in literature [8] [9]–[14] for the SpMV operation. Some of the important and popular storage formats are described in the next sub sections.

To better understand the storage formats an example sparse matrix A of size 5×5 and with elements described in Fig. 1 is used.

$$A_{5\times5} = \begin{bmatrix} 0 & 4 & 0 & 7 & 0 \\ 2 & 0 & 3 & 0 & 6 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 6 & 0 \end{bmatrix}$$

Fig. 1. Sample Matrix for study

A. Coordinate (COO)

COO is one of the earliest and most basic format for storing sparse matrices [15] [16] and is very simple and reliable. This format stores the row index, column index and number of non-zero (nnz) values in three one-dimensional array - one for storing non zero values, one for storing row indices and one for column indices. Fig. 2 shows the COO format for matrix A given in Fig. 1.

$$Data = \begin{bmatrix} 4 & 7 & 2 & 3 & 6 & 5 & 2 & 1 & 6 \end{bmatrix}$$

$$Row_i = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 2 & 3 & 4 & 4 \end{bmatrix}$$

$$Col_i = \begin{bmatrix} 1 & 3 & 0 & 2 & 4 & 1 & 4 & 0 & 3 \end{bmatrix}$$

Fig. 2. COO Format With Data, Row And Column Field

The storage requirement for COO to store a matrix of dimension $M \times N$ with NNZ non-zero values is

$$COO_{storage} = 3 \times NNZ$$

B. Compressed Sparse Row (CSR)

Compressed sparse row (CSR) format is the most popular and widely used sparse formats [17]. SpMV with CSR format gives good performance and is used in libraries like BLAS and LAPACK. Like COO, CSR also needs three 1-D arrays for storing data; one holds the *nnz* values, another one is for the number of *nnz* values per row, and one for column indices. Fig. 3 shows the CSR format for matrix A given in Fig. 1.

The storage requirement for CSR to store a matrix of dimension $M \times N$ with NNZ non-zero values is

$$CSR = 2 \times NNZ + M + 1$$

$$Data = \begin{bmatrix} 4 & 7 & 2 & 3 & 6 & 5 & 2 & 1 & 6 \end{bmatrix}$$

$$Rowptr = \begin{bmatrix} 0 & 2 & 5 & 6 & 7 & 9 \end{bmatrix}$$

$$Col_i = \begin{bmatrix} 1 & 3 & 0 & 2 & 4 & 1 & 4 & 0 & 3 \end{bmatrix}$$

Fig. 3. CSR Format With Data, Row And Column Field

C. Compressed Sparse Column (CSC)

The Compressed Sparse Column (CSC) [18] [19] is similar to the Compressed sparse row (CSR). The main diference between them is that CSC uses column pointer instead of row pointer. *CSC* uses three 1-D array for storage; one for nnz values, one for column pointer and one for row indices. Fig. 4 shows the CSC format for matrix A given in Fig. 1.

$$Data = \begin{bmatrix} 4 & 7 & 2 & 3 & 6 & 5 & 2 & 1 & 6 \end{bmatrix}$$

$$colptr = \begin{bmatrix} 0 & 2 & 4 & 5 & 7 & 9 \end{bmatrix}$$

$$row_i = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 2 & 3 & 4 & 4 \end{bmatrix}$$

Fig. 4. CSC Format With Data, Row And Column Field

The storage requirement for CSC to store a matrix of dimension $M \times N$ with NNZ non-zero values is

$$CSC = 2 \times NNZ + N + 1$$

D. ELLPACK

Another commonly referenced format is ELLPACK. It is well suited for semi-structured and unstructured meshes and for vector architectures [7]. It is also a good storage format for diagonal matrices. It is particularly suited in cases where the maximum number of non-zero values per row does not differ more with the average of non-zero elements in all rows. ELLPACK uses two 2D matrices where one is for storing the *nnz* values in row-wise order and another is for storing column indices of the *nnz* values. If our matrix size is $M \times N$ and $max(N_{nzr})$ presenting the maximum non zero values per row then the column indices 2D array has a size of $M \times max(N_{nzr})$. So the storage for ELLPACK shows in Fig. 5 for matrix A given in Fig. 1.

$$Data = \begin{bmatrix} 4 & 7 & \\ 2 & 3 & 6 \\ 5 & & \\ 2 & & \\ 1 & 6 & \end{bmatrix} Col_i = \begin{bmatrix} 1 & 3 & \\ 0 & 2 & 4 \\ 1 & & \\ 4 & & \\ 0 & 3 & \end{bmatrix}$$

Fig. 5. ELLPACK Format With Data And Column Field

At the implementation level, there are two ways to store the ELLPACK format. One way is to use 2-D arrays. In this

case, the storage requirement for ELLPACK to store a matrix of dimension $M \times N$ with NNZ non-zero values is

$$ELLPACK_{storage} = 2(max(N_{nzr}) \times m)$$

E. Compressed Sparse Row-DU (CSR-DU)

In the CSR-DU [4] format, instead of using col_index and row_ptr (or column_ptr) like in CSR (or CSC), *ctl* a single byte array is used. This array consists of four elements: *uflags*, *usize*, *ujump*, *ucis*. *uflags* and *usize* (to identify the type of unit and size), *ujump* (a variable length which denotes the first column index of each unit), and *ucis* (an array which denotes the distance between the column index of the first element of the unit and the column index of the previous element). Fig. 6 shows the CSR-DU format for matrix A given in Fig. 1.

units	uflags	usize	ujump	ucis
0	U8,NR	2	1	2
1	U8,NR	3	0	2,2
2	U8,NR	1	1	-
3	U8,NR	1	4	-
4	U8,NR	2	0	3

$Data = [\ 4\ \ 7\ \ 2\ \ 3\ \ 6\ \ 5\ \ 2\ \ 1\ \ 6\ ]$

Fig. 6. CSR-DU Format with data and ctl values

F. Compressed Sparse Row-VI(CSR-VI)

This format is extension of the CSR format. In CSR-VI [4], the *values* array of CSR is replaced with two arrays, *val_unique*: which contains the unique values of the matrix and *val_ind*: the index of the value in the vals_unique array for each of the nnz matrix element. Fig. 7 shows the CSR-VI format for matrix A given in Fig. 1.

$val - indx$ = [1 3 0 2 4 1 4 0 3]
$val - unique$ = [4 7 2 3 6 5 1]

Fig. 7. CSR-VI Format with Values

G. Compressed Sparse eXtended (CSX)

CSX [2] is also an extension of CSR-DU where CSX support different classes of regularities. In CSX, run time code generation method is employed. In CSX, five types of substructures are considered: horizontal, vertical, diagonal, anti-diagonal, and 2-D substructures. For any substructures, CSX uses run length encoding on delta values for more aggressive index compression. For 2-D substructures, transformation is used to convert different substructures to horizontal substructures. Fig. 8 shows the CSX format for matrix A given in Fig. 1.

$delta - values$ = [2 2 2 3]
$indices$ =[1 3 0 2 4 1 2 0 3]

Directions	Elements	
	y	x
Horizontal	y_o	$x_o + i\delta$
Vertical	$y_o + i\delta$	x_o
Diagonal	$y_o + i\delta$	$x_o + i\delta$
Anti-Diagonal	$y_o + i\delta$	$x_o - i\delta$

Fig. 8. CSX with data directions and transformation

III. Related Works

Here presenting some of related works, where researcher want to optimize SpMV with respect to sparse storage format.

Kenli Li, analysis and optimize SpMV using probabilistic method, they present probability mass function (PMF) for non zero elements in sparse matrix, they also evaluate the performance and efficiency of COO, CSR, ELL and HYB and find out the optimistic solution for SpMV using probability mass function (PMF), they also observed that different most of the matrices have their own most appropriate storage format for achieving best performance [15].

Daniele Buono, works with shared memory multiprocessor, they evaluate a new methodology to implement SpMV on shared memory multiprocessor which has two phases, one is for building a scalar matrix and on the second phase at first they reduce the scalar matrix by row by providing numerous opportunities to memory location, in this paper they use CSR as there baseline because CSR is inefficient with very sparse and graphs [20].

Zhang, on there paper they want to optimize data locality of sparse matrix vector multiplication algorithm by improving the sparse matrix storage format, they proposed cache oblivious extension quadtree storage structure (COEQT), where the sparse matrix is recursively divided into sub module and that will also fit into the cache by doing this it will improve the data locality [21].

Xiaowen Feng, proposed a new storage format for Sparse matrix name Segmented Interleave Combination (SIC), Instead of using compressed sparse row (CSR) they proposed this format because they find out a problem if the non zero variables are very high then thread divergence can cause and that will affect the performance while using CSR, so they combined the CSR values and proposed a new storage format name SIC [22].

Arash Ashari, proposed Blocked row column (BRC) storage format that optimize the Sparse Matrix vector with addressing the thread divergence, redundant computation in GPUs in the paper they also find out the optimizing challenges which are thread divergence, load imbalance, non-coalesced, and memory access [10].

Yang, introduce a SpMV that represent large graphs which has noticeable performance, they marge the ideas form Transposed Jagged Diagonal storage (TJDS) [23] with coo [24].

Francisco Vazquez, evaluated a new approach based on ELLR-T kernel on GPU using CUDA where they find out the performance of this specific kernel based on the thread block size and the number of accumulated threads, they also proposed a auto tuning model based on memory access for GPU for ELLR-T kernels [25].

Baskaran, they optimize SpMV on compilation and run time basis, in compilation time they include synchronization

free parallelism, thread mapping, optimize memory access and data redundancy on their optimization they proposed a new Blocked storage format and implement it on GPUs [26].

Shizhao Chen, perform a comprehensive study on representation of sparse matrix on Intel Knights landing XeonPhi and ARM-based FT-200PLUS architecture, they found that best representation of sparse relay on architecture and the unit of program, in their paper they use very well known CSR,CSR5,ELL,SELL and HYB sparse matrix storage format [27].

Weifeng Liu, they introduce a new storage format CSR5 (compressed sparse row) which provides high throughput SpMV on CPUs,GPUs and Xeon Phi, in the paper they also mention an efficient storage format should agree with two criteria, one is it should avoiding structure-dependent parameter tuning, another one is it should support fast sparse matrices vector for regular and irregular matrices [28].

IV. Metric

A. Motivation

For the SpMV kernel, the actual and maximum performance strongly depends on the characteristics and nature (substructure) of a sparse matrix. Any generalized storage format like CSR, COO, ELLPACK, etc. does not make any assumption about the shape and metadata of an input sparse matrix [2]. From our research, we found that some storage formats perform better on some sets of sparse matrices and perform poorly on some. In Fig. 9, performance evaluation for the basic and essential storage formats with GFLOPS (Floating point operation per second) is presented. As described earlier, performance of the SpMV operation depends on data access pattern, data dependency among storage arrays in some formats, load balancing when parallelizing the operation on shared memory architecture, sharing of the memory hierarchy, choosing the right storage format for specific sparse matrices. For choosing the right storage format, we need to know the nature of the sparse matrix, but storage format do not make any assumptions about the sparse matrix substructure.

In our research, we find out that shape and substructure of the sparse matrix plays an important role in choosing the appropriate storage format. For example, any horizontal shaped sparse matrix, CSR will perform better than ELLPACK. Another example is ELLPACK which will perform better on the diagonal shape of data. Additionally, we need to use the thickness of the diagonal. If the thickness of diagonal increases then the performance of ELLPACK will be decreased, because for ELLPACK the 2D array has a size of *m*×*Nmnzr* where *Nmnzr* is the maximum non zero values per row. If the number of non-zero values differs more with the average then ELLPACK performance will be decreased. Other characteristics (described in later sections) like symmetry, density, sparsify, etc. also helps in deciding the most appropriate storage formats for sparse matrix.

Besides COO, CSR (and its variances), CSC and ELLPACK, this extends to other proposed storage formats in literature. Table III describe, proposed sparse storage formats and their sparse matrix characteristics. From Table III Bell and Garland's HYBRID(HyB) format which is a hybrid of ELLPACK and COO, is generally the fastest format for unstructured matrices [7]. Hiroki Yoshizawa claims that performance of SpMV with compressed sparse row (CSR) storage format depends on selection of parameter. In the paper, they also mention that conjugate gradient method is one of the popular iterative methods for solving SpMV; they put their focus on optimizing thread mapping for the CSR; they propose an efficient algorithm for automatic selection of optimal parameter on GPU, which is generally performing better then CUSPARSE when the diagonal density is higher [29]. F. Vazquez [25] proposed and evaluated implementation of ELLR-T which is based on ELLPACK storage format, and the evaluation goes with some parameters like number of rows, total number of *nnz* elements, average number of entries with respect to rows, difference of the maximum number of entries in a row and average entries with respect to row, percentage standard deviation of entries. Their proposed approach performs better with less diagonal density data.

B. Bench Mark

In this paper, we will use a wide range of sparse matrix with different structural and numerical properties. We collect our sparse matrix from University of Florida sparse matrix collection (UFSMC) [30], which is the standard and most popular benchmark in SpMV research. Table I shows the benchmarks used for the experimental evaluation.

C. Metrics

To better understand the substructure of a sparse matrix and to automate the decision of choosing the best storage format given a sparse matrix, a set of newer metrics are needed to complement the existing metadata used to describe the substructure of a sparse matrix. This subsection describes these newer metrics in more detail.

1) *rowThickness* - In a given sparse matrix, the maximum number of *nnz* values per row is defined here as *rowThickness* and is given by Equation1, where r_i indicates the number of *nnz* values in row *i*.

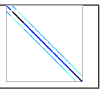
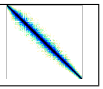
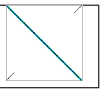
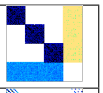
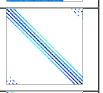
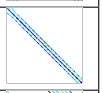
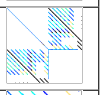
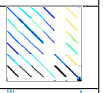
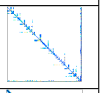
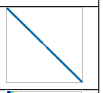
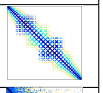
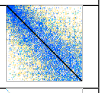
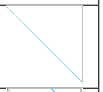
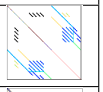
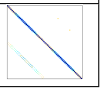
$$rowThickness = max(r_1, r_2, r_3, ...r_N) \quad (1)$$

The SpMV performance for any storage format depends on *rowThickness*.For example, if $rowThickness \ll avg(r_1, r_2, r_3, ...r_N)$ the performance of ELLPACK storage format decreases. In this case, CSR, COO or other formats will be the better choice for developers. On the other hand when $rowThickness \approx avg(r_1, r_2, r_3, ...r_N)$ ELLPACK performs better.

2) *Pattern Symmetry* - The sparse (square) matrix will be pattern symmetric if the existence of *nnz* entries match across the diagonal [30]. To better understand pattern symmetry, let's consider a square matrix P of size $M \times M$. This matrix will be pattern symmetric if $\forall i, jP(i, j) \neq 0 \rightarrow P(j, i) \neq 0$.

3) *Numeric Symmetry* - The sparse matrix will be numeric symmetric if the *nnz* entries numerically match across the diagonal [30]. Consider a square sparse matrix P of size $M \times$M. Matrix P will be pattern symmetric if $\forall i, jP(i, j) = P(j, i)$ or $P = P^T$.

TABLE I. THE BENCHMARKS USED FOR THE EXPERIMENTAL EVALUATION.

No.	Matrix Name	Rows	Columns	Non-zeros	Plot
1	sherman3	5,005	5,005	20,033	
2	airfoil_2d	14,214	14,214	259,688	
3	bbmat	38,744	38,744	177,1722	
4	ill_stokes	20,896	20,896	131,368	
5	gt01r	7,980	7,980	430,909	
6	cavity26	4,562	4,562	138,040	
7	lowthrust	18,476	18,476	224,897	
8	windtunnel	40,816	40,816	803,978	
9	circuit	12,127	12,127	48,137	
10	qa8fm	66,127	66,127	1,660,579	
11	ga3as3h12	61,349	61,349	5,970,947	
12	nd24k	72,000	72,000	28,715,634	
13	bcsstm	15,439	15,439	15,439	
14	robot	2,358	2,358	18,218	
15	dw2048	2,048	2,048	10,114	

4) *Horizontal Symmetry* - A sparse matrix P of size $M \times N$ will be horizontal symmetric if $\forall i \forall j P(i,j) = P(M-i, M-j)$.
5) *Vertical Symmetry* - A sparse matrix P of size $M \times N$ will be vertical symmetric if $\forall i \forall j P(i,j) = P(N-i, N-j)$.
6) *Row or Column-major Order* - A Sparse matrix entries stored in a file like $\forall e =< row, col, val >$. Row-major order for that stored sparse matrix will be $\forall e_{row}(e_{row} \leq e_{row+1})$ and column-major order will be $\forall e_{col}(e_{col} \leq e_{col+1})$.
7) *Density and Sparsity* - Sparsity of a sparse matrix will be $sparsity = zeroElements/(M \times N)$.On the other hand, $Density = nnz/(M \times N)$ will give us density of a sparse matrix. Density and sparsity are related as $sparsity = 1 - density$. Thus, any matrix with $sparsity \geq 0.5$ is generally considered as sparse otherwise the matrix is considered dense.
8) *Diagonal Density* - The *diagonal density* for any square sparse matrix is calculated by finding out the count of *nnz* values on the diagonal. Equation2 describe the diagonal density of a sparse matrix.

$$count(\forall i P(i,i) \neq 0) \tag{2}$$

9) *Upper Triangle Density* - The upper triangle density for any sparse matrix calculated by finding out the count of upper triangle *nnz* values where row indices are less then column indices. Equation 3 describe the upper triangle density of a sparse matrix.

$$count(\forall i \forall j P(i,j) = P(i,j) : j > i) \tag{3}$$

10) *Lower Triangle Density* - The lower triangle density for any sparse matrix calculated by finding out the count of lower triangle *nnz* values where row indices are greater then column indices. Equation describe the lower triangle density of a sparse matrix.

$$count(\forall i \forall j P(i,j) = P(i,j) : i > j) \tag{4}$$

D. Metric Generation

In this subsection, we describe the algorithms for generating the metrics to describe the substructure and shape of the sparse matrix. These metrics help to suggest the best storage format for a given sparse matrix. For generating the metrics, at first we will find out the *rowThickness*, in *rowThickness* we will find the maximum number of non zero values per row. Algorithm 1 describe rowThickness function.

Algorithm 1 Maximum Number per rows(*rowThickness*)

```
1:  function ROWTHICKNESS(col)
2:      count[col]++
3:      maxcol ← 0
4:      for i ← 1, col do
5:          if count[i] > maxCol then
6:              maxCol ← count[i]
7:          end if
8:      end for
9:      return maxCol
10: end function
```

Next, for getting more specific knowledge about the sparse matrix we need to find out the information about *pattern symmetric*, our algorithm will find the pattern symmetric by matching the upper triangle and lower triangle value, if it is the same the sparse matrix will be *pattern symmetric*. Algorithm 2 describe *pattern Symmetric* function.

After that, we will find out the percentage of *numeric, horizontal and vertical symmetric*. In *numeric symmetric*, our algorithm will find the percentage of the same upper triangle and lower triangle *nnz* values which are separated by a diagonal. Algorithm 3 describe *numeric Symmetric* function.

Algorithm 2 Pattern Symmetric

```
1: function PATTERNSYMMETRIC(row_indices, col_indices)
2:     for row ← 1, row_indecies do
3:         for col ← 1, col_indecies do
4:             if row < col then
5:                 up ← up + 1
6:             end if
7:             if row > col then
8:                 lw ← lw + 1
9:             end if
10:        end for
11:    end for
12:    if up! = lw then
13:        patternSym == 1
14:    else
15:        patternSym == 0
16:    end if
       return patternSym
17: end function
```

Algorithm 3 Numeric Symmetric

```
1: function NUMERICSYMMETRIC(row_indices, col_indices)
2:     for row ← 1, row_indecies do
3:         for col ← 1, col_indecies do
4:             if data[row][col]! = data[col][row] then
5:                 nCount ← nCount + 1
6:             end if
7:         end for
8:     end for
9:     numeric = (nCount/nnz) * 100
10:    return numeric
11: end function
```

In *horizontal and vertical* algorithms, we will check the percentage of the same value which is presented horizontally and vertically. Algorithm 4, 5 describes the horizontal and vertical symmetric function.

Algorithm 4 Horizontal Symmetric

```
1: function HORIZONTALSYMMETRIC(row_indices, col_indices)
2:     for row ← 1, row_indecies/2, AND, nrow ← row_indecies − 1, nrow ← nrow − 1 do
3:         for col ← 1, col_indecies do
4:             if data[row][col]! = data[nrow][col] then
5:                 hCount ← hCount + 1
6:             end if
7:         end for
8:     end for
9:     horizontal = (hCount/nnz) * 100
10:    return horizontal
11: end function
```

Row and column-major order is also an important characteristic to find out the appropriate storage format for a particular sparse matrix. Our algorithm will find the row and column-major order by checking the row and column values from the data set are sorted or not. If row indices are sorted then it will print *row-major order* if column indices are sorted then it will print *column-major order*. Algorithm 6 describe

Algorithm 5 Vertical Symmetric

```
1: function VERTICALSYMMETRIC(row_indices, col_indices)
2:     for row ← 1, col_indecies/2, AND, ncol ← col_indecies − 1, ncol ← ncol − 1 do
3:         for col ← 1, col_indecies do
4:             if data[row][col]! = data[ncol][col] then
5:                 vCount ← vCount + 1
6:             end if
7:         end for
8:     end for
9:     vertical = (vCount/nnz) * 100
10:    return vertical
11: end function
```

the row and column-major order. After finding the *row and column-major order*, we will approach for sparsity and density, Algorithms will find out the *sparsity* by dividing the zero-elements with row multiple column and *density* by dividing the *nnz* by row multiple column. Algorithm 7 describe the sparsity and density functions. After that we will find out the *density* of the *diagonal, upper, lower triangle*. For calculating the percentage of *diagonal density*, our algorithm at first find out the diagonal values where row and column indices are same then *nnz* elements will be divided by those diagonal values and for upper and lower triangle density our tools will find out the upper and lower values where row indices are less or greater then column indices, respectively then *nnz* elements again divided by the upper and lower values. Algorithm 8 describe the diagonal density function where upper and lower triangle also describe. After evaluating those results metrics will help to suggest developers for most suitable storage formats for their particular sparse matrix.

Algorithm 6 Row and Column major

```
1: function ROWCOLMAJOR(nnz)
2:     for row ← 1, nnz − 1 do
3:         if rowdata[row] > rowdata[row + 1] then
4:             dataorder == 1
5:         else
6:             dataorder == 0
7:         end if
8:     end for
9:     for col ← 1, nnz − 1 do
10:        if coldata[col] > coldata[col + 1] then
11:            dataorder == −1
12:        else
13:            dataorder == 0
14:        end if
15:    end for
16:    return dataorder
17: end function
```

V. Experimental Evaluation

A. Experimental Setup

The experiments have been performed on Aziz Super-Computer at King Abdulaziz University, Saudi Arabia. In experiment, one nodes with 16 processors is being requested.

Algorithm 7 Sparsity and Density of Matrix data

```
function SPARSIFY(row, col, nnz)
    rowCol = [row] * [col]
    zeroElement = rowCol − nnz
    sparsify = zeroElement / rowCol
    return sparsify
end function
function DENSITY(row, col, nnz)
    rowCol = [row] * [col]
    density = nnz / rowCol
    return density
end function
```

Algorithm 8 Upper, Lower Triangle and Diagonal Density

```
function DIAGONALDENSITY(row_indices, col_indices)
    for row ← 1, row_indecies do
        for col ← 1, col_indecies do
            if row < col then
                upper = nnz/(up ← up + 1)
            else if row > col then
                lower = nnz/(lw ← lw + 1)
            else if row = col then
                diagonal = nnz/(dia ← dia + 1)
            end if
        end for
    end for
    return upper, lower, diagonal
end function
```

TABLE II. METRICS WITH SPARSE MATRIX SPMV PERFORMANCE(GFLOPS)

mtx	Diagonal Density(%)	Numeric Symmetry(%)	Row Thickness	Avg *nnz values*	COO	CSR	IDX	ELL
sherman	4.003	25	7	4	1.514	1.788	1.715	1.718
lll_stokes	9.158	11	12	9	9.419	14	13.186	11.949
airfoil_2d	18.27	5.5	23	18	13.033	17.588	16.318	15.365
gt01r	54	1.9	75	53	20.352	26.369	24.153	20.228
cavity	30.29	3.3	62	30	9.506	11.222	10.799	9.145
bbmat	45.73	2.2	132	45	39.555	49.206	40.264	22.618
lowthrust	6.364	16	7184	6	4.641	8.726	8.947	6.360
windtunnel	66.9	0.99	93	66	17.089	59.388	47.609	98.01
circuit	3.969	25	5682	3	2.699	4.103	4.009	3.017
qa8fm	13.06	7.7	14	13	19.675	33.065	32.265	28.224
bcsstm	1	1.00E+02	1	1	0.901	1.346	1.315	1.322
robot	4.162	24	23	4	0.762	0.881	0.886	0.898
dw2048	4.938	20	8	4	0.847	0.922	0.958	0.936
nd24k	199.9	0.5	483	199	47.167	71.32	58.514	462.15
ga3as3h12	49.16	2	1024	49	29.524	81.224	43.407	105.96

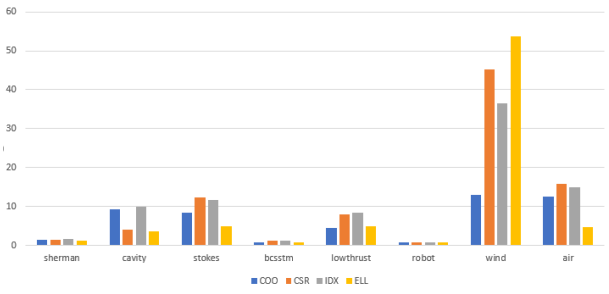


Fig. 9. SpMV performance in GFLOPS

TABLE III. PROPOSED STORAGE FORMATS BY AUTHORS

Authors name	Storage format	mtx	performance	Metrics
Bell and Garland [7]	HYB	Lowthrust	15.873	Pattern Symmetric =YES Numeric Symmetric=16% Horizontal and Vertical Sym = 7.9% Column Major Order =Yes Sparsity= 0.997 Density = 0.0003444 Diagonal Density = 6.364
		Circuit	5.987	Pattern Symmetric =YES Numeric Symmetric=25% Horizontal and Vertical Sym= 11% Column Major Order =Yes Sparsity= 0.997 Density = 0.0003273 Diagonal Density = 3.969
		Windtunnel	5.988	Pattern Symmetric =YES Numeric Symmetric=0.99% Horizontal and Vertical Sym = 7.1% Column Major Order =Yes Sparsity= 0.9984 Density = 0.001639 Diagonal Density = 66.9
Hiroki Yoshizawa [29]	CSR-T	qa8fm	5.988	Pattern Symmetric =YES Numeric Symmetric=7.7% Horizontal and Vertical Sym = 0.041% Column Major Order =Yes Sparsity= 0.9889 Density = 0.0111 Diagonal Density = 13.06
		ga3a3h2	9.001	Pattern Symmetric =YES Numeric Symmetric=2% Horizontal and Vertical Sym = 1% Column Major Order =Yes Sparsity= 1.006 Density = -0.005677 Diagonal Density = 49.16
		nd24k	12.214	Pattern Symmetric =YES Numeric Symmetric=0.5% Horizontal and Vertical Sym = 0.25% Column Major Order =Yes Sparsity= 0.9838 Density = 0.01619 Diagonal Density = 199.9
F.Vazquez [25]	Approach based on ELLR-T	rbsa480	3.4	Pattern Symmetric =YES Numeric Symmetric=2.8% Horizontal and Vertical Sym = 1.3% Column Major Order =Yes Sparsity= 0.9258 Density = 0.07417 Diagonal Density = 35.9
		dw2048	2.1	Pattern Symmetric =YES Numeric Symmetric=20% Horizontal and Vertical Sym = 7.2% Column Major Order =Yes Sparsity= 0.9976 Density = 0.002411 Diagonal Density = 4.938
		mhd32000	9.8	Pattern Symmetric =YES Numeric Symmetric=4.7% Horizontal and Vertical Sym = 2.4% Column Major Order =Yes Sparsity= 0.9934 Density = 0.006643 Diagonal Density = 21.26

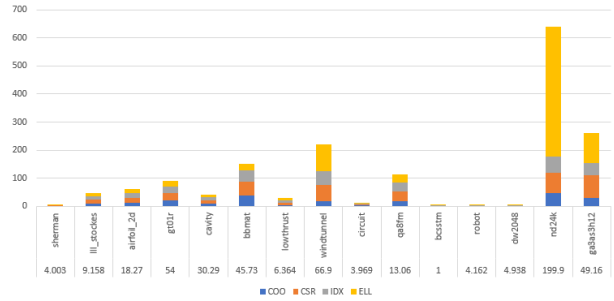


Fig. 10. SpMV performance evaluation respect to Diagonal Density

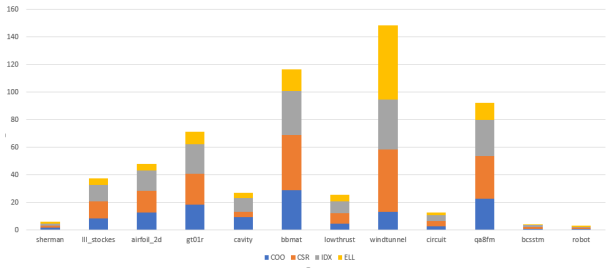


Fig. 11. SpMV performance(GFLOPS)

Each compute node has Dual socket hex-core processor running at 2.4GHz with 256GB RAM. Linux version 2.6.32-358.23.2.e16.x86_64 and GCC compiler version 4.4.7 is used in our experiment.

B. Metrics and Performance Analysis

In this section, Presents metrics evaluation using selected benchmark in previous chapter and also evaluate those similar benchmark according to the metrics behavior. In this approach, first generated metric is rowThickness. rowThickness is used to find out the maximum non-zero values per row. Row thickness is important for any sparse matrix because there is a relation between rowThickness and SpMV performance. If the row thickness does differ from the average non-zero values, then the ELLPACK storage format will not be a great choice for any data set. In this case, the experiment showed that CSR or any other storage format is more efficient then ELLPACK. From Table II section "Row Thickness" gives the generated rowThickness values for every sparse matrix. The sparse matrix characteristic's pattern symmetric is used in our approach for getting a clear visualization of the sparse matrix. If the sparse matrix is pattern symmetric any system can easily figure out that the row thickness does not very much higher from the average non-zero values.

Numeric, Horizontal and vertical symmetries are the other symmetric metrics of a sparse matrix, by using those parameters we can clearly understand the behavior and the structure of any sparse matrix. From the experiment, found that sparse matrix storage formats perform better with less numerically symmetric data. From the experiment, also found that if the percentage of numerically symmetric is less then horizontal and vertical ELLPACK perform better then other sparse matrix storage formats. In Table IV, will find out the symmetry information of selected sparse matrix. Row or column-major order showed in Table IV is another metric that is used in our approach. Experiment showed that, when a sparse matrix is in row-major order CSR and IDX perform better than other storage formats. The Table II sections "Sparsity and Density" gives the status of any matrix, those will justify the sparsity of any matrix using $sparsity = 1 - density$. Another section

TABLE IV. STORAGE FORMATS WITH METRICS

mtx file	metadata	mtx file	metadata
bcsstm	Pattern Symmetric =YES Numeric Symmetric=1e+02% Horizontal and Vertical Sym = 50% Column Major Order =Yes Sparsity= 0.9999 Density = 6.477e-05 Diagonal Density = 1	sherman	Pattern Symmetric =YES Numeric Symmetric=25% Horizontal and Vertical Sym = 9.7% Row Major Order =Yes Sparsity= 0.9992 Density = 0.0007997 Diagonal Density = 4.003
robot	Pattern Symmetric =YES Numeric Symmetric=24% Horizontal and Vertical Sym = 12% Column Major Order =Yes Sparsity= 0.9982 Density = 0.001765 Diagonal Density = 4.162	airfoil_2d	Pattern Symmetric =YES Numeric Symmetric=5.5% Horizontal and Vertical Sym = 2.7% Row Major Order =Yes Sparsity= 0.9987 Density = 0.001285 Diagonal Density = 18.27
gt10R	Pattern Symmetric =YES Numeric Symmetric=1.9% Horizontal and Vertical Sym = 0.93% Row Major Order =Yes Sparsity= 0.9932 Density = 0.006767 Diagonal Density = 54	bbmat	Pattern Symmetric =YES Numeric Symmetric=2.2% Horizontal and Vertical Sym = 1.1% Row Major Order =Yes Sparsity= 0.9988 Density = 0.00118 Diagonal Density = 45.73
Cavity26	Pattern Symmetric =YES Numeric Symmetric=3.3% Horizontal and Vertical Sym = 1.7% Row Major Order =Yes Sparsity= 0.9934 Density = 0.00664 Diagonal Density = 30.29	lll_stokes	Pattern Symmetric =YES Numeric Symmetric=11% Horizontal and Vertical Sym = 5.5% Row Major Order =Yes Sparsity= 0.9996 Density = 0.0004383 Diagonal Density = 9.158

from Table II diagonal density provide the description about diagonal density. Diagonal density is very important for any sparse matrix. Storage formats performance can determined by the percentage of diagonal density. Experiments showed that for higher percentage of diagonal density ELLPACK perform better then CSR. Fig. 10 showed performance of SpMV for different storage formats with different sparse matrix respect to diagonal density. Fig. 11 presents the performance graph respect to COO, CSR, IDX and ELLPACK. ELLPACK perform peak performance with *windtunnel* sparse matrix, where the percentage of numerical symmetry is 0.99%, diagonal density 66.9% and column major order. ELLPACK perform worst performance with bcsstm sparse matrix, where the percentage of numerical symmetry is 1e+02%, diagonal density 1% and column major order. After experimental analysis we found that ELLPACK perform best with higher diagonal density and give lowest performance where diagonal density is very low. CSR perform best among the sparse matrix with bbmat, where the percentage of numerical symmetry is 2.2%, diagonal density 45% and row-major order data.While the percentage of numerical symmetry 24%, diagonal density is 4.162% and column-major order then CSR perform worst.Our experimental result showed that CSR perform worst when the nature of the sparse matrix is column-major order. Any row-major order sparse matrix with better percentage of diagonal density CSR perform peak performance. For any column-major order sparse matrix we can use another solution which also compressed the data

but here the format used column pointer instead of row pointer, named CSC. On the other hand COO performed average for all the sparse matrix, COO give us its peak performance with bbmat but the performance is not greater than CSR. COO also perform worst with bcsstm, where the percentage of numerical symmetry is 1e+02%, diagonal density 1% and column-major order. For bcsstm sparse matrix CSR gives better performance than other three storage format used in our literature. Another newly storage format which we used in our tools is IDX, these storage format perform better then all others with lowthrust sparse matrix, where the percentage of numerical symmetry 16%, diagonal density is 6.364% and column-major order. IDX performance is always very close to CSR.

VI. Conclusion

SpMV operation's performance affects the efficiency of large number of applications like CFD, computer graphics, robotics, structural problems, acoustics problems, etc. The performance of SpMV operations depends on many factors. Matrix characteristics, storage formats, software and hardware implementations are most common problems for the performance of SpMV operation.

In this paper, we focused on matrix characteristics, an approach has been proposed and evaluated to assume the matrix characteristics by generating several metrics. Our approach allows the developers to choose the most suitable storage formats for given sparse matrix and improve the performance of SpMV operation. The aim of our approach is suggest the best suitable storage format for any sparse matrix by analyzing our generated metrics. We have evaluate our approach using 15 real world sparse matrices from University of Florida sparse matrix collection (UFSMC) [30] with four most popular storage formats. Along with some well-known metrics we also generate some new metrics in our approach. In our evaluation we observed that, CSR perform best when the row thickness is very large that means CSR is a better choice for horizontal substructures. ELLPACK perform best with higher diagonal density (49.16%) when the numeric symmetry (2%) is very less.

In future we will extend our research with newer metrics to describe the sparse matrix with metadata more accurately. We will incorporate other common dependency for SpMV operation and increasing both sparse matrix and storage format. We will enhance our research by incorporating machine learning.

Acknowledgment

Experiment for the work presented in this paper was supported by High Performance Computing Center (Aziz Supercomputer) at King Abdulaziz University.

References

[1] G. Goumas, K. Kourtis, N. Anastopoulos, V. Karakasis, and N. Koziris, "Performance evaluation of the sparse matrix-vector multiplication on modern architectures," *J. Supercomput.*, vol. 50, no. 1, pp. 36–77, oct 2009. [Online]. Available: https://doi.org/10.1007/s11227-008-0251-8

[2] K. Kourtis, V. Karakasis, G. Goumas, and N. Koziris, "CSX: An extended compression format for SpMV on shared memory systems," *ACM SIGPLAN Not.*, vol. 46, no. 8, pp. 247–256, 2011.

[3] W. DGROPP, D. KKAUSHIK, D. EKEYES, and G. BFSMITH, "Towards Realistic Performance Bounds for Implicit CFD Codes," *Parallel Comput. Fluid Dyn. 1999*, pp. 241–248, 2000.

[4] K. Kourtis, G. Goumas, and N. Koziris, "Optimizing Sparse Matrix-Vector Multiplication using index and value compression," *Conf. Comput. Front. - Proc. 2008 Conf. Comput. Front. CF'08*, pp. 87–96, 2008.

[5] A. Buluç, J. T. Fineman, M. Frigo, J. R. Gilbert, and C. E. Leiserson, "Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks," *Annu. ACM Symp. Parallelism Algorithms Archit.*, pp. 233–244, 2009.

[6] J. Willcock and A. Lumsdaine, "Accelerating sparse matrix computations via data compression," *Proc. Int. Conf. Supercomput.*, pp. 307–316, 2006.

[7] N. Bell and M. Garland, "Implementing sparse matrix-vector multiplication on throughput-oriented processors," in *Proc. Conf. High Perform. Comput. Networking, Storage Anal. - SC '09*, no. 1. New York, New York, USA: ACM Press, 2009, p. 1. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1654059.1654078

[8] Yousef Saad, "Iterative Methods for Sparse Linear Systems: Second Edition - Yousef Saad - Google Books," 2003.

[9] M. Heller and T. Oberhuber, "Adaptive Row-grouped CSR Format for Storing of Sparse Matrices on GPU," 2012. [Online]. Available: http://arxiv.org/abs/1203.5737

[10] A. Ashari, N. Sedaghati, J. Eisenlohr, and P. Sadayappan, "An efficient two-dimensional blocking strategy for sparse matrix-vector multiplication on GPUs," in *Proc. 28th ACM Int. Conf. Supercomput. - ICS '14*. New York, New York, USA: ACM Press, 2014, pp. 273–282. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2597652.2597678

[11] D. Guo and W. Gropp, "Applications of the streamed storage format for sparse matrix operations," *Int. J. High Perform. Comput. Appl.*, vol. 28, no. 1, pp. 3–12, 2014.

[12] M. Maggioni and T. Berger-Wolf, "AdELL: An adaptive warp-Balancing ELL format for efficient sparse matrix-Vector multiplication on GPUs," *Proc. Int. Conf. Parallel Process.*, pp. 11–20, 2013.

[13] C. Zheng, S. Gu, T. X. Gu, B. Yang, and X. P. Liu, "BiELL: A bisection ELLPACK-based storage format for optimizing SpMV on GPUs," *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2639–2647, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2014.03.002

[14] J. Hartmanis and J. V. Leeuwen, *High Performance Embedded Architectures and Compilers 5th*, J. v. L. Gerhard Goos, Juris Hartmanis, Ed., 2010, vol. 9, no. 3.

[15] K. Li, W. Yang, and K. Li, "Performance analysis and optimization for SpMV on GPU using probabilistic modeling," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 196–205, 2015.

[16] R. Shahnaz, A. Usman, and I. R. Chughtai, "Review of Storage Techniques for Sparse Matrices," in *2005 Pakistan Sect. Multitopic Conf.* IEEE, dec 2005, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/4133468/

[17] Y. Nagasaka, A. Nukada, and S. Matsuoka, "Adaptive Multi-level Blocking Optimization for Sparse Matrix Vector Multiplication on GPU," *Procedia Comput. Sci.*, vol. 80, pp. 131–142, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2016.05.304 https://linkinghub.elsevier.com/retrieve/pii/S187705091630655X

[18] I. P. Stanimirovic and M. B. Tasic, "Performance comparison of storage formats for sparse matrices," *Ser. Math. Informatics*, vol. 24, no. 1, pp. 39–51, 2009. [Online]. Available: http://facta.junis.ni.ac.rs/mai/mai24/fumi-24_39_51.pdf

[19] E. Montagne and A. Ekambaram, "An optimal storage format for sparse matrices," *Inf. Process. Lett.*, vol. 90, no. 2, pp. 87–92, 2004.

[20] D. Buono, F. Petrini, F. Checconi, X. Liu, X. Que, C. Long, and T.-C. Tuan, "Optimizing Sparse Matrix-Vector Multiplication for Large-Scale Data Analytics," *Proc. 2016 Int. Conf. Supercomput. - ICS '16*, pp. 1–12, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2925426.2926278

[21] J. Zhang, J. Wan, F. Li, J. Mao, L. Zhuang, J. Yuan, E. Liu, and Z. Yu, "Efficient sparse matrix-vector multiplication using cache oblivious extension quadtree storage format," *Futur. Gener. Comput. Syst.*, vol. 54, pp. 490–500, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.future.2015.03.005

[22] X. Feng, H. Jin, R. Zheng, K. Hu, J. Zeng, and Z. Shao, "Optimization of sparse matrix-vector multiplication with variant CSR on GPUs," *Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS*, pp. 165–172, 2011.

[23] D. Hutchison and J. C. Mitchell, *Lecture Notes in Computer Science*, 1973, vol. 9, no. 3. [Online]. Available: http://www.mendeley.com/research/lecture-notes-computer-science-2/

[24] X. Yang, S. Parthasarathy, P. Sadayappan, H. Yoshizawa, D. Takahashi, E. Montagne, A. Ekambaram, M. E. Epstein, I. Rodan, G. Griffenhagen, J. Kadrlik, M. C. Petty, S. A. Robertson, and W. Simpson, "Fast sparse matrix-vector multiplication on GPUs," *Proc. VLDB Endow.*, vol. 90, no. 2, pp. 130–136, mar 2012. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1098612X15572062 http://dl.acm.org/citation.cfm?doid=1938545.1938548

[25] F. Vázquez, J. J. Fernández, and E. M. Garzón, "Automatic tuning of the sparse matrix vector product on GPUs based on the ELLR-T approach," *Parallel Comput.*, vol. 38, no. 8, pp. 408–420, 2012.

[26] M. M. Baskaran, R. Bordawekar, M. Manikandan, and B. Rajesh Bordawekar, "Optimizing Sparse Matrix-Vector Multiplication on GPUs Using Compile-time and Run-time Strategies," Tech. Rep., 2008. [Online]. Available: http://domino.watson.ibm.com/library/CyberDig.nsf/home.

[27] S. Chen, J. Fang, D. Chen, C. Xu, and Z. Wang, "Adaptive Optimization of Sparse Matrix-Vector Multiplication on Emerging Many-Core Architectures," *2018 IEEE 20th Int. Conf. High Perform. Comput. Commun. IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Sci. Syst.*, pp. 649–658, 2018.

[28] W. Liu and B. Vinter, "CSR5 : An Efficient Storage Format for Cross-Platform Sparse Matrix-Vector Multiplication Categories and Subject Descriptors," pp. 339–350.

[29] H. Yoshizawa and D. Takahashi, "Automatic Tuning of Sparse Matrix-Vector Multiplication for CRS Format on GPUs," in *2012 IEEE 15th Int. Conf. Comput. Sci. Eng.* IEEE, dec 2012, pp. 130–136. [Online]. Available: http://ieeexplore.ieee.org/document/6417285/

[30] T. A. Davis and Y. Hu, "The university of Florida sparse matrix collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 1–25, 2011.

Survey on Domain Specific Languages Implementation Aspects

Eman Negm[1], Soha Makady[2], Akram Salah[3]
Faculty of Computers and Artificial Intelligence
Cairo University, Giza, Egypt

***Abstract*—Domain Specific Languages (DSLs) bridge the gap between the business model and the technical model. DSLs allow the technical developer to write programs with the business domain notations. This leads to higher productivity and better quality than General Purpose Languages (GPLs). One of the main challenges of utilizing DSLs in the current software process is how to reduce the implementation cost and the knowledge required for building and maintaining DSLs. Language workbenches are environments that provide high level tools for implementing different language aspects. The purpose of this paper is to provide a survey on the different aspects of implementing DSLs. The survey includes structure, editor, semantics, and composability language aspects. Furthermore, it overviews the approaches used for each aspect and classify the current workbenches according to these approaches.**

Keywords*—Domain Specific Language (DSL); language workbench; language implementation aspects; software language engineering

I. Introduction

Domain Specific Languages (DSLs) [1] are languages that are designed and implemented to express and solve a specific class of problems. This class represents the domain of the language. SQL (Structured Query Language) is a DSL specialized in database domain, HTML (Hyper Text Markup Language) is another DSL that concerns with the web domain. Throughout the paper, *MiniIoT* DSL is used to describe the different aspects of DSL implementation, it is a fake DSL for Internet of Things (IoT) domain. Listing 1 shows a sample of *MiniIoT* code that defines sensors and actuators to send alarm and open the extinguisher in case of the temperature exceeds some threshold in a specific building.

DSLs have many advantages over General Purpose Languages (GPLs) for representing a specific domain. On one hand, it provides higher abstractions for the given domain which raises the productivity and the quality of the development process [2]–[4]. The user, who utilizes *MiniIoT*, will use concepts like devices, sensors, and actuators to write IoT programs. In case of utilizing one of the GPL like Java and C, the user is forced to use concepts that are not related to his domain like classes, fields, and arrays to represent his problem. The latter representation consumes more time and effort and may generate more errors than the direct *MiniIoT* representation. DSL also provides better validation and verification for the output programs since it utilizes domain specific constraints. The DSL developer could define constraints to verify that the generated program is meaningful. For example, *MiniIoT* will not allow to configure the same sensor in two different locations in the same time. In addition, the error messages are also more meaningful since it utilizes the domain concepts. *MiniIoT* will display error messages like "*This sensor is already located in building No. 11 in this time*". On the other hand, the use of domain notations in the DSLs allows more involvement for the domain expert in the development process. This leads to bridge the gap between the business model which is owned by the domain expert and the technical model which is owned by the programmer. As a result, The quality of the final product is enhanced [5], [6].

Listing 1: MiniIoT Code Sample

```
Define Sensor temperatureSensor measures
    Temperature located in Building 11
Define Actuator alarmActuator that send Alarm
    to number 911 located in Building 11
Define Actuator extinguisherActuator that open
     fire extinguisher located in Building 11
if ( temperatureSensor.temperature >
    TEMPERATURE_THURSHHOLD ) {
 alarmActuator.execute
 extinguisherActuator.execute
}
```

Despite the advantages mentioned above, DSL doesn't have the expected role in the current software development life cycle. There are many reasons for this, one of the main reasons is the cost of designing, implementing, and composition of a new DSL. Designing a DSL, that covers all domain aspects, is not an easy task. Implementing a new language from scratch, in case of no suitable one is available, is a very difficult task if the normal GPL techniques are used. Building a compiler or an interpreter from scratch consumes a lot of time and effort, plus special technical skills are required. Most of the current applications involve more than one domain, consequently the modularity and composability of the DSL are very important aspects when creating a new DSL or even choose an existing one.

Language workbenches, the term is proposed by Martin Fowler [7], are comprehensive environments that try to provide a solution for the above problems by providing high level tools. These tools facilitate the development, maintaining, and composition of DSLs. Most of language workbenches apply the idea of modular language and language extension. They enable the language developer to create languages' modules that are fully integrated with each other.

Although the DSL concept is not a new one, there are no standard approaches for DSLs development. There are many aspects that should be handled by the language workbenches to

TABLE I. DSL WORKBENCHES

Workbench Name	URL
Ensō	http://www.enso-lang.org/
Intentional Software	http://www.intentsoft.com/
LISA	—
Mas	http://www.Mas-wb.com/
MetaEdit+	http://www.metacase.com/
MontiCore	http://www.monticore.de/
MPS	http://www.jetbrains.com/mps/
Neverlang	http://neverlang.di.unimi.it/
Rascal	http://www.rascal-mpl.org/
Silver	http://melt.cs.umn.edu/silver/
Spoofax	http://www.spoofax.org/
SugarJ	http://www.sugarj.org/
Whole Platform	http://whole.sourceforge.net/
Xtext	http://www.eclipse.org/Xtext/

achieve their goals. Each workbench follows its own approach to implement these aspects. Some of these approaches are published in scientific papers and others are in the workbench documentation. Up to our knowledge, no recent study done to discuss and classify these approaches.

This paper discusses the different approaches for achieving the various DSL implementation aspects. It includes structure, editor, semantics, and composability language aspects. There are other aspects that are not included in this survey like language validation and testing aspects. The language workbenches are classified based on the approaches used in each aspect. The survey includes 14 language workbenches (Table I). Unfortunately, there are some workbenches that the authors couldn't fetch their approaches in some aspects due to poor documentation or due to their being commercial workbenches.

The rest of the paper is structured as follows, Section II describes some concepts of the language implementation and workbenches. Section III looks at related surveys on DSL implementation. Sections IV, V, VI, and VII introduce the different aspects of the DSL implementation. They illustrate each aspect and the main approaches, that are used in the current workbenches, to achieve these aspect. Section IX concludes the paper and presents prospective ongoing research directions.

II. BACKGROUND

Programing languages are a set of languages used to give instructions to the computers to preform a specific task. There are two types of programming languages: General Purpose Languages (GPLs) and Domain Specific Languages (DSLs). GPL is a language that is used to write programs for any domain. It does not contain any constructs that are related to a specific domain. Java, C++, and Python are examples of GPLs. DSL is a language that concerns with a specific class of problems that represents a specific domain. It contains abstractions and optimizations for the given domain. A specialized editor is often provided with the DSL to provide a specialized editor services and error messages. SQL, HTML, and CSS are examples of DSLs.

DSL could be implemented as internal or external DSL. Internal DSL is the DSL that is embedded into a general purpose language. This type is limited with the grammar of the host language. Additionally, there is a lack in the editor support since the editor is not aware of the grammar and constraints of the embedded DSL. In contrast, external DSL are implemented independently from any host language. It has its own grammar and editors which makes the language more flexible and the editor could assist the programmer during the development process with the domain knowledge. Throughout the paper, The term DSL is used to refer to the external DSL.

DSL development process involves five stages: 1) Domain Analysis: this stage includes understanding and analyzing the concerned domain, getting the concepts and relations of the domain, and determining the boundaries of the domain, 2) Design: this stage defines the design of the DSL that could encode the problems of the domain determined in the previous phase, 3) Implementation: the concrete language is implemented in this stage using one of the existing tools, 4) Evaluation: where the new DSL is evaluated to determine if it satisfies the business's need or not, and 5) Maintenance: it is a continues stage to update the new developed DSL to satisfy the continues changes in the business requirements. This survey concerns with the approaches used in the implementation stage only.

There are two general approaches for programing languages implementation: compilation and interpretation. In the compilation approach, the compiler translates the programs, written in a high level language, into a low level language that will be executed by the machine. In the interpretation approach, the interpreter executes the actions written in the given program directly on the machine without any translation. Building a compiler or an interpreter from scratch is a complex task that needs a lot of effort and time. It also needs special technical skills. DSLs are lightweight languages than GPLs since they cover limited domain. It should be implemented faster with less effort and knowledge. Therefore, using the classical GPL implementation approaches for DSL implementation is not realistic.

Language Workbench is an environment for DSLs development. It provides high level tools for implementing, evaluating, and maintaining DSLs. Language Workbench reduces the effort and knowledge needed for building DSLs and hides the complexity of the GPL implementation approaches. Fowler [7] lists three components that should be supported by any language workbench to create a new DSL: 1) The abstract representation that includes defining the language structure, 2) The editor that allows the user to manipulate the abstract representation, and 3) The generator that transforms the abstract representation into an executable coded. The current approaches used to support the above three components are described in Sections IV, V, and VI. In addition, the language workbench should support the integration among DSLs which is described in Sections VII.

Throughout the paper, the language developer is the person who is responsible for developing the DSL, and the user is the person who utilizes the DSL to write programs. Fig. 1 shows the relation among language developer, language user, and language workbench.

III. RELATED WORK

Although, there are many research directions in the domain specific languages field, there is a limited number of survey papers that cover the aspects of implementing DSLs. Van

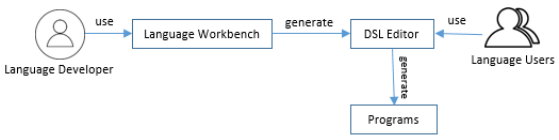


Fig. 1. DSL Implementation Stockholders

Deursen et al. [8] lists the implementation approaches for the semantic aspect only. Kosar et al. [9] provides a preliminary study on ten DSL implementation approaches. A unified DSL is designed to be implemented by the ten approaches to compare the DSL implementation effort and the end-user effort to build programs with the given DSL. Similar work is done by [10] to evaluate four DSL implementation approaches based on unified state machine DSL. It concludes that no single approach is valid for all scenarios. Oliveira et al. [11] provide a theoretical survey on DSLs that does not describe any implementation techniques. Erdweg et al. [12] focus on the language composability aspect and how it is covered by different language workbenches.

A systematic mapping study is provided by [13] that includes research questions related to the techniques and methods of DSLs, the existing DSLs with their related domains, and the tools for DSL creation and usage. The study lists some techniques used in DSLs and it provides statistics about the DSL research types and their respective domains. The results show that defining external DSLs for the different domains gains a lot of attention. This indicates that utilizing of DSLs is very useful and applicable in many domains.

Language Workbench Challenge 2013 (LWC'13) published a work where the authors propose a feature model for languages workbenches and classify the workbenches according this model [14], [15]. It provides a unified challenge (i.e. a DSL for questionnaires) to be implemented by ten workbenches that are included in LWC'13. The paper compares the features provided by the different workbenches in the different aspects. It shows that no one language workbench supports all the required features. The authors is concerned with the supported features of each workbench more than the approaches used to implement these features.

Kosar et al. [16] and Mernik [17] performed another systematic mapping study on the research papers published from 2005 to 2013. The authors include the papers published after their survey paper in 2005 [18].The systematic mapping study includes research questions that try to catch the research space and trends of the DSL field within the given period. One of the main conclusions of this study is that the DSLs will be the main programming languages for the next period. Additionally, it reports that one of the open problems in DSL field is how to facilitate the DSL development for domain expert. This survey discusses the current approaches that contribute in solving the above problem. Thanhofer-Pilisch et al. [19] preformed another systematic mapping study that focuses on DSL evolution only.

As per the previous paragraphs, no recent survey papers cover the latest approaches of implementing DSLs. Unlike the above work, this survey describes the recent approaches used in four different aspects. It focuses on the internal approaches used in the current workbenches rather than the provided features of each workbench. The provided survey could be a starting point for the DSL researchers to catch the current status of DSL implementation aspects.

IV. Language Structure Aspect

The language structure aspect includes defining the structure of the new DSL. This structure represents the concepts and the relations that reflect the target domain. In this aspect, the language developer defines the concepts that are included in the *MiniIoT* language. Fig. 2 shows a subset of the concepts of *MiniIoT* DSL and their relations. *MiniIoT* contains two types of devices: sensors and actuators. The device is located in some place. The Sensor generates a specific observation and the actuator does some action.

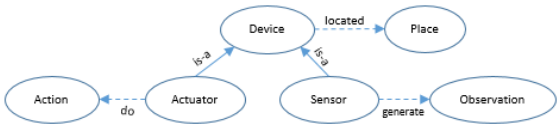


Fig. 2. *MiniIoT* Concepts

There are two main approaches, used in the current language workbenches, for defining the language structure: Grammar-driven approach and Model-driven approach. The subsequent sections describe each approach in details. Table II classifies the included workbenches according to the above mentioned approaches.

TABLE II. DSL Workbenches classification according to language structure aspect

Approach	Workbench	Used Techniques
Grammar-driven Approach	*Spoofax*	SDF
	SugarJ	SDF
	Silver	Attribute Grammar
	LISA	Attribute Grammar
	Neverlang	Grammar DSL
	MontiCore	Grammar DSL
	Rascal	Grammar DSL
	Xtext	Grammar DSL
	Ensō	Object Grammar
Model-driven Approach	*Whole Platform*	Meta-model DSL
	MPS	Meta-model DSL
	Mas	Meta-model DSL
	MetaEdit+	GOPPRR Model
	Intentional Software	Meta-model DSL

A. Grammar-Driven Approach

In the grammar-driven approach, the definition of the language depends on defining the grammar of this language. Context Free Grammar(CFG) is used as the formal definition for the concrete syntax of the language. The meta languages, that are used to define CFG, are reused to define the DSLs with some modifications to be more appropriate for DSL definition. Backus–Naur Form (BNF) is a one of the main formal meta-languages that are used to define CFG. Extended Backus–Naur Form (EBNF) is a meta-language that extends BNF by adding more notations and symbols like multiplicity symbol.

Attribute grammar is a context free grammar with attributes and evaluation rules. It can contain syntax and semantics

information while (E)BNF contains syntax only. The values of the attributes are evaluated by a compiler or a parser according to the given evaluation rules. Attribute grammar defines two types of attributes: synthesized attributes and inherited attributes. The values of the synthesized attributes are computed from the values of the attributes of the children nodes. They pass semantic information down in the abstract syntax tree. However, the values of the inherited attributes are computed using the values of the attributes of the parent nodes. They pass semantic information up in the abstract syntax tree. *Silver* [20] and *LISA* [21] are examples of systems that utilize the attribute grammar to define the structure of the DSL.

Syntax Definition Formalism [22] (SDF) is another meta-language for defining the grammar of a language. SDF is richer than (E)BNF and more modular. It allows to divide the grammar into modules which facilitates the language's embedding and reusing. In addition, SDF provides more declarative grammar definition and declarative disambiguation. Moreover, It enables to define both lexical and context-free syntax. For the above reasons, SDF is more suitable to be utilized for DSL definition. SDF implementation provides a scannerless parser for the SDF definition. *Spoofax* [23] and *SugarJ* [24] are examples of workbenches that use SDF for defining language structure.

Another technique for defining the grammar of the language is to develop a new DSL for the grammar definition. The main advantage of this technique is that the new developed DSL is designed to be easier and more user friendly than the above standards. In addition, most of the proposed languages support modularity and reusability. *Xtext* [25] workbench proposed a grammar DSL which is an EBNF-like DSL developed by *Xtext* itself. This grammar DSL defines the concrete syntax and how it is mapped to the semantic model of the new DSL. *Rascal* [26] workbench has a meta-programming DSL that includes notations for defining the syntax of the new DSLs.

Neverlang [27], [28] also proposes a DSL that represents the syntax by a set of productions coded in BNF. *MontiCore* [29], [30] proposes Grammar Definition Language for defining both concrete and abstract syntax. The benefit of using the same language for both elements is to reduce the inconsistency and redundancy between the two elements. One of the drawbacks of this technique is that it does not support languages that have multiple concrete syntax and single abstract syntax.

Object grammar [31], [32] is another technique proposed in *Ensō* workbench. It provides a declarative mapping between text and object graph. Object grammar extends EBNF by adding constructs to build an object graph while parsing the text. It also defines how to transform the generated graph back into text. Object grammar produces a graph as an output for the parsing process instead of the parse tree in the above techniques. The graph is described by a schema. One of the main advantages of Object Grammar that it supports language reusabilty easier than traditional grammar-based techniques.

B. Model-Driven Approach

The Model-driven approach does not use grammar rules to define the language structure. In contrast, it utilizes the meta-modeling for defining the structure of the language. The program, written by the user, is the model that conforms to the meta-model defined by the language developer. A Language structure meta-model is used to describe the abstract syntax not the concrete syntax. It defines the concepts, relations, and constraints of the language. This meta-model does not describe how the user will edit these concepts and relations. The concrete syntax is determined by a projection process that describes how to edit the concepts and relations defined in the previous meta-model. The projection process is described in more details in the Language Editor Aspect (Section V-B). One of the main advantages of this technique is that it provides higher abstraction level than the grammar-driven techniques. The language developer doesn't have to know the grammar details or the parser techniques. He defines a semantic model rather that a concrete syntax which makes the definition process more easier.

Most of the language workbenches, that follow this approach, propose a DSL to define of the language structure meta-model. *MPS* [33], *Whole Platform* [34], *Mas*, and *Intentional Software* [35] workbenches follow this approach. *MPS*'s DSL is called *Structure Language* which is a DSL for representing and editing the Abstract Syntax Tree (AST) of the program. It defines the concepts, attributes and relations of the nodes of the AST. *Whole Platform* has a modeling framework that utilizes a DSL called *Models Language*. *Models Language* defines the language as a set of entities, features, and types. *Mas* also has its own DSL to define the structure of the language. It a web-based language that allows the users to define the abstract syntax of the DSL.

Intentional Software Workbench [36] proposed a tree representation for the software programs that merges the abstract syntax and the concrete syntax. It represents any program as a so-called intentional program tree which is a kind of syntax tree. This tree consists of a set of nodes and references; the nodes represent the program elements and the references represent relations among elements. The intentional program tree includes different types of nodes that hold semantic and syntactic information. This mix between abstract and concrete syntax facilitates the textual representation of the semantic model as described in Section V-B.

MetaEdit+ [37] is a language workbench that supports only graphical DSLs. It proposes a model called GOPPRR Model to define the abstract syntax of the graphical DSL. GOPPRR stands for Graph, Object, Property, Port, Role, and Relation which are the elements of the language structure in *MetaEdit+*. The user creates the GOPPRR model by filling in a set of forms or by specifying the model graphically.

V. Language Editor Aspect

In language editor aspect, the language developer defines how the user views and edits the language structure defined in the above aspect. In this aspect, the developer of the *MiniIoT* language should define the editor of the sensors and the actuators concepts. The editor should allow the user to define new sensors and actuators and to define the relations among them. In addition, It should support all the language editor services like highlighting, error checking, and code completion. Fig. 3 shows a sample textual editor for *MiniIoT* script that defines sensors and actuators. This editor is created by *MPS* workbench.

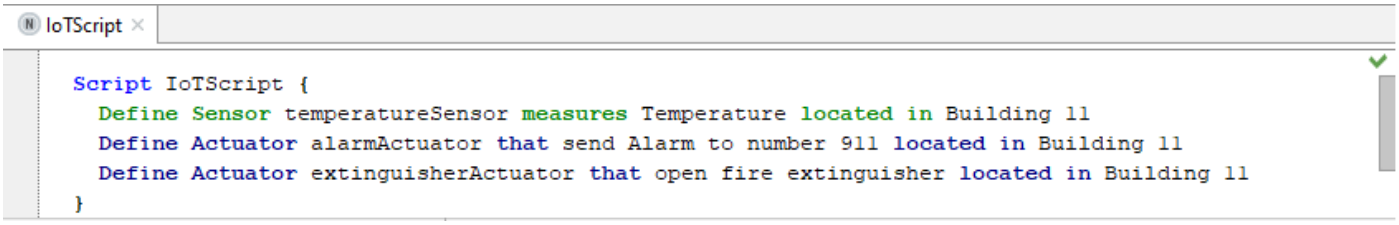


Fig. 3. *MiniIoT* Editor

TABLE III. DSL WORKBENCHES CLASSIFICATION ACCORDING TO LANGUAGE EDITOR ASPECT

Editing Mode	Workbench	Supported Language Notations			
		Textual	Tabular	Symbols	Graphical
Parser-based	*Spoofax*	✓			
	Neverlang	✓			
	MontiCore	✓			
	Silver	✓			
	Rascal	✓			
	Ensō	✓			✓
	SugarJ	✓			
	Xtext	✓			
	LISA	✓			
Projectional	*Whole Platform*	✓	✓	✓	✓
	MPS	✓	✓	✓	
	Mas	✓	✓		
	MetaEdit+		✓	✓	✓

The user should be able to edit either the grammar-driven or the model-driven definition. The usability of the language editor is one of the key factors for the success of any DSL. There are two editing approaches implemented in the language workbenches: the parser-based editing and the projectional editing (i.e. parser-less editing). Table III classifies the workbenches according to the above editing approaches.

A. Parser-based Editing

Parser-based editing depends on a parser that translates a textual code into an AST. Users interact with the concrete syntax by entering a sequence of characters into a text buffer. The parser then matches this sequence with the language grammar to construct the abstract syntax tree (AST) of the program. This editing approach is the main approach that are applied in the grammar-driven workbenches. The parser is automatically generated after the grammar of the DSL is defined by the language developer. The generated parser is utilized to build the final editor of the DSL. Most of the parser-based workbenches generate the editor as an Eclipse plugin based on the generated parser. Consequently, The parser-based editors support textual DSLs only. *Ensō* is an exceptional case, it is a grammar-driven workbench that supports graphical notations since it depends on the object grammar that defines the bidirectional mapping between the text and the object graph (Section IV-A).

Spoofax and *SugarJ* use SDF for the grammar definition, SDF generates scannerless generalized LR parser [38]. *Spoofax* also has a configuration DSL called *Editor SerVice (ESV)* to configure the editor services. *Silver* has a parser and context-aware scanner generator called *Copper*. *LISA* tool generates a source code of the scanner, parser, interpreter, and compiler in Java. *Silver* and *LISA* do not support editor generation. *Xtext* and *MontiCore* workbenches also generate a parser based on the given grammar. In addition, they generate an Eclipse plugin as a textual editor for the language. *Neverlang* has a modular LR parser generator called *DEXTER* but it does not support editor generation.

B. Projectional Editing

Projectional editing is an editing approach that depends on the abstract representation as the main source of information. It is also called parser-less editing since no parsing is needed to build the AST. Unlike parser-based editing, projectional editing allows the user to directly build and edit the AST. No need for transformation from the concrete syntax to the abstract syntax. Instead, the projection process generates a visual representation from the AST. This representation could be textual, graphical, or tabular. The workbenches that apply this approach are called projectional workbenches, they follow the model-driven approach for the language structure definition. One of the main advantages of the projectional editing is the ability to represent the same abstract representation with different visual representations according to the user's need.

As in the language structure aspect, new meta-languages are defined for projecting the abstract representation to the end user and defining the editor services like coloring and error marking. *MPS*, *Whole Platform*, *Intentional Software*, and *Mas* are examples for projectional language workbenches. *MPS* has a DSL for creating the projectional editors called *Editor Language*. It defines the editor as a set of cells [39], [40], each cell can contain a static symbol or a user defined symbol. These cells allow the language developer to define different types of notations for the DSL (i.e. textual, tabular, symbols, or graphical) since they reflect the AST directly without parsing.

Whole Platform has a Model-Based Editing (Mbed) framework that provides the *Editors DSL*. It is a DSL that allows the user to create editors based on the language structure defined by *Models DSL*. Mbed is implemented based on the Eclipse Graphical Editing Framework[1] (GEF). Whole Plateform editors are generated as an Eclipse plugin.

Intentional Software uses the intentional program tree described in (Section IV-B) to facilitate the textual editing. This is done by continuous loop of unparsing, editing, and parsing steps. The unparsing step transforms the tree to a sequence of tokens, each token is a sequence of characters or

[1]http://www.eclipse.org/gef

an extended formate (e.x. image). The token also holds other information like formating information and author information. The generated token sequence is the textual representation of the given tree (i.e. program). The user can edit the token sequence by adding new tokens or deleting existing ones. Then, the updated sequence is parsed to generate an updated tree. The final step is to unparse the updated part of the tree to derive tokens that will be displayed through the editor. The main goal of this method is to keep the advantage of the model-driven approach while simulating the normal textual editors behavior.

C. Language Notations

This section discusses the types of notations that could be supported by the DSL editors. A DSL could include textual, tabular, symbols, and graphical notations. It could be restricted to only one type or it can mix different types of notations. The types of notations supported by the editor is totally dependent on the used editing approach. Table III lists the notations types supported by different workbenches.

Parser-based workbenches (*Spoofax*, *Neverlang*, *MontiCore*, *LISA*, *Silver*, *SugarJ*, and *Xtext*) support textual notations only since they are based on parser-based editing. However, *Ensō* supports graphical notations since it depends on the object grammar that include bidirectional mapping between the text and the object graph. *MetaEdit+* is the only workbench that does not support textual notations, it supports graphical, symbol, and tabular notations only. Projectional workbenches simplify the integration among different types of notations since they are parser-less. Consequently, *MPS*, *Mas*, and *Whole Platform* support mixing among different notations in the same DSL.

D. Projectional Editing vs. Parser-based Editing

Table IV lists a comparison between the projectional editing and the normal parser-based editing. Language modularity, notational freedom, and program representation flexibility could be achieved easily by projectional editing while it more difficult to be implemented using parser-based editing. Projectional editing is still not compatible with many systems in the current infrastructure like source control systems. One of main advantages of the normal parser-based editing over the projectional one is avoiding the tool lock-in, the user is not limited to specific editors with specific versions. The user can open the code and edit it with any textual editor. Additionally, the code is always open even it contains incomplete or wrong syntax. Finally, the usability of the textual parser-based editors overcomes the usability of the projectional ones. The main reason for this situation is that the user works directly with the program tree in the projectional editors rather than the character sequence in the parser-based editors. Consequently, a special handling should be done for deletion, insertion, and copy/past actions done by the user which is not the case for the parser-based editors.

TABLE IV. PARSER-BASED EDITING VS PROJECTIONAL EDITING

Criteria	Parser-based Editing	Projectional Editing
Editor Usability	Normal text editors	Specialized editors
Language Modularity	Done by merging Grammars (Ambiguity)	Done by merging ASTs
Notational Freedom	Support textual notations only	Support non-textual notations
Language Evolution	Code is always opened	Editors should take special care for incompatible models
Infrastructure Integration	Can be easily integrated with current tools	Not compatible with many tools and copy/past from projectional editor to/from text editor is still an issue.
Tool Lock-in	Any text editor can work	Users are limited to specific editors with specific versions.
Program Representation	Only textual representation	Providing several projections

VI. LANGUAGE SEMANTICS ASPECT

With the structure and the editor aspects, the language developer defines the structure of the DSL and how to edit this structure. In the language semantics aspect, the language developer defines how the program, written in the given DSL, will be executed. Language semantics describes the meaning of the language notations. It describes the steps, the computer should follow, to execute the given program. There are two ways for defining the semantics of a DSL: translational semantics and interpretive semantics.

A. Translational Semantics

Translational semantics defines the meaning of the language by translating this language into another target language. The compiler is an example for translational semantics that describes the semantics of a high level language by translating it into a low level language. For DSLs, the target language is often one of the general purpose languages, and the AST represents the model of the language. Listings 2, 3, 4, and 5 show a sample code that describes the meaning of the *MiniIoT* code shown in Fig. 3 using Java language.

Listing 2: *MiniIoT* Java code generated - Device.java

```
public class Device {
 protected String place;
 public String getPlace() {
  return place;
 }
 public void setPlace(String place) {
  this.place = place;
 }
}
```

Listing 3: *MiniIoT* Java code generated - Actuator.java

```
public class Actuator extends Device {
 private String action;
 public Actuator(String place, String action)
     {
```

```
  super();
  this.place = place;
  this.action = action;
 }
 public String getAction() {
  return action;
 }
 public void setAction(String action) {
  this.action = action;
 }
}
```

Listing 4: *MiniIoT* Java code generated - Sensor.java

```
public class Sensor extends Device {
 private String observation;
 public Sensor( String place, String
     observation) {
  super();
  this.place = place;
  this.observation = observation;
 }
 public String getObservation() {
  return observation;
 }
 public void setObservation(String observation
     ) {
  this.observation = observation;
 }
}
```

Listing 5: *MiniIoT* Java code generated - IoTScript.java

```
public class IoTScript {
 public static void main(String[] args) {
  Sensor temperatureSensor = new Sensor("
      Temperature", "Building 11");
  Actuator alarmActuator = new Actuator("send
      Alarm to number 911", "Building 11");
  Actuator extinguisherActuator = new Actuator
      ("open fire extinguisher", "Building 11")
      ;
 }
}
```

The translation could be implemented by two ways: model transformation and code generation [41]. The model transformation (i.e. model-to-model) translates the model of the source language into the model of target language independently of the concrete syntax of both languages. The classical approach for implementing model transformation is to construct the target AST while traversing the source AST. *Xtext* and *MPS* support this approach. Another approach is to build a relation between the source AST and the target AST. Then, utilizing this relation to implement the translation between them [42]. The advantage of this approach is to support bidirectional mapping between the source and the target ASTs.

The code generation (i.e. Model-to-Text) translates the model of the source language into the source code of target language directly. Template languages are used for implementing code generation where the source code of the target language is embedded as a text within the template language. Consequently, the tool used for editing template language is not aware of the target language. *Xtext* supports code generation by template languages. By composing the template language and the target language, the tool becomes aware of both languages which facilitates the translation definition. *MPS* and *Spoofax* utilize language composition and template languages to support code generation.

Xtext uses *Xtend*[2] language for code generation and model transformation. *Xtend* is a general purpose language very similar to Java syntax but with less linguistics redundancy. *Xtext* supports code generation and model transformation by generating the target text or the target AST while traversing the source AST.

MPS also supports code generation and model transformation. However, the code generation is only used at the end of the chain where the AST is translated into a GPL code to be passed into the compiler. This translation is done by *textgen* language. *MPS* supports model transformation by templates and macros. The template code is an instance of the target model that defines the actual transformation. These templates contain a set of macros that are used to define the dependences between the target model and the source model. Macros also define queries over the source model that are utilized to build the target one.

Whole Platform supports code generation by providing a Java model generation framework that translates the model to Java compilation units. The framework provides a set of APIs, that is based on Eclipse platform Java tools[3], to allow the language developer to define the translation from the model to the Java compilation units. *Whole Platform* supports model transformation by a traversal framework that facilitates traversing the source model to implement the translation process to the target model.

Spoofax has a DSL for program transformation called *Stratego* [43]. It utilizes rewrite rules and rewriting strategies to support code generation and model transformation. The rewriting strategies contain a set of rewrite rules in a specific order and conditions. If the right-hand sides of the rewrite rules are the final text of the target language, then it will preform a code generation. However, If the right-hand sides are parts of the model of the target language, then it will preform a model transformation. *SugarJ* also uses Stratego for defining the language's translational semantics.

B. Interpretive Semantics

Unlike translative semantics, interpretive semantics defines the meaning of the programs by executing them directly without translation to another language. The language developer defines how the different language constructs will be evaluated (i.e. executed). The semantic actions, defined by the language developer, are executed while traversing the AST. The developer of *MiniIoT* in this case will build an interpreter that generates an executable code directly, no intermediate Java code will be generated.

[2]https://www.eclipse.org/xtend/

[3]http://www.eclipse.org/jdt/ index.html

Xtend is utilized by *Xtext* to allow the language developer to build interpreters for DSLs. *MPS* utilizes the BaseLanguage, a Java similar language, to define interpreters. Baselanguage is defined by *MPS* itself, so it could be extended to support more features for building interpreters. The developer uses Xtend in *Xtext* or BaseLanguage in *MPS* to define how the DSL statements will be executed and how the expressions will be evaluated. In addition, he defines the changes in the program's variables and states.

Whole Platform supports interpretive semantic through its traversal framework, the language developer utilizes the APIs provided by this framework to traverse the model and define the corresponding execution actions. *Spoofax* and *SugarJ* follow a different way to build interpreters. It utilizes the rewrite rules to define the current state of the program and to define the transformation among different states to reach the final execution state.

VII. Language Composability Aspect

DSL is a language that is specialized in a specific domain, while the real life programs may contain more than one domain. Merging more than one DSL in the same program is called language composition. Language composability aspect defines how the new DSL will be composed with other DSLs. Assuming that the developer of *MiniIoT* has a new requirement for extending *MiniIoT* to support robots functionalities. *MiniRobot* is a ready DSL for robotic application development. Accordingly, the developer will decide to compose *MiniRobot* DSL with *MiniIoT* DSL to satisfy the new requirements.

Erdweg et al. [12] proposed a theoretical classification for the different types of language composition. The authors specify four types of language composition: Language Extension, Language Unification, Self-Extension, and Extension composition. These types describe the relations among the languages that will be composed. The composition process includes combining the different language aspects: Language Structure, Language Editor, Language Semantics, and Language Validation.

Parser-based workbenches support composability by merging the grammars of the combined languages. *Xtext* supports only language extension. It defines the grammar in a modular way by the grammar DSL, where the module could extend another module. The grammar in *Xtext* only inherits from one base grammar, consequently it could not support language embedding, extension composition, and language unification. The limitation of *Xtext* composability is due to utilizing ANTLR's LL(*) algorithm [44]. Alternatively, workbenches that depend on generalized parsing techniques support different types of grammar composition.

Spoofax and *SugarJ* uses SDF and Stratego for composing different language aspects. SDF applies scannerless generalized LR parsing, which enables language unification. *Rascal* is another example for workbenches that utilizes generalized parser. *Spoofax*, *SugarJ*, and *Rascal* support language semantics and validation composition by combining the rewrite rules.

MontiCore supports composability by grammar inheritance and language embedding. The grammar inheritance enables language extension without changing the base language. The extension adds new productions or overrides existing ones. The language embedding between two grammars is done by defining external nonterminals that should be filled by the embedded language.

Neverlang depends on feature orientation and modular language development to implement language composability. It proposes a feature oriented composition model that defines the language's implementation as a set of components. The model defines the relations among the components as dependencies. The dependency is a property that is required by the component but it is not defined in the same component. The dependency is represented by a placeholder in the syntax level or by a semantic property in the semantic level.

Ensō achieves composability by implementing a merge operator between the two object grammars that define the combined languages. The merge operator is a union operator that merges the object graphs of the two languages and overrides the duplicate objects and attributes by the values of the second language. This allows the second language to extend and modify the first one.

Projectional workbenches support composability aspect easier than parser-based ones, since the composition is based on the integration between the abstract representations rather than the grammar rules. *MPS* supports the idea of modular language [45]. Modular language is a bridge between large languages and small languages. It depends on a small core language and a set of language modules (extensions) come with its own syntax, editor, and IDE tooling. The composition in *MPS* is very similar to object oriented programming.

VIII. Discussion

The survey indicates that the approaches of developing DSL could be classified into two main classes: Model-based approaches class and Text-based approaches class. The model-based approaches represent the language as a meta model and the program as an instance model of the given meta model. The meta model is the abstract representation of the language. The definition of the meta model represents the language structure aspect. The definition for the process of editing the meta model represents the language editor aspect. The composition aspect is achieved by merging the meta models of the given languages.

On the other hand, the text-based approaches represent the language as a set of grammar rules and the program as a text that follows the given rules. The definition of the grammar rules represents the language structure aspect. The definition of the parser that translates the text into an abstract syntax tree represents the language editor aspect. The composition aspect is achieved by merging the given grammar rules.

The language semantic aspect defines the translation or the interpretation of the abstract representation to other representation or executable code. Since the language semantic aspect depends on the abstract representation, the same techniques are used in model-based approaches and text-based approaches to implement this aspect.

The text-based approaches are similar to the approaches used for developing general purpose languages, accordingly they are more mature than the model-based approach. In

addition, they are applied by more language workbenches than model-based approaches. Alternatively, the model-based approaches are considered as a new direction for implementing programming languages. They facilitate the implementation of many language aspects and they provide editing capabilities more than text-based approach.

IX. Conclusion

This paper introduces a survey of the different aspects of implementing a new DSL. The survey covers structure, editor, semantics, and composability language aspects. It lists the approaches used for achieving each aspect and describes how different workbenches apply this aspect.

The survey concludes that there are no standards for applying the different DSL implementation aspects. More research is needed to set standards for DSL implementation. Additionally, the key aspect of the DSL implementation is the structure aspect. The technique used in the structure aspect determines the techniques used in the other aspects. Finally, no one existing approach facilitates all DSL implementation aspects.

Future work should consider other aspects like validation and testing aspects. Further research is needed to cover the missing workbenches in each aspect. The survey shows that the projectional editing approach is a very promising approach that could facilitate the implementation of many aspects. Accordingly, further studies will be done to address the current challenges of applying the projectional editing approach.

References

[1] R. Lämmel, "A story of a domain-specific language," in *Software Languages*. Springer, 2018, pp. 51–86.

[2] R. B. Kieburtz, L. McKinney, J. M. Bell, J. Hook, A. Kotov, J. Lewis, D. P. Oliva, T. Sheard, I. Smith, and L. Walton, "A software engineering experiment in software component generation," in *Proceedings of the 18th international conference on Software engineering*. IEEE Computer Society, 1996, pp. 542–552.

[3] S. Kelly and J.-P. Tolvanen, "Visual domain-specific modelling: Benefits and experiences of using metacase tools," in *International Workshop on Model Engineering, at ECOOP*, vol. 2000. Citeseer, 2000, pp. 1–9.

[4] A. N. Johanson and W. Hasselbring, "Effectiveness and efficiency of a domain-specific language for high-performance marine ecosystem simulation: a controlled experiment," *Empirical Software Engineering*, vol. 22, no. 4, pp. 2206–2236, 2017.

[5] T. Kosar, M. Mernik, and J. C. Carver, "Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments," *Empirical software engineering*, vol. 17, no. 3, pp. 276–304, 2012.

[6] T. Kosar, S. Gaberc, J. C. Carver, and M. Mernik, "Program comprehension of domain-specific and general-purpose languages: replication of a family of experiments using integrated development environments," *Empirical Software Engineering*, vol. 23, no. 5, pp. 2734–2763, 2018.

[7] M. Fowler, "Language workbenches: The killer-app for domain specific languages," 2005.

[8] A. Van Deursen, P. Klint, and J. Visser, "Domain-specific languages: An annotated bibliography," *ACM Sigplan Notices*, vol. 35, no. 6, pp. 26–36, 2000.

[9] T. Kosar, P. E. Martı, P. A. Barrientos, M. Mernik *et al.*, "A preliminary study on various implementation approaches of domain-specific language," *Information and software technology*, vol. 50, no. 5, pp. 390–405, 2008.

[10] N. Vasudevan and L. Tratt, "Comparative study of dsl tools," *Electronic Notes in Theoretical Computer Science*, vol. 264, no. 5, pp. 103–121, 2011.

[11] N. Oliveira, M. J. Pereira, P. Henriques, and D. Cruz, "Domain specific languages: A theoretical survey," *INForum'09-Simpósio de Informática*, 2009.

[12] S. Erdweg, P. G. Giarrusso, and T. Rendel, "Language composition untangled," in *Proceedings of the Twelfth Workshop on Language Descriptions, Tools, and Applications*. ACM, 2012, p. 7.

[13] L. M. do Nascimento, D. L. Viana, P. A. S. Neto, D. A. Martins, V. C. Garcia, and S. R. Meira, "A systematic mapping study on domain-specific languages," in *Proceedings of the 7th International Conference on Software Engineering Advances (ICSEA'12)*, 2012, pp. 179–187.

[14] S. Erdweg, T. Van Der Storm, M. Völter, M. Boersma, R. Bosman, W. R. Cook, A. Gerritsen, A. Hulshout, S. Kelly, A. Loh *et al.*, "The state of the art in language workbenches," in *International Conference on Software Language Engineering*. Springer, 2013, pp. 197–217.

[15] S. Erdweg, T. Van Der Storm, M. Völter, L. Tratt, R. Bosman, W. R. Cook, A. Gerritsen, A. Hulshout, S. Kelly, A. Loh *et al.*, "Evaluating and comparing language workbenches: Existing results and benchmarks for the future," *Computer Languages, Systems & Structures*, vol. 44, pp. 24–47, 2015.

[16] T. Kosar, S. Bohra, and M. Mernik, "Domain-specific languages: A systematic mapping study," *Information and Software Technology*, vol. 71, pp. 77–91, 2016.

[17] M. Mernik, "Domain-specific languages: A systematic mapping study," in *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 2017, pp. 464–472.

[18] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM computing surveys (CSUR)*, vol. 37, no. 4, pp. 316–344, 2005.

[19] J. Thanhofer-Pilisch, A. Lang, M. Vierhauser, and R. Rabiser, "A systematic mapping study on dsl evolution," in *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2017, pp. 149–156.

[20] E. Van Wyk, D. Bodin, J. Gao, and L. Krishnan, "Silver: An extensible attribute grammar system," *Science of Computer Programming*, vol. 75, no. 1, pp. 39–54, 2010.

[21] M. Mernik, M. Lenič, E. Avdičaušević, and V. Žumer, "Lisa: An interactive environment for programming language development," in *International Conference on Compiler Construction*. Springer, 2002, pp. 1–4.

[22] J. Heering, P. R. H. Hendriks, P. Klint, and J. Rekers, "The syntax definition formalism sdf—reference manual—," *ACM Sigplan Notices*, vol. 24, no. 11, pp. 43–75, 1989.

[23] L. C. Kats and E. Visser, *The spoofax language workbench: rules for declarative specification of languages and IDEs*. ACM, 2010, vol. 45, no. 10.

[24] S. Erdweg, T. Rendel, C. Kästner, and K. Ostermann, "Sugarj: library-based syntactic language extensibility," in *ACM SIGPLAN Notices*, vol. 46, no. 10. ACM, 2011, pp. 391–406.

[25] M. Eysholdt and H. Behrens, "Xtext: implement your language faster than the quick and dirty way," in *Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion*. ACM, 2010, pp. 307–309.

[26] P. Klint, T. Van Der Storm, and J. Vinju, "Easy meta-programming with rascal," in *International Summer School on Generative and Transformational Techniques in Software Engineering*. Springer, 2009, pp. 222–289.

[27] W. Cazzola and E. Vacchi, "Neverlang 2–componentised language development for the jvm," in *International Conference on Software Composition*. Springer, 2013, pp. 17–32.

[28] E. Vacchi and W. Cazzola, "Neverlang: A framework for feature-oriented language development," *Computer Languages, Systems & Structures*, vol. 43, pp. 1–40, 2015.

[29] H. Krahn, B. Rumpe, and S. Völkel, "Monticore: a framework for compositional development of domain specific languages," *International journal on software tools for technology transfer*, vol. 12, no. 5, pp. 353–372, 2010.

[30] A. Butting, R. Eikermann, O. Kautz, B. Rumpe, and A. Wortmann, "Systematic composition of independent language features," *Journal of Systems and Software*, vol. 152, pp. 50–69, 2019.

[31] T. van Der Storm, W. R. Cook, and A. Loh, "Object grammars," in *International Conference on Software Language Engineering*. Springer, 2012, pp. 4–23.

[32] T. Van Der Storm, W. R. Cook, and A. Loh, "The design and implementation of object grammars," *Science of Computer Programming*, vol. 96, pp. 460–487, 2014.

[33] M. Voelter and V. Pech, "Language modularity with the mps language workbench," in *2012 34th International Conference on Software Engineering (ICSE)*. IEEE, 2012, pp. 1449–1450.

[34] R. Solmi, "Whole platform," Ph.D. dissertation, PhD thesis, University of Bologna, 2005.

[35] C. Simonyi, M. Christerson, and S. Clifford, "Intentional software," in *ACM SIGPLAN Notices*, vol. 41, no. 10. ACM, 2006, pp. 451–464.

[36] D. Waggoner, M. A. Jensenworth, P. Kwiatkowski, and C. Simonyi, "System and method for combining text editing and tree encoding for computer programs," Jun. 13 2017, uS Patent 9,678,724.

[37] S. Kelly, K. Lyytinen, and M. Rossi, "Metaedit+ a fully configurable multi-user and multi-tool case and came environment," in *International Conference on Advanced Information Systems Engineering*. Springer, 1996, pp. 1–21.

[38] E. Visser *et al.*, *Scannerless generalized-LR parsing*. Universiteit van Amsterdam. Programming Research Group, 1997.

[39] M. Voelter, J. Siegmund, T. Berger, and B. Kolb, "Towards user-friendly projectional editors," in *International Conference on Software Language Engineering*. Springer, 2014, pp. 41–61.

[40] F. Steimann, M. Frenkel, and M. Völter, "Robust projectional editing," in *Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering*. ACM, 2017, pp. 79–90.

[41] M. Völter, S. Benz, C. Dietrich, B. Engelmann, M. Helander, L. C. Kats, E. Visser, and G. Wachsmuth, "Dsl engineering-designing, implementing and using domain-specific languages (2013)," *URL: http://voelter. de/dslbook/markusvoelter-dslengineering-1.0. pdf, http://dslbook. org*, 2013.

[42] OMG, "Mof 2.0 query/view/transformation (qvt) adopted specification." 2005, oMG document ptc/05-11-01.

[43] M. Bravenboer, K. T. Kalleberg, R. Vermaas, and E. Visser, "Stratego/xt 0.17. a language and toolset for program transformation," *Science of computer programming*, vol. 72, no. 1-2, pp. 52–70, 2008.

[44] T. J. Parr and R. W. Quong, "Antlr: A predicated-ll (k) parser generator," *Software: Practice and Experience*, vol. 25, no. 7, pp. 789–810, 1995.

[45] M. Voelter and K. Solomatov, "Language modularization and composition with projectional language workbenches illustrated with mps," *Software Language Engineering, SLE*, vol. 16, no. 3, 2010.

Fine-tuning Resource Allocation of Apache Spark Distributed Multinode Cluster for Faster Processing of Network-trace Data

Shyamasundar L B*[1]
Department of CSE
CMR Institute of Technology
Bengaluru, Karnataka, India

V Anilkumar[2]
CSIR-Fourth Paradigm Institute
NAL Belur Campus
Bengaluru, Karnataka, India

Jhansi Rani P[3]
Department of CSE
CMR Institute of Technology
Bengaluru, Karnataka, India

***Abstract*—In the field of network security, the task of processing and analyzing huge amount of Packet CAPture (PCAP) data is of utmost importance for developing and monitoring the behavior of networks, having an intrusion detection and prevention system, firewall etc. In recent times, Apache Spark in combination with Hadoop Yet-Another-Resource-Negotiator (YARN) is evolving as a generic Big Data processing platform. While processing raw network packets, timely inference of network security is a primitive requirement. However, to the best of our knowledge, no prior work has focused on systematic study on fine-tuning the resources, scalability and performance of distributed Apache Spark cluster (while processing PCAP data). For obtaining best performance, various cluster parameters like number of cluster nodes, number of cores utilized from each node, total number of executors run in the cluster, amount of main-memory used from each node, executor memory overhead allotted for each node to handle garbage collection issue, etc., have been fine-tuned, which is the focus of the proposed work. Through the proposed strategy, we could analyze 85GB of data (provided by CSIR Fourth Paradigm Institute) in just 78 seconds, using 32 node (256 cores) Spark cluster. This would otherwise take around 30 minutes in traditional processing systems.**

Keywords*—Big data; packet data analysis; network security; distributed apache spark cluster; Yet Another Resource Negotiator (YARN); parameter tuning

I. Introduction

Big data could be defined as data with high variety, volume, velocity and veracity information assets [1]. It claims optimal, cost effective and innovative techniques of processing information. Which in turn provides better insight and much better decision making. Such processing is difficult to conduct using centralized approaches in a highly scalable, high-throughput and fault-tolerant way [2].

In the proposed work, the PCAP data analysis will be implemented on top of a SPARK cluster which is deployed over Hadoop YARN (Yet Another Resource Negotiator).

The proposed work analyzes large amount of network capture data that has been collected for a period of four months (which constitutes to Big data). The collected data will be in *.pcap* format. The complex PCAP analysis includes processing huge amount of data collected. The processing must be performed on large stored datasets by an analyst to detect security incidents or to perform security audits.

A. Apache Hadoop versus Apache Spark

Apache Hadoop uses MapReduce processing framework, by using YARN for cluster management and Hadoop Distributed File System (HDFS) for distributed storage. Hadoop MapReduce provides fault-tolerant and distributed execution of 'jobs', which includes the following processing steps:

1) Reading the input data from HDFS blocks and splitting them to mappers.
2) Map: Applies an user-defined function and outputs one file per mapper.
3) Combining output from the mappers (using user defined functions), which is optional.
4) Partition, shuffle, sort and merge the data into the reducers.
5) Reduce the data.
6) Output one file to each reducer and store data into HDFS.

Tasks where Hadoop MapReduce is chosen over SPARK:

- *Processing huge data sets in a linear manner:* MapReduce allows processing huge amount of data parallelly, where large chunk of data is broken into smaller ones. They are processed separately on the data nodes. The results are gathered automatically across multiple nodes and then a single result is returned. If the dataset is bigger than the available RAM, then MapReduce may outperform SPARK.
- *When immediate results are not expected:* MapReduce will be a good solution only if the processing speed is not critical.

Tasks where Apache Spark is chosen over Mapreduce:

- *Faster data processing:* In-memory data processing makes Spark much faster than processing using MapReduce. In essence, 10 times faster processing of data in storage and 100 times faster using Random Access Memory (RAM) [3] [4] [5].
- *Iterative data processing:* If the data is processed often, then SPARK can be chosen over Mapreduce. Multiple map operations can be done using Spark's Resilient Distributed Datasets (RDDs), whereas MapRe-

duce has to write intermediate results back to a disk, which increases input/output (I/O) overhead.

- *Near real-time data processing:* If immediate insights are needed, then Spark should be opted since it performs in-memory processing.
- *Processing graphs:* Computational model of Spark is good to perform iterative computations since it has GraphX, which is dedicated for graph computation.
- *Machine learning tasks:* Spark has built-in machine learning library - MLib which has out-of-the-box algorithms that run in memory. Whereas, Hadoop Mapreduce requires a third-party to provide a library for machine learning. Also in SPARK, there is a provision to tune and adjust the algorithms.
- *Dataset joins:* Spark can perform combinations much faster, while Hadoop requires many shuffles and sorts for joining datasets.

B. Motivation for the Work

In recent times, Apache Spark in combination with Hadoop is evolving as a generic big data processing platform. While processing raw network packets, timely inference of network security is a primitive requirement. However, to the best of our knowledge, no prior work has focused on systematic study on fine-tuning the resources, scalability and performance of Apache Spark cluster in a multi-node environment (while processing PCAP data). For obtaining best performance, various cluster parameters like number of nodes in the cluster, number of cores utilized from each node, total number of executors run in the cluster, amount of RAM used from each node, YARN executor memory overhead allotted for each node to handle garbage collection issue, etc. have been fine-tuned, which is the focus of the proposed work.

II. Background and Related Work

Apache spark is specially designed for handling big data processing problems, which can analyze data in a very less time. It is a cluster computing platform that is open source and designed for processing big data. It provides an user friendly application program interface (API) to write queries and handle the jobs [6]. It has basic functionalities such as memory management, task scheduling, interaction with storage systems, fault recovery and RDDs, which are the main programming abstraction. For parellel processing of data, a set of functions are distributed across memory, which involves transformations and actions. *Transformations* include reduceByKey, distinct, map, intersection, join, union, filter, aggregateByKey, sortByKey and so forth. *Actions* include collect, saveAsTextFile, count, first, countByKey, takeSample, foreach and so forth [7].

Apache Spark elegantly handles iterations, memory availability and is suitable for processing both stream and batch jobs. Overall it outperforms Hadoop by orders-of-magnitude for several applications.

As illustrated in Fig. 1, Spark has three major layers. *Spark Core* - a generalized layer where all the basic functions are defined and all the other extensions and functionalities are developed on top of Spark Core. *Spark Ecosystem* contains additional libraries to operate on top of DataFrames and Spark Core. These components give power in the fields of machine learning, Structured Query Language (SQL) capabilities, real time processing of big data etc.

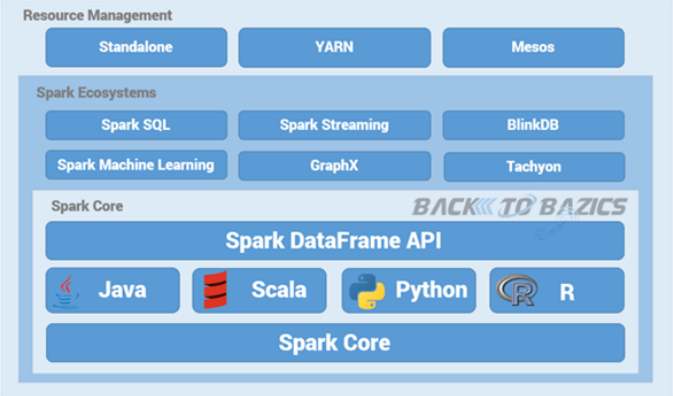


Fig. 1. Apache Spark Architecture and Ecosystem

A. Job Execution on a Spark Cluster

When a job is submitted by a driver process, first the request is sent to the *YARN Resource manager*. YARN checks for data locality and performs task scheduling by finding best available slave nodes. Then the submitted job will split into several stages. Each stage in-turn splits into multiple tasks, based on available resources and data locality. *Driver* daemon will send the necessary details related to the job to each node, prior to execution of task. Currently executing tasks are tracked by the driver and updates are sent to the *master* node. This can be checked with Master Node's user interface. Aggregate values from all the nodes are shared with the master node, once the job is completed. Hadoop and Spark coexist in the same cluster as shown in Fig. 2.

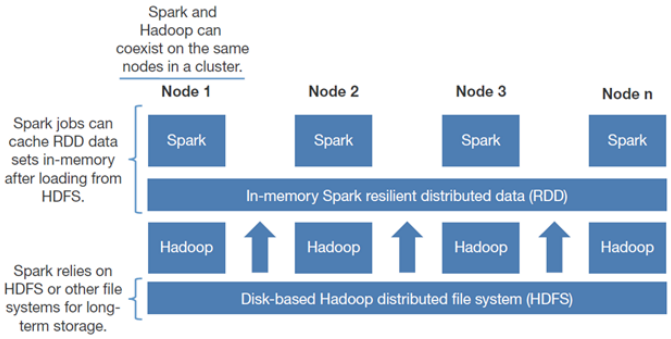


Fig. 2. Apache Spark and Apache hadoop co-existing on the same cluster

There exists an *ApplicationMaster* process for each application instance in YARN. Application requests resources from *ResourceManager* and informs *NodeManagers* to initiate containers on its behalf, after allocating the resources. The *Spark driver* will run in the *ApplicationMaster* on top of cluster host, once the resources have been allocated. The process is shown in Fig. 3 and 4.

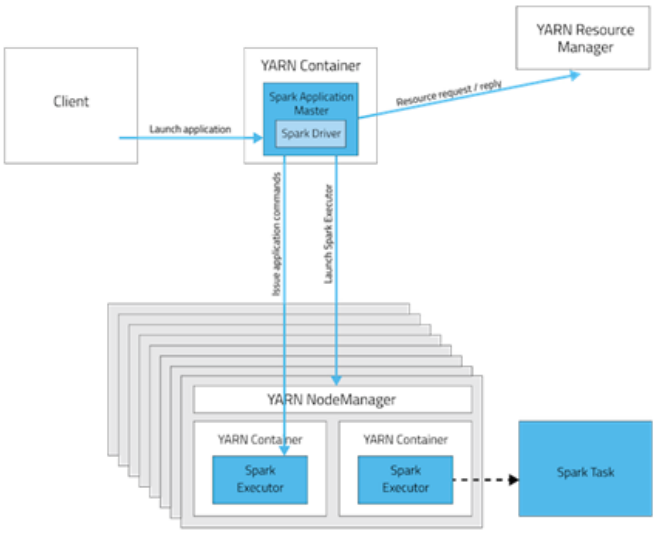


Fig. 3. Spark daemons running when deployed on top of YARN

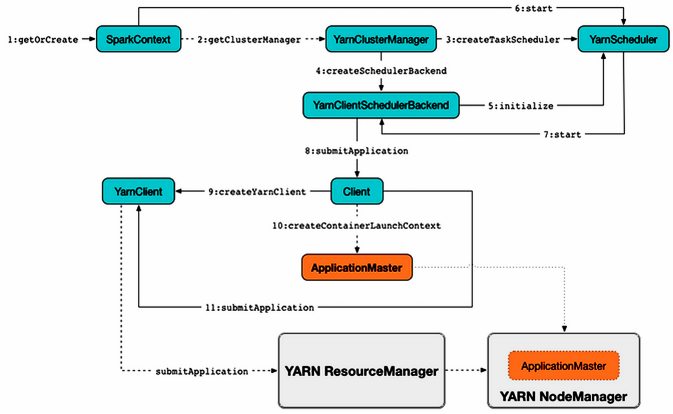


Fig. 4. Submitting Spark Application to YARN Cluster (aka Creating SparkContext with yarn Master URL and client Deploy Mode)

Spark includes all the necessary transformations that are pipe-lined in just a single stage for boosting the performance. Then, the data must be *shuffled* among different stages [8]. When a shuffle is done, in-memory output will be shuffled from previous stage to the storage system. Extra data will be stored to disk, if insufficient memory is allocated and intermediate data is transferred across the network. When employing more machines, due to shuffling overhead, it will achieve speed-ups much smaller than 'm', the number of machines (nodes); In general, overhead incurred due to shuffling is directly proportional to number of machine pairs, i.e.,

$$O(m^2)$$

.

Bachupally [9] has analyzed the network trace data and found anomalous connections made to the network by using HDFS. Some features of PCAP data with size 131 Megabytes (MB) were extracted using Wireshark and stored in HDFS as Comma Seperated Value (CSV) format. However, Wireshark has been used and very small data has been handled. Wireshark [10] is an opensource software and one of the most famous analysis tools for network packets. But there are some limitations.

In [11], authors present four different network monitoring tools that can monitor and analyze the network traffic. The disadvantage of Wireshark has been discussed. According to them, it will not detect malicious activities on the network and it means that Wireshark may not be useful to study network security. In addition, Wireshark cannot handle the large packet data. If the user has a large capture file more than 100MB, Wireshark will become slow while loading, filtering and alike actions.

In [12], attention is paid on the benefits of Apache Spark when compared to Hadoop MapReduce. In case of MapReduce, data will be read from disk and results are written to HDFS after a specific iteration. Then, the data will be read from HDFS for further iteration. The entire process utilizes lot of time and disk space. This results in the issue of lower fault tolerance and high latency of the entire system. To overcome these issues, Apache Spark is being used instead of MapReduce. The authors also focused on time-series data analysis using SPARK environment in real-time. Patterns were generated out of analysis, which in-turn gave a clear glimpse of characteristics and statistics of data. Thus, making MapReduce less efficient compared to SPARK. But the dataset under consideration is not related to packet capture data and the dataset used in the proposed work requires resource allocation to be tuned, which has not been explored.

In [13], SQL queries have been executed to analyze about 85 Gigabyte (GB) of network packet dataset provided by the University of New Brunswick. The way to analyze huge size of PCAP files on Hadoop and visualize the analyzed results on web browser by using Hue (Hadoop User Experience) has been discussed.The experiments were conducted using MapReduce, whereas experiments in the proposed method are done using SPARK. It has been set up on a standalone mode in Hadoop and SQLs are executed on a local machine, and not on a distributed cluster.

In [14] and [15], a method is proposed where input data is converted into RDDs. This allows in-memory computation on huge clusters of Spark, in a fault-tolerant way. Lazy transformations are applied to RDDs, which in-turn creates new RDDs and store them into HDFS or onto the driver. This method has been incorporated in the proposed work.

In [16], researchers integrate a network monitoring architecture into Apache Spark. NetFlow data has been processed as the way in traditional processing approaches. Moreover, implementation using stream processing will find out novel information which could not be found using traditional packet monitoring approaches.

Prakasam [17] expresses that, due to shortage of options in inter-stage communication facilities of processes, it makes MapReduce unsuitable for interactive workloads (Interactive Data mining, Stream data processing and Analysis) and iterative (Machine Learning and Graph Processing) processing. He dives into architecture of MapReduce and its disadvantages. At the end, alternatives such as Apache Spark and Apache Tez were used because of their suitability in interactive and iterative processing.

In [18] and [19], authors conclude that execution time of a particular job on Apache Spark platform can vary significantly depending on the input data type and size, method and implementation of the algorithm and computing capability (e.g., number of nodes, memory size, Central Processing Unit - CPU speed etc). This makes it extremely difficult to predict job performance, which is often needed to optimize resource allocation. Performance prediction can help to locate execution stages with abnormal resource usage pattern [20]. Although these issues are considered, the nature of PCAP data is different from their data and this requires different resource allocation mechanisms based on the number of cores to be allotted for data processing. This issue is resolved in the proposed work.

In [21], authors claim to propose a distributed Intrusion Detection System (IDS) based architecture that is capable of detecting anomalies in the network in real-time using Apache Spark framework and Netmap. But, Center for Applied Internet Data Analysis - CAIDA's dataset [22] has been used to detect Distributed Denail of Service (DDoS) attacks, which is an offline dataset. Their setup takes more than 3 minutes to detect an attack. In real scenarios, > 3 minutes of time to find an attack is not an optimal solution. Proposed framework has a greater performance speed and the cluster can handle 3 GB of data in seconds, since it is possible to process 85 GB of data in 78 seconds. Also, the query that has been posed doesn't execute the jobs in parallel, which is the reason that their jobs take more time for processing.

In [23], authors claim that the analysis is done on real time data. But offline dataset "KDD 1999" has been used and supervised ML technique is used for processing the data. Also, the size of the dataset is not huge enough to fit into big data, which could be processed using traditional systems itself, without the need of Apache SPARK. The size of the dataset is just 743MB. But dataset considered in the proposed work is 85GB. Moreover, there is no mention about the configurations of SPARK cluster, regarding number of nodes in the cluster, number of cores, size of RAM on each system, etc... This makes it difficult to compare performance of the proposed work with their results.

None of the above referenced papers concentrate on fine-tuning the resource allocation of spark cluster, which enhances the performance by controlling excess usage of resources or limiting garbage collection.

The initial cluster startup takes around 12 seconds, independent of size of the cluster. In real world, applications run for dozens of minutes, which is an acceptable overhead. During each execution of the job, intermediate data will be produced which are stored on local disks and not on HDFS (as in [24]). This helps to yield the best performance.

From these research reviews, it is found out that one of the most challenging task is to handle the large packet data and analyze it. Thus, in this study, the designed analysis environment could handle the large packet data by using Apache SPARK on top of Hadoop YARN cluster.

III. Theory Framework and Modeling

A. Cluster Setup and Spark Application Submission

Spark scripts read a configuration file, describing the application and the Spark cluster configuration, provided by the user. This submits one or more jobs to the workload manager. Once a job is chosen by the job scheduler, a definite number of nodes are reserved for HDFS (if requested) and Spark cluster. Different services will be started after resource allocation procedure. The HDFS namenode service will be started if Distributed File System (DFS) is requested. Then all the HDFS datanode services will be started and connected to the NameNode. Spark cluster setup is done in a similar manner (wait for master node to be ready and then start the worker nodes). In Apache Spark, *master* node is the standalone Spark manager and *worker* nodes are nothing but Spark worker services, the place where the executors are launched.

B. Resource Allocation Schemes

Apache Spark doesn't provide any storage (like HDFS) or any Resource Management capabilities. It is just a unified framework for processing large amount of data near to real time [25]. It accesses Hadoop data store (HDFS) and runs on top of existing Hadoop cluster. Spark allows applications in Hadoop cluster to run up to 10x faster when running on disk and 100x faster when running in memory. This is possible by reducing the actual number of read/ write operations to and from the disc. Spark also gives the feature to quickly write applications in Scala, Java, or Python. This in return helps developers to create and run their applications in the programming languages they are more familiar with, thereby making it easy to build parallel apps.

Worker nodes will have the available resources (memory, CPU cores and disk). Master node is responsible for allocating the available resources to the required applications. Each application create executor processes where the tasks run in parallel. Resource allocation can be done using the following three mechanisms:

- **Default:** Here the applications are submitted without specifying any details of resource allocation. All the applications run in a First In First Out (FIFO) manner and every application consume the resources from all the worker nodes. Hence, when a single application is running, it will utilize all the worker nodes and create the executors.
- **Static:** Here, the user will specify the number of cores, executors, memory etc. that an application can have. The resources are shared among multiple applications submitted by one or more users.
- **Dynamic:** Here, applications may release executors that are idle and give back some of the resources to the Spark cluster. The free resources can be taken back in future, if required.

However, each one of the resource allocation schemes have some problems.

- Firstly, when only one application is running with default resource allocation scheme, it consumes all

the resources. Hence, resources are not shared among applications.

- Secondly, in case of static resource allocation scheme, user manually sets the resources that each application will use.
- Thirdly, with dynamic resource allocation scheme, the initial amount of resources is still set by the user. Hence, incorrect resource allocation will cause severe performance issues.

  Finally, if any production cluster demands user-specific deadlines, then default allocation of resources may not work because application having a strict deadline has to wait in FIFO queue. Also, inappropriate allocation of resources in both dynamic and static resource allocation schemes will affect the deadlines.

C. SPARK_CSV (Comma Separated Value) Library Package

In the proposed scheme, network trace files in PCAP format are converted to CSV format to ease the purpose of querying the data files on top of SPARK framework. This package reads CSV files as Spark DataFrames. The API accepts many different options when reading the files:

- **path:** Location of files are specified here.
- **header:** When this variable has been set true, the first line in the file will be used for naming the columns and will be excluded from the data. By default, it's value is false.
- **delimiter:** Columns will be delimited using ',' (comma) by default. But any character can be set as delimiter.
- **quote:** The quote character is " (double quote) by default. But any character can be set. Delimiters inside the quotes will be ignored.
- **escape:** Escape character is \ (backslash) by default. One can set it to any other character.
- **parserLib:** It is "commons" by default and one can set it to "univocity" for CSV parsing.
- **mode:** Determines the mode of parsing. It is PERMISSIVE by default. Possible values include:

  PERMISSIVE: It parses all the lines. Missing tokens are inserted by null and extra tokens will be ignored.

  DROPMALFORMED: Drop the lines that have more or fewer tokens than expected or contain tokens that do not match with the schema

  FAILFAST: If any malformed line is encountered, it aborts with *RuntimeException.*
- **charset:** It will be 'UTF-8' by default and can be set to any other charset names.
- **inferSchema:** Column types are inferred automatically and by default it is set to false.
- **comment:** Lines starting with a specific character are ignored. "#" is the default. One can disable comments by having value for the variable as *null.*
- **nullValue:** A string is specified that indicates null value. Any other fields that match this string are set as nulls.
- **dateFormat:** Indicates format of the strings to use while reading timestamps or dates.

D. Tuning Resource Allocation

There exist some situations where even-though a 100-node cluster is setup and an application is run, only two tasks are executing. These kind of situations are not unfair, when a number of parameters influence resource utilization of Spark. So, in the proposed work, for the best performance the target is to make best use of the available resources from the cluster.

If resource allocation is not configured correctly, submitted job may consume all of the cluster resources and in-turn other applications will starve for resources.

The steps for a Spark job in a cluster mode include:

- *SparkContext* will connect to YARN cluster manager from the driver node.
- Resources are allocated to the cluster manager across other applications.
- Spark will acquire executors on the nodes in a cluster where each application will get it's own executor processes. Application code (python files/jar/python egg files) will be sent to executors.
- Tasks will be sent by *SparkContext* to these executors.

This has been shown in Fig. 5.

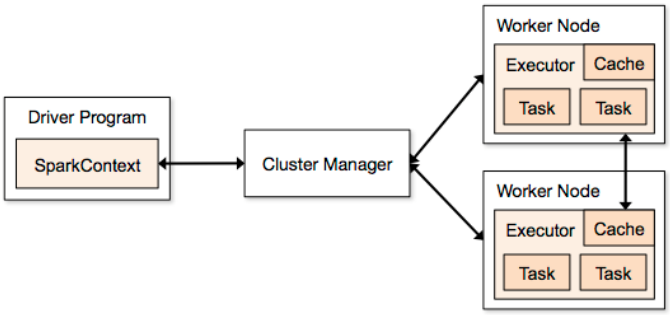


Fig. 5. Steps involved in cluster mode for a Spark Job

From the above mentioned steps, it is very clear that the total number of executors and their allocated memory setting will play an important role in spark's job performance. If executors are run with too much memory, it results in delays due to excessive garbage collection.

E. Utilized Dataset Description

In this paper, a collection of raw network packets obtained from a "/24" darknet setup has been utilized. A darknet is a set of globally routed and valid Internet Protocol (IP) addresses, which are not assigned to any host or devices. As the IP address space is not assigned to any hosts or devices, under ideal condition, no genuine traffic is expected to reach the darknet (Thus only malicious traffic has been captured,

which constitutes to 85GB collected over a period of four months). However, Internet measurement and analysis in the past have shown that substantial amount of traffic arrives at darknet IP address space, and they are typically triggered by malicious activities like Internet wide port scanning, botnet recruitment and expansion process, reflections from IP spoofed Denail-of-Service (DoS) and DDoS attacks, etc. The darknet packets were collected continually for a period of 122 days (4 months) from 01/06/2017 to 30/09/2017. It consists of about 500 million (õ.5 billion) Transmission Control Protocol (TCP) three-way handshake packets: Synchronize (SYN) request, SYN/Acknowledgment (ACK) response and ACK confirmation. The packets were originated from all over the world and were subjected to source IP address spoofing by responding to incoming SYN requests with appropriate SYN/ACK responses and validating the third ACK packet suing the sequence and acknowledgment number of these packets.

F. Utilized Testbed Description

The experiments were performed on a dedicated multi-node network testbed. It consists of 60 rack-mountable servers of 1U size and interconnected using 1 Gigabit per second (Gbps) Ethernet switches. Each node consists of one Intel Xeon E3-1230 v3 processor @ 3.30 GHz, with 8 MB cache. The processor consists of 4 cores and hyper threading is enabled on each core leading to an effective number of 8 CPUs available for concurrent job execution. Each node is also equipped with 2 numbers of 3 terabyte SATA hard disk drives at 7200 rotations per minute for storing the network packets. The testbed is hosted in a self-contained data center spread across three racks and equipped with inbuilt cooling, uniterrupted power supply, fire detection and suppression, water leakage detection, etc. As described under section 4, in order to determine the performance and scalability, multiple experiments are conducted with configurations of 1, 2, 4, 8, 16 and 32 nodes (leading up to 256 SPARK executors concurrently) in the testbed, with turbo mode off. HDFS and Apache SPARK are configured on the nodes. Each node runs Linux Operating System and Java version 8.

IV. Description of Models Involved, while Evaluating the Performance of Spark

A. Model to Estimate Execution Time

Since a job will be executed in several stages and each stage will be containing several tasks, jobs and stages can be represented as in equations 1 and 2, respectively.

$$Job = \{Stage_a \mid 0 \leq a \leq X\} \tag{1}$$

$$Stage_a = \{Task_{a,b} \mid 0 \leq b \leq Y\} \tag{2}$$

Where, the number of stages running within a job is X and number of tasks running within a stage is Y. Since different stages inside the job are sequentially executed, one can represent the time taken for executing a job as summation of execution times at each stage, *startup time* for the job and *cleanup time* for the job. This has been represented in equation 3.

$$SparkJobTime = StartupTime + \sum_{k=1}^{X}(Stage\_Time_k) + CleanupTime \tag{3}$$

In each of the stages, one core of CPU will execute a single task at a time. If a cluster has *W* worker nodes, one can calculate *P*, the number of parallel tasks as per equation 4.

$$W = \sum_{a=1}^{W}(Num\_Cores_a) \tag{4}$$

Where, the number of cores in CPU of a slave (worker) node *a* is Num_Cores_a and *W* is number of slave nodes within the cluster. Therefore, in each stage of execution, tasks will execute in batches and each batch contains *W* parallel running tasks. However, in a heterogeneous cluster, if the computing capacity of each slave is different, there will be an inherent uncertainty during execution which results in a significant variation in the execution time of the submitted job.

Hence, the time that has been spent in each stage can be calculated by doing a summation of the time taken for execution of sequential tasks in each stage, startup time of each stage and cleanup time of each stage, as represented in equation 5.

$$Stage\_Time = StartupTime + \max_{c=1} W \sum_{a=1}^{S_c}(Task\_Time_{c,a}) + CleanupTime \tag{5}$$

Where total number of CPU cores (number of tasks running in parallel) is *W*, number of tasks running sequentially on each core is S_c.

Finally, since several tasks within a stage run with the same execution pattern, the time taken for executing a task can be calculated as per equation 6.

$$Task\_Time = Deserialize\_Time + Run\_Time + Serialize\_Time \tag{6}$$

Where, time taken for deserialization of input data is *DeserializeTime*, time taken for serialization of input data is *SerializeTime* and the actual amount of time spent to perform several operations on input data like *map, filter, etc.* is *Run_Time*.

B. Model to Estimate Memory Consumption

Since Spark performs in-memory computations, sufficient memory must be allocated to avoid execution slowdown, while creating the RDDs. Sometimes during configuration settings, with a lack of memory will lead to unexpected termination of the program execution. To avoid these adverse effects, a model to calculate the minimum amount of memory required for creating RDDs is proposed. Specifically, if *Y* tasks are running, once can define the total amount of memory

required for execution of a job as summation of time required for executing each task. This has been defined by equation 7.

$$JobMemRDD = \sum_{a=1}^{Y}(TaskMemRDD_a) \tag{7}$$

C. Model to Estimate I/O Cost

Within a stage, current RDDs are generated by making use of previously generated RDDs using the transformation operation, *Shuffle_Map*. The result data is generated by performing *Result* operation. The cost of I/O operations involved in the above tasks can be categorized into two types, namely: *ShuffleReadCost* and *ShuffleWriteCost*. *ShuffleWriteCost* is the cost incurred while storing the interim data onto the disk buffer and *ShuffleReadCost* is the cost incurred while fetching interim data from the slave nodes.

Shuffle stage is the most I/O intensive stage that involve data transmission (reading data from the slaves) and fetching (storing data onto the disk) in a frequent manner. Stage wise I/O cost can be calculated by following equations 8 and 9.

$$Stage\_IO\_Write_a = \sum_{b=1}^{Y}(Task\_IO\_Write_{a,b}) \tag{8}$$

$$Stage\_IO\_Read_a = \sum_{b=1}^{Y}(Task\_IO\_Read_{a,b}) \tag{9}$$

D. Model to Predict the Performance

At first, calculate the time taken for execution of a job. For that, the actual number of tasks being executed is found out by equation 10.

$$Y = \frac{Input\_Size}{Block\_Size} \tag{10}$$

Where, *Input_Size* is nothing but total size of the dataset and *Block_size* represents the actual one data block stored in HDFS. As discussed, tasks will run batch wise within a job and the total number of tasks running is each batch is computed as per equation 4. If a cluster has nodes with different computing capabilities, The average amount of time taken to execute a task in a stage for a worker node w, can be calculated as per equations 11 and 12.

$$Task\_Run\_Time_{w,a} = Deserialize\_time_{w,a} + Run\_Time_{w,a} + Serialize\_Time_{w,a} \tag{11}$$

$$Avg\_Task\_Time_w = \frac{1}{n_w}\sum_{a=1}^{n_w}(Task\_Run\_Time_{w,a}) \tag{12}$$

Where, the number of tasks being run in the worker node within a stage of the submitted job w is n_h. During the experimentation, there were slight differences in the execution time of different batches within a particular stage. This can be calculated as per equation 13.

$$RatioOfTimeDifference_w = \frac{\frac{1}{n_w - W_w}\sum_{a=W_w+1}^{n_w}(Task\_Time_{w,a})}{\frac{1}{W_w}\sum_{b=1}^{W_w}(Task\_Time_{w,b})} \tag{13}$$

Where, the number of tasks that are running in a worker node w is n_w and the number of tasks running in a batch is W_w. Since tasks are executed on different worker nodes in parallel, for predicting the time taken for execution of a particular stage, the *StartupTime* and *CleanupTime* remain constant. Then, time taken for execution of a stage and task can be calculated as in equations 14 and 15.

$$Est\_Stage\_Time = StartupTime + \max_{c=1} W \sum_{a=1}^{S_c}(Avg\_Task\_Time_{c,a}) + CleanupTime \tag{14}$$

$$Est\_Task\_Time_{c,a} = \begin{cases} Avg\_Task\_Time_c, & a = 1 \\ Avg\_Later\_Task\_Time_c, & a > 1 \end{cases} \tag{15}$$

Where, the total number of cores of CPU is W as per equation 4. The number of sequential tasks that are running in each core is S_c.

$Avg\_Task\_Time_c$ gives us the average time to execute batch of a CPU core within a worker node as in equation 12. The average time taken to execute following tasks of batches, $Avg\_Later\_Task\_Time_c$ can be calculated as in equation 16.

$$Avg\_Later\_Task\_Time_c = RatioOfTimeDifference_w * Avg\_Task\_Time_w \tag{16}$$

For predicting the I/O costs, the average of shuffle R/W costs incurred by a task is computed. Then, I/O cost incurred for a particular stage b can be calculated as specified in equations 17 and 18.

$$Est\_Stage\_IO\_Write_b = \sum_{w=1}^{W}(Y_w * \frac{1}{n_w}\sum_{a=1}^{n_w}(Task\_IO\_Write_{w,a})) \tag{17}$$

$$Est\_Stage\_IO\_Read_b = \sum_{w=1}^{W}(Y_w * \frac{1}{n_w}\sum_{a=1}^{n_w}(Task\_IO\_Read_{w,a})) \tag{18}$$

Where, the number of slaves (worker nodes) is W, Y_w is the total number of tasks that are running on worker node w and n_w is the number of tasks that are running on worker node w, at stage b.

Finally, calculate average RDD memory footprint required for each stage as defined in equation 19.

$$Est\_RDD\_Mem = \sum_{w=1}^{W}(\frac{Y_w}{n_w}\sum_{a=1}^{n_w} Task\_RDD\_Mem_{w,a}) \quad (19)$$

E. Estimation and Evaluation of Spark Performance, with an Example

Now, once could understand some basic analytics and flow of execution of an application in Spark, with an example.

Spark will schedule an application for execution, in a distributed cluster platform. RDDs generated within Spark will be used to identify the way temporary data and inputs are generating during computation phase. RDDs will be divided into equal sized partitions that could be configured to store on disk, in memory, or both. Multiple operations are pipelined within every partition to execute several operations in parallel. Since each partition will handle a task, the number of jobs running will be equal to the number of partitions created.

As illustrated in Fig. 6, considering a *join* operation, inputs are taken from multiple partitions. A stage will be created for the input data, so that the execution of the tasks are made to run in parallel manner. Considering the example above, *map* operation executes in parallel taking the RDD from Hadoop, on all the partitions, which will be aggregated to Stage 1 (grey part of the diagram). The *filter* operation also holds the same logic (green part of the diagram). Then, the *join* operation needs to wait for stages 1 and 2 to finish and after the results are obtained from these stages, a separate stage 3 will be scheduled. The intermediate results generated from mapped and filtered RDDs are logical RDDs, representing intermediate results.

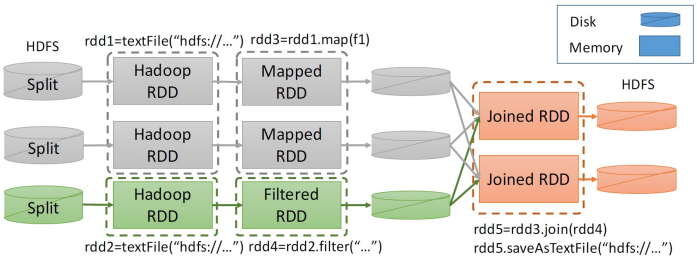


Fig. 6. Scheduling a Spark application for execution

If fewer machines are allocated, stages 1 and 2 cannot run concurrently and will run sequentially, thus the execution slows down. The end user of Spark application will have the flexibility to configure the storage location of these RDDs and also the size of them. If the RDDs stored in memory, care has to be taken so that entire RDDs fit into the memory. Otherwise, some partitions are spilled to the disk, or recomputed later, or even discarded.

In the phase of *Application Profiling*, several times an application is executed and entire dataset is divided into subsets of input data with an increased size (In real scenarios, this can be done using sampling), for evaluation of its dependency upon the important tuning parameters. During evaluation of performance, an entire dataset will be considered to estimate each stage's execution time from the collected information and then the entire application's performance is evaluated by considering the estimated time of execution for each stage.

The *Application Profiling* stage analyzes how tuning parameters will affect the overall performance of a Spark application in terms of storage and time taken to complete the entire task.

Now, consider the parameter *storage.memoryfraction* to analyze the parameters related to memory that affects the performance of a Spark application. In essence, Spark will divide available at every slave node into two important areas. The first part will account to 60% of the memory available by default, that will be allocated to caching of RDDs. The second part will account to 20% of remaining memory to the computations of *shuffle* stage. Remaining memory will be used for the computations of functions defined by the user. *storage.memoryfraction* parameter defines the memory being allotted to cache RDDs. If in case, some parts of RDDs does not fit in the main memory, those RDDs are spilled to the disk, or recomputed later, or even discarded. This will definitely affect the performance of the application.

To deeply understand the situation where entire RDDs does not fit into the main memory and to understand the effect of *storage.memoryfraction* parameter, some experiments were conducted. The entire dataset used for the experiments constituting to 85GB and minimal memory was considered, where all the RDDs will not fit into the main memory. From the results, it was observed that, the required time for completion of processing task increased dramatically (as per Fig. 7) when the memory utilized for caching RDDs, exceeds some value (i.e., 0.76).

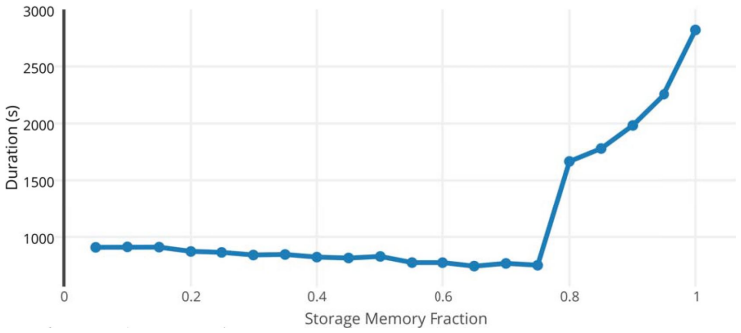


Fig. 7. Tuning *storage.memoryfraction* parameter

This concludes that, there was insufficient memory for *shuffle* operations and most of the memory was utilized for caching RDDs. In these cases, RDDs are spilled to the disk which incurs I/O overhead. Considering these cases, for the experiments conducted in the proposed work, care is taken that entire RDDs fit into the main memory to achieve the best possible performance.

V. Results and Discussion

In the proposed work, 4 months of network trace data which contains 85GB of data has been analyzed. This data has

been processed in stages of 1 month, 2 months and 4 months. 32 nodes have been used from the test bed, each having 8 cores of CPU with 32GB of RAM. This data has been processed using different use-cases considering 5 cores, 7 cores and 8 cores of CPU from each node. With these setups, 7 cores per node achieved best results.

All Spark executors of an application will have same number of cores that are fixed. The total number of cores could be specified either with *–executor-cores* flag while invoking *spark-shell, spark-submit and pyspark* from command line, or by setting *spark.executor.cores* property in *spark-defaults.conf* file or through a *SparkConf* object. The cores control the number of tasks that an executor can concurrently run. For example, "*–executor-cores 7*" implies that each executor will run with a maximum number of seven tasks at the same time.

The *spark.executor.instances* configuration property or *–num-executors* command-line flag controls the total number of executors requested. One can avoid setting of this property by switching to dynamic allocation with *spark.dynamicAllocation.enabled* property, which enables a Spark application by requesting executors during backlog of pending tasks. This frees up executors when idle.

It is also important to tune how the resources requested by Spark could fit into what is available with YARN. The relevant properties of YARN are:

- *yarn.nodemanager.resource.memory-mb:* This controls maximum total memory that can be used by the containers present on each node.
- *yarn.nodemanager.resource.cpu-vcores:* This controls maximum total cores that can be used by the containers present on each node.

Fig. 8 shows a geographical distribution of the origin of the darknet malicious packets mapped using IP2Location mapping from the source IP address of the SYN requests.

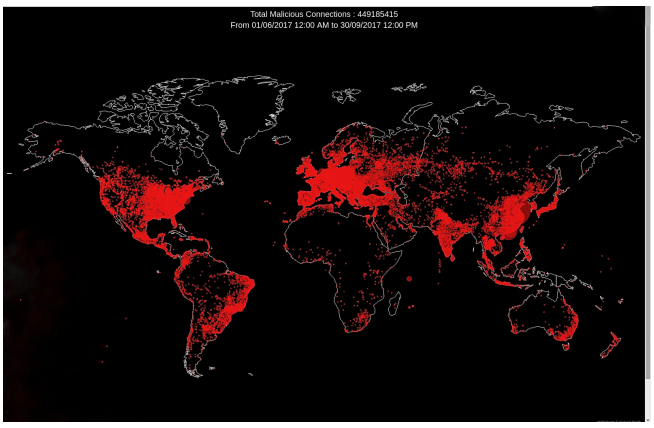


Fig. 8. Total malicious connections for a period of four months: 449 million packets

To make all these a little more concrete, consider an example of configuring a Spark application to utilize most of the cluster resources; The proposed framework has a cluster with 32 nodes running NodeManagers, each equipped with 8 cores and 32GB of memory. To acheive the best performance, the NodeManager capacities, *yarn.nodemanager.resource.cpu-vcores* and *yarn.nodemanager.resource.memory-mb* should probably be set to 7 (cores) and 31GB * 1024 = 31744 (megabytes), respectively.

Consider the case *yarn.nodemanager.resource.cpu-vcores* with a value 8 and *yarn.nodemanager.resource.memory-mb* with value 32768MB (32GB * 1024). One must always avoid allocating 100% of the available resources to YARN containers, since a node requires some of the resources to run Hadoop and Operating System (OS) daemons. In this case, one has to leave a core and a gigabyte for these system processes. Hence, set the parameters *yarn.nodemanager.resource.cpu-vcores* and *yarn.nodemanager.resource.memory-mb* to 7 and 31744 respectively.

Running *yarn.nodemanager.resource.cpu-vcores* with a value *5* and having just enough memory required to run a single task will throw away the benefits that could come from running several tasks in a single Java Virtual Machine (JVM).

The results have been summarized in Tables I-IX. The same results have been visualized in Fig. 9-11.

Tables I-IX represent the time taken to process data in batches of one month (24GB), two months (42GB) and four months (85GB). This data has been analyzed by making use of 7 CPU-cores per node in the first set of experiments, then in the next set using 5 CPU-cores per node and finally all the 8 CPU-cores are used to process the data.

The difference in run-time between the number of CPU-cores utilized is mainly observed when the experiments are run on four months of data using a single node. Consider the following 3 cases:

- If we use 5 cores of CPU from a node, it takes 1346 seconds to process 85GB of data.
- If we use 8 cores of CPU from a node, it takes 1090 seconds to process 85GB of data.
- If we use 7 cores of CPU from a node, it takes 971 seconds to process 85GB of data.

In the first case, since some of the CPU cores are not at all utilized, it takes more time to process the data. Whereas in the second case, we are not leaving any core of CPU for Operating System and YARN daemons and all the resources are utilized only for Spark process. Hence it takes a bit more time to process the same 85GB of data (when compared to last case). Finally, leaving one core of CPU for OS and YARN daemons and utilizing the remaining cores to process the data will yield the best run-time performance.

TABLE I. TIME TAKEN FOR PROCESSING ONE MONTH DATA (24GB), USING 7 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (7 cores)	281
2 nodes (14 cores)	171
4 nodes (28 cores)	103
8 nodes (56 cores)	62
16 nodes (112 cores)	37
32 nodes (224 cores)	22

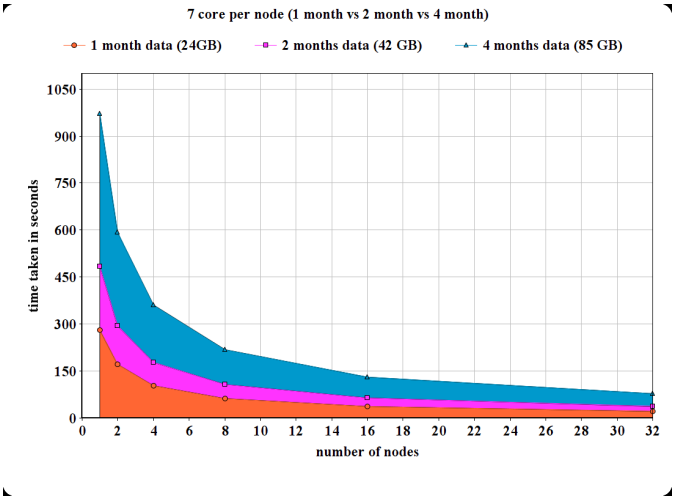


Fig. 9. 7 cores per node

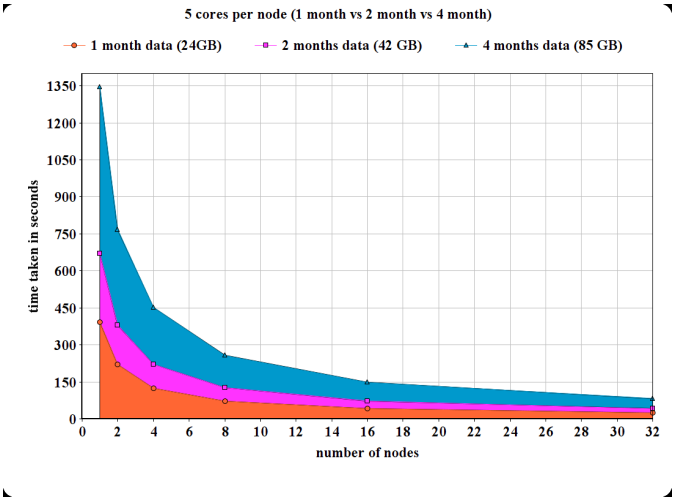


Fig. 10. 5 cores per node

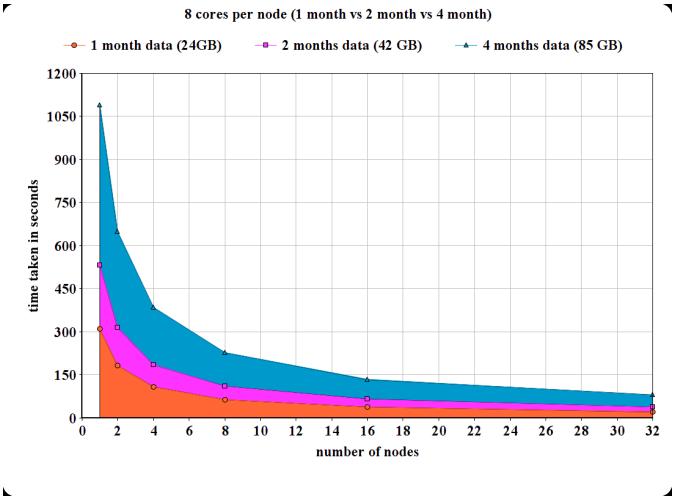


Fig. 11. 8 cores per node

TABLE II. TIME TAKEN FOR PROCESSING OF TWO MONTHS DATA (42GB), USING 7 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (7 cores)	483
2 nodes (14 cores)	294
4 nodes (28 cores)	178
8 nodes (56 cores)	107
16 nodes (112 cores)	64
32 nodes (224 cores)	38

TABLE III. TIME TAKEN FOR PROCESSING OF FOUR MONTHS DATA (85GB), USING 7 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (7 cores)	971
2 nodes (14 cores)	592
4 nodes (28 cores)	360
8 nodes (56 cores)	218
16 nodes (112 cores)	131
32 nodes (224 cores)	**78**

TABLE IV. TIME TAKEN FOR PROCESSING ONE MONTH DATA (24GB), USING 5 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (5 cores)	393
2 nodes (10 cores)	220
4 nodes (20 cores)	124
8 nodes (40 cores)	72
16 nodes (80 cores)	42
32 nodes (160 cores)	24

TABLE V. TIME TAKEN FOR PROCESSING OF TWO MONTHS DATA (42GB), USING 5 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (5 cores)	670
2 nodes (10 cores)	380
4 nodes (20 cores)	221
8 nodes (40 cores)	127
16 nodes (80 cores)	73
32 nodes (160 cores)	41

TABLE VI. TIME TAKEN FOR PROCESSING OF FOUR MONTHS DATA (85GB), USING 5 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (5 cores)	1346
2 nodes (10 cores)	768
4 nodes (20 cores)	452
8 nodes (40 cores)	257
16 nodes (80 cores)	148
32 nodes (160 cores)	83

TABLE VII. TIME TAKEN FOR PROCESSING ONE MONTH DATA (24GB), USING 8 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (8 cores)	310
2 nodes (16 cores)	183
4 nodes (32 cores)	108
8 nodes (64 cores)	64
16 nodes (128 cores)	38
32 nodes (256 cores)	22

When the same job was made to run on a traditional system without SPARK, having configurations of Intel Xeon E3-1230 v3 processor @ 3.30 GHz, with 8 MB cache, it took 29 minutes

TABLE VIII. TIME TAKEN FOR PROCESSING OF TWO MONTHS DATA (42GB), USING 8 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (8 cores)	531
2 nodes (16 cores)	315
4 nodes (32 cores)	186
8 nodes (64 cores)	110
16 nodes (128 cores)	65
32 nodes (256 cores)	38

TABLE IX. TIME TAKEN FOR PROCESSING OF FOUR MONTHS DATA (85GB), USING 8 CPU-CORES PER NODE

No. of nodes (cores)	Time taken in seconds
1 node (8 cores)	1090
2 nodes (16 cores)	648
4 nodes (32 cores)	385
8 nodes (64 cores)	228
16 nodes (128 cores)	135
32 nodes (256 cores)	80

and 8 seconds to process 85 GB of data.

VI. CONCLUSIONS AND FUTURE WORK

The nature of processing PCAP data is different from processing other data formats. This requires tuning resource allocation based on the number of cores, number of executors and the amount of memory to be allotted for faster data processing. This issue has been addressed in the proposed work.

In the literature, small amount of work is done in the field of processing network trace data using SPARK technology. 85GB of data has been analyzed in just 78 seconds using 32 node (256 cores) SPARK cluster, which would otherwise take around 30 minutes in traditional processing systems. Best results were achieved by allotting 7 (out of 8) cores of CPU per node and 31744MB of memory (leaving one GB of memory for OS and YARN daemons).

Spark avoids the file system to a greater extent. It retains most of the data distributed in memory, across multiple phases of the same job. RDDs hide the details of fault-tolerance and distribution for huge collections of items. RDDs apply the same operation of map, filter and join to many data items. RDDs are computed lazily for the first time they are used, so that it can later pipeline the transformations; In the proposed work, the focus is on the cases where aggregate memory can hold the entire input RDD within main memory, so that the queries (job) submitted can be executed faster.

A. Future Work

In all experiments, enough memory resources has been allocated so that RDDs can fit into the main memory. Thorough investigation of storage alternatives, especially when RDDs cannot fit in the main memory, are out of scope and left for future work.

As a future work, one can utilize the techniques to analyze the large network packet data in near real-time and apply some machine learning algorithms to develop near real-time automatic detection systems against network attacks.

REFERENCES

[1] Mashooque Memon, Safeeullah Soomro, Awais Jumani, and Muneer Kartio. Big data analytics and its applications. *Annals of Emerging Technologies in Computing*, 1, 10 2017.

[2] A. A. Cárdenas, P. K. Manadhata, and S. P. Rajan. Big data analytics for security. *IEEE Security Privacy*, 11(6):74–76, Nov 2013.

[3] Apache spark. https://spark.apache.org/.

[4] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4):431 – 448, 2018.

[5] Peter P. Nghiem and Silvia M. Figueira. Towards efficient resource provisioning in mapreduce. *Journal of Parallel and Distributed Computing*, 95:29 – 41, 2016. Special Issue on Energy Efficient Multi-Core and Many-Core Systems, Part I.

[6] Apache hadoop [Online]. http://hadoop.apache.org/.

[7] B. Akil, Y. Zhou, and U. Röhm. On the usability of hadoop mapreduce, apache spark apache flink for data science. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 303–310, Dec 2017.

[8] R. Tous, A. Gounaris, C. Tripiana, J. Torres, S. Girona, E. Ayguadé, J. Labarta, Y. Becerra, D. Carrera, and M. Valero. Spark deployment and performance evaluation on the marenostrum supercomputer. In *2015 IEEE International Conference on Big Data (Big Data)*, Oct 2015.

[9] Y. R. Bachupally, X. Yuan, and K. Roy. Network security analysis using big data technology. In *SoutheastCon 2016*, pages 1–4, March 2016.

[10] Wireshark. https://www.wireshark.org/.

[11] D. Mistry, P. Modi, K. Deokule, A. Patel, H. Patki, and O. Abuzaghleh. Network traffic measurement and analysis. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pages 1–7, April 2016.

[12] R. C. Maheshwar and D. Haritha. Survey on high performance analytics of bigdata with apache spark. In *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pages 721–725, May 2016.

[13] Keisuke Kato and Vitaly Klyuev. Hadoop environment for the analysis of large network packets. In *Proceedings of the 2nd International Conference on Applications in Information Technology*, pages 56–59.

[14] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[15] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 2–2, Berkeley, CA, USA, 2012. USENIX Association.

[16] M. Čermák, T. Jirsík, and M. Laštovička. Real-time analysis of netflow data for generating network traffic statistics using apache spark. In *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, pages 1019–1020, April 2016.

[17] Prakasam Kannan. Beyond hadoop mapreduce apache tez and apache spark.

[18] S. Chaisiri, B. S. Lee, and D. Niyato. Optimization of resource provisioning cost in cloud computing. *IEEE Transactions on Services Computing*, 5(2):164–177, April 2012.

[19] Parveen Patel, Deepak Bansal, Lihua Yuan, Ashwin Murthy, Albert Greenberg, David A. Maltz, Randy Kern, Hemant Kumar, Marios Zikos, Hongyu Wu, Changhoon Kim, and Naveen Karri. Ananta: Cloud scale load balancing. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, pages 207–218, New York, NY, USA, 2013. ACM.

[20] Y. Tan, H. Nguyen, Z. Shen, X. Gu, C. Venkatramani, and D. Rajan. Prepare: Predictive performance anomaly prevention for virtualized cloud systems. In *2012 IEEE 32nd International Conference on Distributed Computing Systems*, pages 285–294, June 2012.

[21] A. M. Karimi, Q. Niyaz, Weiqing Sun, A. Y. Javaid, and V. K. Devabhaktuni. Distributed network traffic feature extraction for a real-time ids. In *2016 IEEE International Conference on Electro Information Technology (EIT)*, pages 0522–0526, May 2016.

[22] The caida ucsd ddos attack 2007 dataset, 2007. https://www.caida.org/data/passive/ddos-20070804_dataset.xml.

[23] M. Kulariya, P. Saraf, R. Ranjan, and G. P. Gupta. Performance analysis of network intrusion detection schemes using apache spark. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 1973–1977, April 2016.

[24] S. Michael, A. Thota, and R. Henschel. Hpchadoop: A framework to run hadoop on cray x-series supercomputers. *Cray USer Group (CUG)*, 2014.

[25] K. Wang and M. M. H. Khan. Performance prediction for apache spark platform. In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pages 166–173, Aug 2015.

Rich Style Embedding for Intrinsic Plagiarism Detection

Oumaima Hourrane[1], El Habib Benlahmer[2]
Laboratory Information Technology and Modeling,
Faculty of Science Ben Msik,
Hassan II University of Casablanca, Morocco

***Abstract*—Stylometry plays an important role in the intrinsic plagiarism detection, where the goal is to identify potential plagiarism by analyzing a document involving undeclared changes in writing style. The purpose of this paper is to study the interaction between syntactic structures, attention mechanism, and contextualized word embeddings, as well as their effectiveness on plagiarism detection. Accordingly, we propose a new style embedding that combines syntactic trees and the pre-trained Multi-Task Deep Neural Network (MT-DNN). Additionally, we use attention mechanisms to sum the embeddings, thereby experimenting with both a Bidirectional Long Short-Term Memory (BiLSTM) and a Convolutional Neural Network (CNN) max-pooling for sentences encoding. Our model is evaluated on two sub-task; style change detection and style breach detection, and compared with two baseline detectors based on classic stylometric features.**

***Keywords*—Plagiarism detection; style embedding; deep neural network; stylometry; syntactic trees**

I. Introduction

The enormous availability of textual data and documents via the Internet (web pages, databases, digital documents, etc.) makes it more straightforward for unauthorized copying and stealing of intellectual properties of others. To counter this phenomenon, there are two main approaches to detect plagiarism that have been studied and implemented; extrinsic and intrinsic plagiarism detection [3]. The former method consists of comparing suspicious passages or documents against a range of external sources, while the latter method focuses on using local analysis to detect suspicious passages within a document by comparing their writing styles without using any referenced document.

The act of collecting reference set of documents for the extrinsic plagiarism detection is very troublesome from the computational time point of view. In addition, a source of a plagiarized document may not always be available. Therefore, researchers in recent works had addressed these limits by the intrinsic plagiarism detection, which has been introduced first by Meyer Zu Eissen and Stei [10]. The idea is that one would expect that a plagiarized passage from a document is stylistically inconsistent from the non-plagiarize passages, without comparing a suspicious document to the potential sources. Such stylistic inconsistencies can be detected by quantifying certain stylometric features, such as character n-gram profile [11], Parse trees [12], word n-gram and word frequency [13].

Since the major advances in Deep learning-based pre-trained word embeddings [14], [15], [16], It became fundamental to many Natural Language Processing tasks to use these representations as input features to other models [5], [6]. Furthermore, the emergence of the contextualized language representations [17], [20], [19], [18], [4] along with Multi-task Learning concept [21], had given birth to more refinded language representations, the Multi-Task Deep Neural Network (MT-DNN) [1]. By training this latter on various tasks such as Natural Language Inference, Parsing task, and Neural Machine Translation, several semantic and syntactic properties have been learned.

In this paper, our goal is to leverage the properties learned from the pre-trained MT-DNN, and to make our style representation richer by adding along with the MT-DNN another type of embeddings based on syntactic trees, where each word is represented by POS tags from the leaf to the root for the constituency tree, and the dependencies links for the dependency tree. Then, the resulted sequence are served as input to a Bi-LSTM architecture to get the final syntactic embeddings of words. To combine the strengths of the three type of embeddings we employ a weighted sum based on a self-attention mechanism [22]. Finally, we train a fully connected layer followed by a softmax on top of the sequence encoding to classify each sequence. The latter can be a document for the style change detection sub-task, and a sentence for the style breach detection sub-task, more details can be found in Section III.

The rest of this paper is organized as follows: In Section 2, we describe our architecture with more details. In Section 3 we outline our experimental settings and compare our result with two baseline methods. In Section 4, we introduce some related works on intrinsic plagiarism detection. Finally, we summarize our work and future directions.

II. Related Work

In this section, we report previous works that have been done for the intrinsic plagiarism detection task, and other related tasks, such as Style Breach detection [9], and Style Change Detection [7].

By considering the intrinsic plagiarism detection task as a one-class classification problem, [24] construct a wide range of stylometric features and apply a density method to separate outliers from target class, where they assumed that the outliers to be the plagiarized passages from the source documents. They as well employed the meta-learning approach of Koppel and Schler [25] to post-process unreliable stylometric analysis results.

To compute writing style differences, [26] proposed an intrinsic algorithm that describes a new variant by using stop words with TF-weighting as input feature, after the document segmentation into equal size chunks and features representation, they compare each chunk representation with that of the whole suspicious document.

To to find style change and irregularities, [12] analyzed syntactic information of authors focusing on sentence constructions. Namely, they extract POS-sequences representations, next they construct a distance matrix of every distinct pair sentences by applying sequence alignment algorithm. Finally, they employed Gaussian normal distribution function over the mean distances, all thresholds and parameters are optimized by implementing genetic algorithms.

The work of [27] consists of running various step for the intrinsic plagiarism detection. Beginning by chunking to suspicious document into equal size overlapping windows, then they represent each chunk by the relative frequencies of a predetermined set of high-frequency character trigrams. Next, they measure the distance between consecutive windows using is a symmetric adaptation of the normalized distance. Finally, they employed an algorithm for outliers detection based on Principal Components Analysis.

There are two approaches that deal with the style breach detection instead of detecting suspicious passages to be plagiarized, The first approach [28] assume that each sentence vector depends on the previous and next sentence vectors, so they proposed an architecture based on mapping sentences by using a pre-trained encoder-decoder as an embedding. The resulting vectors are used then for the outliers detection step. The second approach is an unsupervised analysis [8], it is focused on retrieving various stylistic features along with some new features naming common English word frequencies.

These following works propose models that answer the question of if a given document had a style change within a document; [30] use character-based Convolutional Neural Networks (CNN) which can be applied to any language. [31] combine TF.IDF representation of the documents with other stylistic features, they then use an ensemble of diverse models including SVM, Random Forest, AdaBoost, MLP and LightGBM. [32] employ two parallel attention networks, feeding the model with the hierarchical structure of the language using a pre-trained statistical parser.

Finally, [33] propose a feature discretization method based on two-step cluster for Naive Bayes. Namely, they used the TF-IDF and query language model as a discrete feature and False Positive/False Negative (FP/FN) threshold to improve the accuracy.

III. Methodology

Our work consist of using different kind of embeddings for the intrinsic plagiarism detection problem. We apply our architecture on two sub-tasks. The first, is the style change detection, where we verify if an input document contains style changes, therefore, it contains sections written by other authors. The second is the style breach detection, where we detect the intrusive sections that are stylistically deviant from the main writing style.

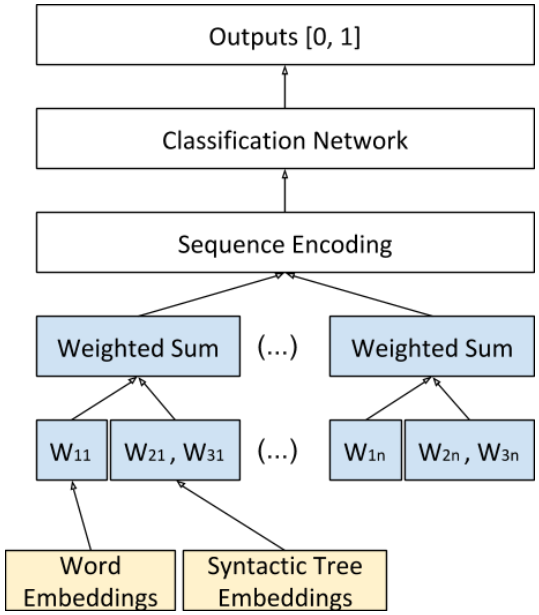


Fig. 1. General Architecture for Intrinsic Plagiarism Detection.

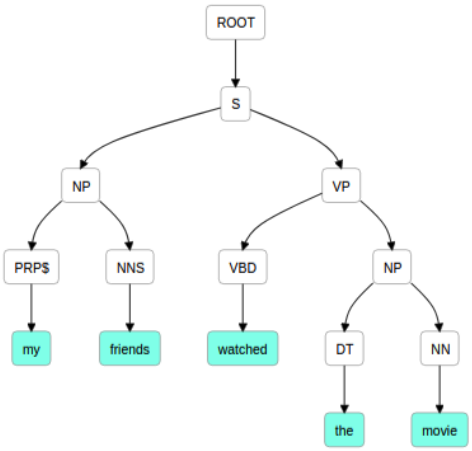


Fig. 2. An example of the constituency tree of 'My friends watched the movie'.

The general framework of our model is illustrated in Fig. 1. Here the input of the model is the embedding of a given sequence while the output indicate respectively if a sequence is plagiarized or not. We describe in the next sections more details about each elements of our architecture.

A. Style Embeddings (SE)

1) Constituency Trees Embedding (CTE): We extract constituency trees for each sentence is a document using the Stanford CoreNLP [29]. We choose this kind of syntactic tree to describe the grammatical aspect of a given text, as it defines the way to hierarchically construct a sentence from words based on constituency relations. Fig. 2 shows an example of a constituency tree.

For each word in a sentence, We define constituency tree feature as a path starting from the root to the corresponding

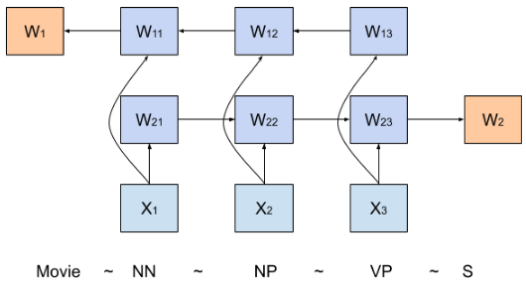


Fig. 3. Constituency tree embedding for a given word using a Bi-directional LSTM.

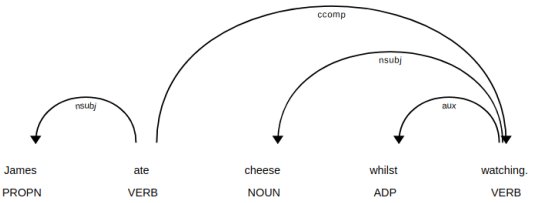


Fig. 4. An example of the dependency tree of 'James ate cheese whilst watching.'.

leaf, for example, the word 'movie' in our example, can be represented by a syntactic sequence (NN, NP, VP).

As it has been proven effective in [34], we construct syntactic embedding by using a Bi-directional LSTM to represent a constituency tree feature of a given word as indicated in Fig. 3. the input is index vectors for syntactic tags initialized with a multivariate normal distribution. Finally, each word is represented with the concatenation of W_1 and W_2; the output vectors of the forward and backward LSTMs, respectively.

2) Dependency Trees Embedding (DTE): The dependency tree is a syntactic tree constructed by dependency grammars, namely, in a given sentence, words are connected to each other by direct links called dependencies.

The advantage of this structure is its ability to deal with language that is morphologically rich and has a free word order, which represents only the information that is necessary for the sentence. Fig. 4 illustrates an example of a dependency tree.

Relations among the words is represented above the sentence with directed and tagged arcs from root to leaves. For each word we encode its dependency sub-tree with its dependent nodes ordered by their positions in the original sentence, using Bi-directional LSTM as indicated in Fig. 5. Similar to the constituency tree, each word is represented with the concatenation of W_1 and W_2; the output vectors of the forward and backward LSTMs respectively.

3) Contextualized word embedding: The second input that we fed into our model is the embedding information of words we denote it an embedding of a single word as W_3. We choose to use the pre-trained MT-DNN [1], which has created new state-of-the-art results across many popular NLP benchmarks. The particularity of this model is that is a one-to-many multi-tasking learning framework that computes the loss across different tasks and applies them at the same time. Moreover,

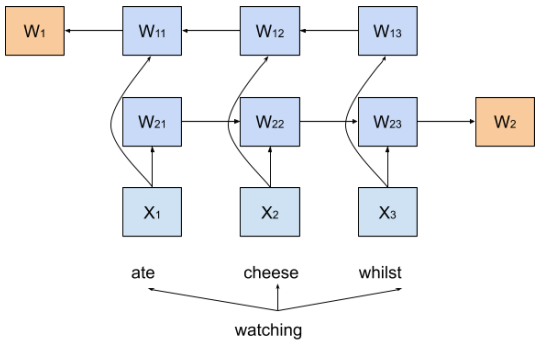


Fig. 5. Dependency tree embedding for a given word using a Bi-directional LSTM.

several stylistic properties have been learned in this model by training it on a parsing task, Natural Language Inference and Neural Machine Translating tasks.

4) Sequence Encoding: For a sequence of n words $\{p_j\}_{j=1}^{n}$, we denote the three embeddings we get from the previous step as $\{W_{i,j}\}_{j=1}^{n}$ and $i = 1, 2, 3$. As our aim is to apply the attention weights within the sequence, we then combine these embedding by taking their weighted sum.

$$W'_j = \sum_{i=1}^{3} \alpha_{i,j} W_{i,j}$$

Where $\alpha_{i,j}$ are scalar weights from a self-attention mechanism [22]: $\alpha_{i,j} = \phi(a.W_{i,j} + b)$, where a and b are learned parameters and ϕ is a Softmax function.

Next, we experiment with two encoders CNN and a standard bidirectional LSTM (BiLSTM-Max) with max-pooling to obtain a global sequence embedding.

For the CNN encoder, a convolution operation involves the application of a filter k to a window of h words to produce a new feature. For instance, $c_{j,i}$ is the i^{th} feature generated from a window of the word embedding $W'_{j:j+h-1}$

$$c_{j,i} = f(a_i.W'_{j:j+h-1} + b_i)$$

Where a_i and b_i respectively are the weight and bias f the j^{th} filter. Finally, a max-pooling operation is applied over these features to get the final sequence representation E^{seq}_{CNN}:

$$E^{seq}_{CNN} = max_{row}(c_{j,i})$$

Secondly, the BiLSTM-Max computes for each direction two sets of m hidden states as follows:

$$\overrightarrow{h_i} = \overrightarrow{LSTM_i}(W_1, W_2, ..., W_i)$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM_i}(W_i, W_{i+1}, ..., W_m)$$

In order to get the final hidden states, for each time-step the hidden states are afterward concatenated, then a max-pooling operation is applied over their components to get the final sequence representation: $h = max([\overrightarrow{h_i}, \overleftarrow{h_i}]_{j=1,2,..,m})$

B. Style Change Detection

This task deals with a particular question, whether or not a document has multiple authors. Certainly, the changes of authorship have to be determined by capturing changes in writing styles, and the problem can be taken by applying a binary classification over the whole document. In that case, each document is represented as a fixed-sized vector of the resulted style embedding, this vector is next fed into a fully connected layer followed by a Softmax layer to predict the existence of style changes. we kept the same dimension of the fully connected layer as the sequence embedding from the previous step. The outputs of the network are 0 or 1. A binary cross-entropy loss function is used to train the entire model. This later is optimized using Adam optimizer [23].

C. Style Breach Detection

This task consists of finding sections within a document where the authorship changes. This problem can simply be seen as a text segmentation problem based on the writing style. In our work, we assume that style breaches may occur on sentences ending so we apply our style-embedding on the sentence level. Then, we applied a sentence outliers detection as commonly used in Intrinsic Plagiarism Analysis [24], where we determine if an input sentence is plagiarized or not. same as in the style change detection sub-task, we fed the resulted style-embedding of the input sentences into a fully connected layer followed by a Softmax layer, then we train and optimize the model by using a binary cross-entropy loss function and Adam optimizer, respectively.

IV. Experiment

A. Datasets

We evaluate the performance of the proposed approach for the style breach detection sub-task on PAN-PC-11 corpus[1], which is addressed for the intrinsic Plagiarism detection. It contains artificially plagiarised passages for each suspicious document and mentions the offsets of plagiarised and non-plagiarized parts. The next sections describe more details about the experiment.

The second corpus we used in the style change detection sub-task is PAN 2018 corpus[2] based on user posts from QA network Stack Exchange. It contains 2980 training problems, 1492 validation problems and 1352 test problems, where for each subset the amount of documents combining styles changes is equal to the number of documents containing no changes.

B. Baseline Models

Concerning the style change detection sub-task, we chose as a baseline model, the method of Zlatkova [2] that is state-of-the-art on PAN 2018 Competition [7]. The authors used stacking ensemble architecture. After the text pre-processing, they chunked the text into three equal fragments and used several lexical and syntactical features, and trained four different classifiers on each fragment, they combined these models with TF-IDF based gradient boosting model and fed them into a Logistic regression meta-classifier to produce the final result.

For the style breach detection sub-task, we chose as a baseline model the approach of Khan [8] which is state-of-the-art on PAN 2017 Competition [9]. The authors used as stylometric feature most frequent POS tags and words, and other dictionaries. Next, they compute the similarity score between each two adjacent sliding windows which is then compared to a predefined threshold in order to decide the style breach between two sentences if it exists.

C. Experimental Settings

Syntactic information related to constituency trees and dependency trees were extracted from Stanford CoreNLP[3]. For a given syntactic tree representation of a single word we chose a window size of max 10 tags in the constituency tree vector and 10-nearest dependents in the dependency tree vector, in order to prevent excessive usage of the memory, while benefiting the performance of the model [22].

As described in Section III-A4, we use a pre-trained contextualize word embedding MT-DNN [1] to serve as part of the input into the model. Moreover, all the embedding types we used in the input have a dimension of 1000.

For the classification network, as indicated in Section III-B. We train a fully connected layer followed by a Softmax on top of the sequence embedding after the max pooling operation. A binary cross-entropy is used as loss function. The initial learning rate is set to $10^{-}4$, dropout to 0.2 and we use Adam for optimization.

D. Results

The overall performance results are depicted in Table I. Our purpose is to apply the combination of different embeddings, though, we report the accuracy of each embeddings to see which one contributes in the final performance, we also report the difference between using CNN max pooling operation for sequence encoding and BiLSTM-Max architecture.

Concerning the style change detection sub-task. We could see that the SE-Combined-BiLSTM achieved the best result with an accuracy of nearly 88%, which has shown that using BiLSTML max pooling operation for sequence encoding achieve good performance for this problem. The same thing on the style breach detection task, the SE-Combined-BiLSTM outperform the other models with an accuracy of 80% Table II, which suggest that using both BiLSTM for sequence embedding is most effective to extract the plagiarized passages that are stylistically different from the original text.

In both sub-tasks, we could see that The SE-DTE Dependency trees embedding outperform the other representations in term of accuracy, followed by SE-CTE Constituency tree embedding, then the SE-MT-DNN Contextualized word embedding. And when combining the three embedding types, the model achieved higher accuracy, providing the evidence that combining a contextualize word embedding with syntactic information can accurately detects the style change with documents. Therefore detects plagiarism intrinsically. Moreover, we

[1]https://webis.de/data/pan-pc-11.html

[2]https://pan.webis.de/clef18/pan18-web/author-identification.html#style-change-detection

[3]https://stanfordnlp.github.io/CoreNLP/

TABLE I. PERFORMANCES OF ALL MODELS ON THE STYLE CHANGE DETECTION SUB-TASK

Models	Accuracy
SE-CTE	0.70
SE-DTE	0.76
SE-MT-DNN	0.69
SE-Combined-CNN	**0.78**
SE-Combined-BiLSTM	0.88
[2]	0.75

TABLE II. PERFORMANCES OF ALL MODELS ON THE STYLE BREACH DETECTION SUB-TASK

Models	Accuracy
SE-CTE	0.66
SE-DTE	0.73
SE-MT-DNN	0.70
SE-Combined-BiLSTM	**0.80**
SE-Combined-CNN	0.76
[8]	0.59

found that BiLSTM-Max for the sequence encoding operation has better performance than their CNN counterparts in both sub tasks, which suggests that BiLSTM can more effectively preserve the syntactic and stylometric information.

V. CONCLUSION AND FUTURE WORKS

In this work, we proposed an architecture for the intrinsic plagiarism detection, based on a rich style representation detailed in Section III. We applied our approach on two sub-task; Style change detection and style breach detection, and evaluate it on PAN-18 and PAN-PC-11 corpus, and compare it with baseline detectors. The performance of the combined style embedding is promising, providing the evidence that combining a contextualize word embedding with syntactic information can accurately detects plagiarism intrinsically within a document. Roughly speaking, We subsequently show how our style embedding approach can be used to shed new light on the usage of structural embedding along with other state-of-the-art word embeddings in stylometry.

There are many future directions to improve our intrinsic plagiarism detector including but not limited to: In future work, it would be interesting to apply our idea to different tasks, such as Authorship Attribution and Extrinsic Plagiarism Detection, in order to explore what kinds of embeddings are most useful for each plagiarism detection tasks. It would also be interesting to further exploit more stylometric features, such as word frequencies and word/character n-grams in our combined style representation, and to apply them on an unsupervised plagiarism detection task that focus on separating the original sequences from the plagiarized ones. In addition, it would be interesting to examine the effect of adding the attention weight within and after the sequence encoding, the difference between the embedding concatenation and the weighted variant of our method, and the attention weights change during the training.

REFERENCES

[1] Liu X, He P, Chen W, Gao J. *Multi-task deep neural networks for natural language understanding.* arXiv preprint arXiv:1901.11504. 2019 Jan 31.

[2] Zlatkova D, Kopev D, Mitov K, Atanasov A, Hardalov M, Koychev I, Nakov P. *An ensemble-rich multi-aspect approach for robust style change detection.* CLEF 2018 Working Nots of CLEF. 2018.

[3] Hourrane, Oumaima, and El Habib Benlahmar. "Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval." Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. ACM, 2017.

[4] Hourrane, Oumaima, et al. "Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers." International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018.

[5] Mifrah, Sara, and El Habib Ben Lahmar. "Semantico-automatic Evaluation of Scientific Papers: State of the Art." Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. ACM, 2017.

[6] Omar Zahour, El Habib Benlahmar, Ahmed Eddaoui and Oumaima Hourrane, "Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network" International Journal of Advanced Computer Science and Applications(IJACSA), 10(10), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101072

[7] Kestemont, Mike, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. *Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection.* In Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al., pp. 1-25. 2018.

[8] Khan, Jamal Ahmad. *Style Breach Detection: An Unsupervised Detection Model.* In CLEF (Working Notes). 2017.

[9] Tschuggnall, Michael, Efstthios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. *Overview of the author identification task at PAN-2017: style breach detection and author clustering.* In Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al., pp. 1-22. 2017.

[10] Zu Eissen, Sven Meyer, and Benno Stein. *Intrinsic plagiarism detection.* European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2006.

[11] Stamatatos E. *A survey of modern authorship attribution methods.* Journal of the American Society for information Science and Technology. 2009 Mar;60(3):538-56.

[12] Tschuggnall, Michael, and Günther Specht. *Plag-inn: Intrinsic plagiarism detection using grammar trees.* International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2012.

[13] Oberreuter G, VeláSquez JD. *Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style.* Expert Systems with Applications. 2013 Jul 1;40(9):3756-63.

[14] Mikolov T, Chen K, Corrado G, Dean J. *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781. 2013 Jan 16.

[15] Pennington, Jeffrey, Richard Socher, and Christopher Manning. *Glove: Global vectors for word representation.* Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[16] Joulin A, Grave E, Bojanowski P, Mikolov T. *Bag of tricks for efficient text classification.* arXiv preprint arXiv:1607.01759. 2016 Jul 6.

[17] Kiros, Ryan, et al. *Skip-thought vectors.* Advances in neural information processing systems. 2015.

[18] Devlin, Jacob, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805 (2018).

[19] Howard, Jeremy, and Sebastian Ruder. *Universal language model fine-tuning for text classification.* arXiv preprint arXiv:1801.06146 (2018).

[20] Peters, Matthew E., et al. *Deep contextualized word representations.* arXiv preprint arXiv:1802.05365 (2018).

[21] Zhang, Yu, and Qiang Yang. *A survey on multi-task learning.* arXiv preprint arXiv:1707.08114 (2017).

[22] Lin, Zhouhan, et al. *A structured self-attentive sentence embedding.* arXiv preprint arXiv:1703.03130 (2017).

[23] Kingma, Diederik P., and Jimmy Ba. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980 (2014).

[24] Stein, Benno, Nedim Lipka, and Peter Prettenhofer. *Intrinsic plagiarism analysis.* Language Resources and Evaluation 45.1 (2011): 63-82.

[25] Koppel, Moshe, and Jonathan Schler. *Authorship verification as a one-class classification problem.* Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[26] Oberreuter G, L'Huillier G, Ríos SA, Velásquez JD. *Approaches for intrinsic and external plagiarism detection.* Proceedings of the PAN. 2011.

[27] Kestemont M, Luyckx K, Daelemans W. *Intrinsic plagiarism detection using character trigram distance scores.* Proceedings of the PAN. 2011.

[28] Safin, Kamil, and Rita Kuznetsova. *Style Breach Detection with Neural Sentence Embeddings.* CLEF (Working Notes). 2017.

[29] Manning, Christopher, et al. *The Stanford CoreNLP natural language processing toolkit.* Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014.

[30] Schaetti, Nils. *Character-based Convolutional Neural Network for Style Change Detection.* Training 2980.1490: 1490.

[31] Zlatkova D, Kopev D, Mitov K, Atanasov A, Hardalov M, Koychev I, Nakov P. *An ensemble-rich multi-aspect approach for robust style change detection.* CLEF 2018 Working Nots of CLEF. 2018.

[32] Hosseinia, Marjan, and Arjun Mukherjee. *A Parallel Hierarchical Attention Network for Style Change Detection.*

[33] Wijaya, Adi, and Romi Satria Wahono. *Two-Step Cluster based Feature Discretization of Naïve Bayes for Outlier Detection in Intrinsic Plagiarism Detection.* Journal of Intelligent Systems 1.1 (2015): 1-8.

[34] Liu, Rui, et al. *Structural embedding of syntactic trees for machine comprehension.* arXiv preprint arXiv:1703.00572 (2017).

UAV Control Architecture: Review

IDALENE Asmaa[1], BOUKHDIR Khalid[2], and MEDROMI Hicham[3]

[1,2,3]Research and Engineering Laboratory LRI, National School of Electricity and Mechanics Casablanca, Morocco
[1]Pluridisciplinary Laboratory of Research and Innovation (LPRI), EMSI Casablanca, Morocco

***Abstract*—Since civil Unmanned Aerial Vehicles (UAVs) are expected to perform a wide rang of mission, the subject of designing an efficient control architecture for autonomous UAV is a very challenging problem. Several contributions had been done in order to implement an autonomous UAV. The key challenge of all these contributions is to develop the global strategy. Robotic control approaches could be classified into six categories: Deliberative, Reactive, Hybrid, Behavior, Hybrid Behavior and subsumption approach. In this paper, we will review the existing control architectures to extract the main features of civil UAVs. The definition, advantage and drawback of each architecture will be highlighted to finally provide a comparative study of the mentioned control approaches.**

Keywords*—*Unmanned Aerial Vehicle; control architecture; deliberative approach; reactive approach; hybrid approach; behavior approach; hybrid behavior approach; subsumption approach

I. Introduction

The Unmanned aerial vehicle or UAV is a particular mobile robot that operates in dynamic tri-dimensional space [1] [2], in order to meet some goals or to achieve a given mission by executing a specific control strategy [3] [2] . Typically, autonomous UAV is a complex system [4] that adapts to different circumstances without human intervention [5] [6]. Autonomous UAV needs continually to sense and perceive its surrounding environment, to decide what to do and to execute the appropriate commands in its operating environment [7] [8].

UAV control architecture represents a global strategy and specific algorithms used to define how the UAV will perform its sensing operations and its perceiving capabilities [7] [9], how it will decide its task to meet and how it will act in specific environment conditions [10]. The control architecture affects the robot's capabilities: the processing time, the need to completely know the operating environment, the ability to deal with a large variety of missions, the capacity to achieve goals in the presence of uncertainties and the autonomy level [11], [12].

Several contributions have been done in term of control architectures [10] [13] [14] [15] [16]. beyond the known control architectures, we found the deliberative approach that has been implemented based on the sense-plan-act paradigm [17] [18]. The reactive architecture has been designed as a set of condition-action pairs [19] [20]. While the hybrid approach has been made as a combination of deliberative and reactive capabilities [21], [22]. Moreover, the behavior approach has been defined as a collection of behavior sequences that performs each of them a specific goal [23] . And finally we found the subsumption approach that has been structured as a set of ordered competence levels, each of which provides a specific capability [24], [25]. The key challenge for all these contributions is to develop an autonomous control system that can make suitable decisions, perform many tasks, plan a feasible path and avoid static and dynamic obstacles [12].

This paper is organized as follows: in the second section the existing control architectures for mobile robot will be reviewed, a comparative study of the state of the art will be proposed in the third section and finally a brief conclusion of this work will be presented in section five.

II. Control architecture: Review

Various control architectures have been designed in order to develop high performance systems [10], [13], [14], [15], [16]. Each of them offers new concepts in attempt to build an autonomous robot. In the present section we review in detail the existing control architectures:

A. The Deliberative based Control Architecture

The deliberative control architecture is a top-down approach [13]. In order to meet a given mission, the deliberative approach reasons about goals and constraints to finally execute low-level commands. Basically, it comprises three generic sequential functionalities: sensing, planning and acting modules [17]. For each mission's goal (see Fig. 1), the sensing module senses the robot's surrounding environment in order to update a given world model. To reach the mission goal, the planning module generates a valid task plan considering the robot's constraints. At last, the acting module transforms the task plan into robot low-level commands then executes these commands. The robot then repeats these sequential functionalities until it reaches its mission goals [26].

This architecture presents a serious source of weakness in certain cases, we mention some of them [27], [28]:

- If one of the modules fails, the whole architecture will break.
- It is ineffective in a dynamic or uncertain environment.
- It requires high performance computational capabilities: memory and processing time to build a complete world model.
- It has a higher chance of failing if the representation of the world model is neither exact nor complete.

B. The Reactive based Control Architecture

Reactive control architecture is a bottom-up approach that was developed in order to hold some drawbacks of the deliberative control architecture [13], [10]. It consists of reactive rules

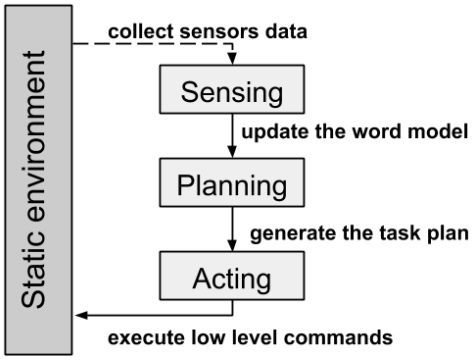


Fig. 1. Deliberative architecture

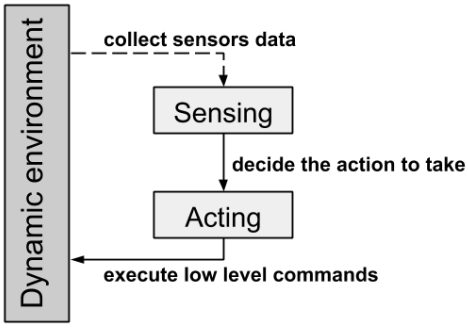


Fig. 2. Reactive architecture

set that deals with environmental changes [29], This architecture implements a control strategy as a collection of condition-action pairs that couples a sensor data to robot action. It can operate in a dynamic environment; without neither building world model nor performing planning functionality; simply by generating control commands based on sensory information [19].

As it shown in Fig. 2, the robot perceives its surrounding world by gathering its sensor data, then automatically decides what action to take. This architecture is computationally simpler than the deliberative approach and has a faster response to dynamic changes without any prior environment knowledge [10].

However, the reactive based approach cannot deal in a situation with high-level goals and complex constraints [30]. Thereby, it exclusively solves the class of problems in which the environment is well known, the goals are well defined and the robot is equipped with enough sensors that allow it to perceive the necessary information [31].

C. The Hybrid Control Architecture

To fulfill a robot mission in a real world, some features of deliberative architecture combined with the reactive architecture are required [13], [10], [32]. The hybrid approach was designed to deal with high-level goals and complex constraints in a dynamic environment [21]. It offers a compromise between reactive and deliberative-based approaches [33]. Typically, the hybrid control architecture (see Fig. 3) employs three-hierarchical levels:

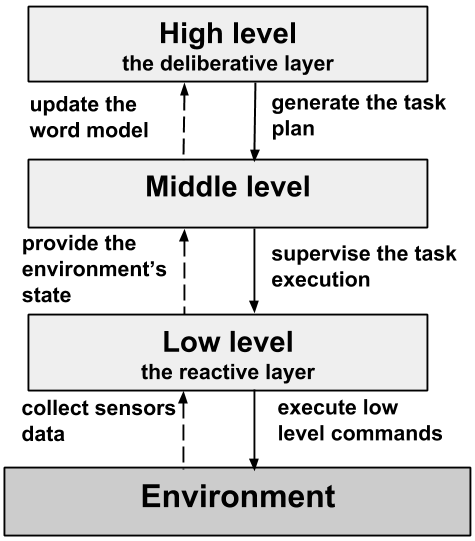


Fig. 3. Hybrid architecture

- High level (the deliberative layer) for decision making. This level performs complex computations to generate a valid task plan that corresponds to a set of actions. Each action defines a specific commands sequence sent to the reactive layer in order to generate the desired action.
- Middle level supervises the interaction between the high level and low level.
- Low level (the reactive layer) for low control senses the environment. It takes care of the immediate safety of the robot such as obstacle avoidance. To generate the robot's motion, low level executes the actions sequence provided by the deliberative layer.

D. The Behavior Control Architecture

Behavior-based control architecture [34]; which is inspired by biological studies; is designed to perform a reactive mapping between perception and action modules [35]. Basically, the behavior approach divides the control strategy into a set of behaviors as it shown if Fig. 4. Each of which is responsible for a particular task [36].

This architecture provides some advantages that make this approach more powerful than the reactive one [37], [38]. Each behavior can provide both reactive and deliberative capabilities. This architecture could deal with an unpredictable situation the robot may face without having to know the environment. Furthermore, it offers parallel and concurrently behaviors collection that act independently to achieve the robot goals. Moreover, it provides a good solution for robot's problem carrying out tasks in multiple and unknown environments.

However, this architecture presents some inconvenient:

- To control a robot, this approach must combine and coordinate several behaviors. But in some cases, it's difficult to choose the prior behavior to execute in first.
- Since behaviors represent low-level control, they may not deal with high-level goals.

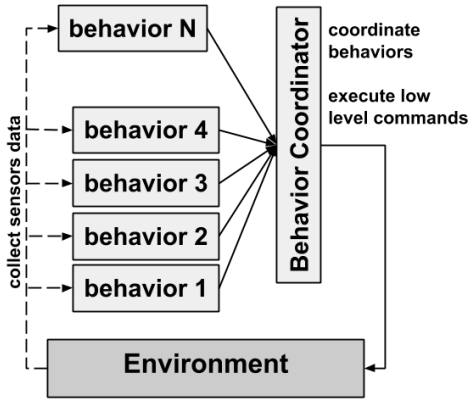


Fig. 4. Behavior architecture

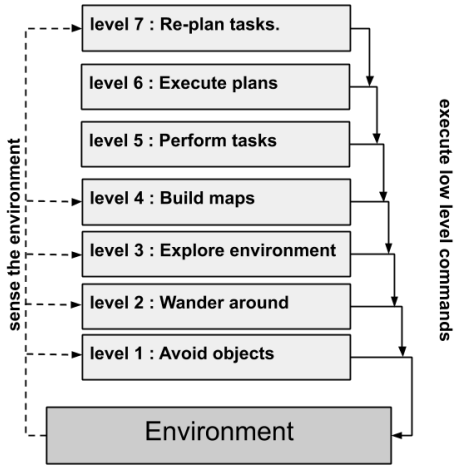


Fig. 5. Subsumption architecture

- The lack of the planning module may result in not solving complicated tasks.

E. The Hybrid Behavior Control Architecture

The main purpose of the hybrid behavior control architecture is to resolve some limits of both the hybrid and the behavior architecture [39] [40] [41]:

- It can deal with high-level goals by generating the appropriate task plan.
- It coordinates the existing behavior to reach the mission objectives.
- It is designed to be flexible.

The hybrid behavior control architecture includes three layers:

High level (called cognitive layer) reasons about mission goals and constraints. It makes task-oriented reasoning and planning, and it decides how to coordinate the current behaviors.

Middle level (or behavior layer) defines all robot's behaviors. Each of which performs a specific task.

Low level was designed to achieve the best performance of each individual behavior and to execute the low commands.

In view of inconvenient of the hybrid behavior approach we can see:

- The same architecture cannot be reused on various robot types.
- This architecture suffers from layers interdependency.

F. The Subsumption Architecture

The Subsumption architecture [24] was designed in order to offer a feasible approach for building a robust robot system. It provides a good solution for control problem dealing with multiple sensors and goals. This approach proposes an extensible layered architecture and provides a unified representation of large scale of robots [42].

The Subsumption approach divides the control problem into a set of ordered layers according to the task achieving behaviors [25]. In each layer, it defines a single competence level that performs a specific behavior such as the ability to move away from an obstacle, to move around an area or to explore the robot's environment. It ties together all those increasing layers to build the robot system.

Brooks [24] has defined seven competence levels (Fig. 5), each of which provides a specific autonomy level:

- The first level makes sure that the control system avoids contact with other objects.
- The second level offers the ability to wander around aimlessly without hitting things.
- The third one provides the robot the ability to explore its environment.
- The fourth level builds an appropriate map of the environment and plans the appropriate trajectories to reach some places;
- The fifth level reasons about goals and performs tasks.
- The sixth one executes plans to reach the desired environment.
- The last level reasons about the behavior and appropriately re-plans tasks.

The key idea of the competence level is to provide an easy way to move to the next higher autonomy level. it is reached by adding a new competence level over the existing higher control layer of the whole architecture.

Subsumption approach presents three major weaknesses [43]:

- During execution, the priority-based mechanism limits the ways the system can be adapted.
- Since higher layers interfere with lower ones, they cannot be designed independently.
- Crucial behaviors cannot always be prioritized.

TABLE I. ANALYSIS OF THE STUDIED CONTROL APPROACHES

	Deliberative	Reactive	Hybrid	Behavior	Hybrid behavior	Subsumption
Global reasoning	Yes	No	Yes	No	Yes	Yes
Reactivity	No	Yes	Yes	Yes	Yes	Yes
Adaptability	No	No	No	Yes	Yes	No
Flexibility	No	Yes	No	Yes	Yes	Yes
Adaptability	No	No	No	Yes	Yes	No
Modularity	No	No	No	Yes	No	No
Robustness	No	No	No	No	No	Yes
Sensor integration	No	No	No	Yes	Yes	No
Extensibility	No	No	No	No	No	Yes
Reusability	No	No	No	No	No	No

TABLE II. DESCRIPTION OF THE SPECIFICATIONS USED IN THE EVALUATION OF THE STUDIED CONTROL APPROACHES

Specification	Description
Global reasoning	the architecture makes suitable decisions and preserves the aircraft safety.
Reactivity	the architecture recognizes changes and avoids dynamic obstacles.
Adaptability	the architecture can carry out a large variety of mission with little reconfiguration requires.
Flexibility	the architecture can add new functionality or change the existing one.
Modularity	build up the whole architecture using a modular approach.
Robustness	the architecture can adapt and repair the task plan when some sensors fail.
Sensor integration	the proposed architecture can be adapted to new sensors.
Extensibility	the ability to improve the current autonomy level by adding additional modules.
Reusability	the same proposed architecture can be used for various UAV type.

III. CONTROL ARCHITECTURES: COMPARATIVE STUDY

In the preceding section, several contributions in term of robotic control architecture have been described. In a nutshell, we provide the comparative table below (Table I) to summarize the advantages and limitations of each of the studied control strategies. The description of the specifications used in the evaluation of the studied control approaches is listed in Table II.

The comparative study, presented in Table I, showed that the deliberative approach is the more promising control strategy for complex mission, operating in a static environment with complete knowledge of the world model. However, specifically for UAV control system, this approach is neither robust, nor flexible, nor extensible, nor reactive.

The reactive approach represents the best architecture choice for missions that require reactive navigation in a dynamic environment such as UAV target tracking. On the other hand, this approach is ineffective due to its incapability to deal with the high-level goal.

The hybrid control approach enjoys the advantage of meeting missions that need both deliberative and reactive capabilities such as Intelligent Surveillance and Reconnaissance (ISR). However, for UAV system, this control architecture may present a strong drawback, in which the decisive layer has lacks the direct access to the reactive layer; in fact, during task planning, the higher level can have old world model. By this means, the task planner may produce invalid task plan and, consequently, it may fail its overall mission.

Hybrid behavior-based approach presents a good solution for UAV control problem that needs great mission capability, higher flexibility, and adaptability to different mission kind. Nevertheless, this approach presents two weakness particularly on the UAV system: it is difficult to be implemented and it suffers from the great dependence to sensor system used on the physical robot.

Subsumption architecture offers the ability to build an intelligent and robust UAV control system with the capacity to improve its global autonomy level. However, while this hierarchical architecture depends to the physical system, it's neither modular nor reusable. In addition, due to the layer's interdependency, this approach presents the lack to integrate new sensors on the overall system. Also, it still difficult to be implemented.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have surveyed different control architecture in attempt to extract the main features that will characterize our autonomous UAV. The comparative study, done in the section three, showed that each approach is more promising in some situations. The deliberative approach can meet complex goals in static environment. The reactive strategy avoids dynamic obstacles. The hybrid architecture combines, in the same architecture, the deliberative and the reactive capabilities. The behavior approach defines a set of modules, each of which can be implemented independently. The hybrid-behavior architecture achieve complex goals by executing independent task. The subsumption approach presents the advantage of improving the autonomy level of the whole architecture.

As conclusion of this work, designing an autonomous UAV for civil applications requires the following features [44]:

- Global reasoning for meeting complex goals [45].

- Reactivity to avoid dynamic obstacles [46].
- Adaptability to carry out a wide range of mission.
- Flexibility for adding new functionality [47].
- Modularity for layers independencies [48].
- Robustness to repair the task plan, if needed [24].
- Sensor integration for the capability to add new sensors to the control architecture [24].
- Extensibility for the ability to improve the current autonomy level of the control architecture [49].
- Reusability for the ability to reuse the same architecture in various UAVs [50].

Our future work will be focused on proposing a new control architecture for autonomous UAV operating in civil domain. The proposed architecture will be characterized by the features cited above (see Table II). The proposed architecture must achieve complex goals, perform complicated tasks, compute a feasible trajectory, avoid obstacles and generate an appropriate flight plan.

REFERENCES

[1] K. P. Valavanis, *Advances in Unmanned Aerial Vehicles: State of the Art and the Road to Autonomy*. Springer Science & Business Media, Feb. 2008. Google-Books-ID: EsjPyblwMdQC.

[2] A. Vuruskan, B. Yuksek, U. Ozdemir, A. Yukselen, and G. Inalhan, "Dynamic modeling of a fixed-wing vtol uav," in *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 483–491, IEEE, 2014.

[3] H. Chen, X. Wang, and Y. Li, "A survey of autonomous control for uav," in *2009 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 2, pp. 267–271, Nov 2009.

[4] S. G Gupta, M. Ghonge, and P. Jawandhiya, "Review of unmanned aircraft system (uas)," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 9, 04 2013.

[5] J. D. Boskovic, R. Prasanth, and R. K. Mehra, "A multilayer control architecture for unmanned aerial vehicles," in *American Control Conference, 2002. Proceedings of the 2002*, vol. 3, pp. 1825–1830, IEEE, 2002.

[6] P. Fahlstrom and T. Gleason, *Introduction to UAV systems*. John Wiley & Sons, 2012.

[7] F. Kendoul, "Survey of advances in guidance, navigation, and control of unmanned rotorcraft systems," *Journal of Field Robotics*, vol. 29, no. 2, pp. 315–378, 2012.

[8] P. Liu, A. Y. Chen, Y.-N. Huang, J.-Y. Han, J.-S. Lai, S.-C. Kang, T.-H. Wu, M.-C. Wen, M.-H. Tsai, *et al.*, "A review of rotorcraft unmanned aerial vehicle (uav) developments and applications in civil engineering," *Smart Struct. Syst*, vol. 13, no. 6, pp. 1065–1094, 2014.

[9] P. Ridao, J. Yuh, J. Batlle, and K. Sugihara, "On AUV control architecture," in *Intelligent Robots and Systems, 2000.(IROS 2000). Proceedings. 2000 IEEE/RSJ International Conference on*, vol. 2, pp. 855–860, IEEE, 2000.

[10] D. Nakhaeinia, S. H. Tang, S. M. Noor, and O. Motlagh, "A review of control architectures for autonomous navigation of mobile robots," *International Journal of Physical Sciences*, vol. 6, no. 2, pp. 169–174, 2011.

[11] S. M. Veres, L. Molnar, N. K. Lincoln, and C. P. Morice, "Autonomous vehicle control systems—a review of decision making," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 2, pp. 155–195, 2011.

[12] M. Hoy, A. S. Matveev, and A. V. Savkin, "Algorithms for collision-free navigation of mobile robots in complex cluttered environments: a survey," *Robotica*, vol. 33, no. 3, pp. 463–497, 2015.

[13] R. G. Simmons, "Structured control for autonomous robots," *IEEE transactions on robotics and automation*, vol. 10, no. 1, pp. 34–43, 1994.

[14] P. P. Narayan, P. P. Wu, D. A. Campbell, and R. A. Walker, "An intelligent control architecture for unmanned aerial systems (uas) in the national airspace system (nas)," in *2nd International Unmanned Air Vehicle Systems Conference*, (Grand Hyatt, Melbourne, Australia), pp. 1–11, 2007.

[15] I. A. Nesnas, A. Wright, M. Bajracharya, R. Simmons, T. Estlin, and W. S. Kim, "CLARAty: An architecture for reusable robotic software," in *Unmanned Ground Vehicle Technology V*, vol. 5083, pp. 253–265, International Society for Optics and Photonics, 2003.

[16] S. Emel'yanov, D. Makarov, A. I. Panov, and K. Yakovlev, "Multilayer cognitive architecture for uav control," *Cognitive Systems Research*, vol. 39, pp. 58–72, 2016.

[17] R. Huq, G. K. Mann, and R. G. Gosine, "Mobile robot navigation using motor schema and fuzzy context dependent behavior modulation," *Applied Soft Computing*, vol. 8, no. 1, pp. 422 – 436, 2008.

[18] J. Pinto, J. Sousa, F. Py, and K. Rajan, "Experiments with deliberative planning on autonomous underwater vehicles," in *IROS 2012 Workshop on Robotics for Environmental Monitoring, Vilamoura, Portugal*, 2012.

[19] R. C. Arkin, "Motor schema — based mobile robot navigation," *The International Journal of Robotics Research*, vol. 8, no. 4, pp. 92–112, 1989.

[20] L. Muratet, S. Doncieux, and J.-A. Meyer, "A biomimetic reactive navigation system using the optical flow for a rotary-wing uav in urban environment," *Proceedings of the International Session on Robotics*, pp. 2262–2270, 2004.

[21] R. Peter Bonasso, R. James Firby, E. Gat, D. Kortenkamp, D. P. Miller, and M. G. Slack, "Experiences with an architecture for intelligent, reactive agents," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 9, pp. 237–256, Apr. 1997.

[22] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 121–128, ACM, 2008.

[23] M. Mataric, "Behavior-based control: Main properties and implications," in *In Proceedings, IEEE International Conference on Robotics and Automation, Workshop on Architectures for Intelligent Control Systems*, pp. 46–54, 1992.

[24] R. Brooks, "A robust layered control system for a mobile robot," *IEEE journal on robotics and automation*, vol. 2, no. 1, pp. 14–23, 1986.

[25] E. Oland, T. S. Andersen, and R. Kristiansen, "Subsumption architecture applied to flight control using composite rotations," *Automatica*, vol. 69, pp. 195–200, 2016.

[26] M. Ghallab, D. Nau, and P. Traverso, *Automated planning and acting*. Cambridge University Press, 2016.

[27] F. Ingrand and M. Ghallab, "Deliberation for autonomous robots: A survey," *Artificial Intelligence*, vol. 247, pp. 10–44, 2017.

[28] H.-D. Burkhard, J. Bach, R. Berger, B. Brunswieck, and M. Gollin, "Mental models for robot control," in *Advances in Plan-Based Control of Robotic Agents*, pp. 71–88, Springer, 2002.

[29] F. R. Noreils and R. G. Chatila, "Plan execution monitoring and control architecture for mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 11, pp. 255–266, Apr. 1995.

[30] F. Belkhouche, "Reactive optimal uav motion planning in a dynamic world," *Robotics and Autonomous Systems*, vol. 96, pp. 114–123, 2017.

[31] U. A. Sheikh, M. Jamil, and Y. Ayaz, "A comparison of various robotic control architectures for autonomous navigation of mobile robots," in *2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE)*, pp. 239–243, IEEE, 2014.

[32] K. Nonami, F. Kendoul, S. Suzuki, W. Wang, and D. Nakazawa, *Autonomous flying robots: unmanned aerial vehicles and micro aerial vehicles*. Springer Science & Business Media, 2010.

[33] A. D. Ryan, D. L. Nguyen, and J. K. Hedrick, "Hybrid control for uav-assisted search and rescue," in *ASME 2005 International Mechanical Engineering Congress and Exposition*, pp. 187–195, American Society of Mechanical Engineers, 2005.

[34] M. Mataric, "Behavior-Based Control: Main Properties and Implications," in *In Proceedings, IEEE International Conference on Robotics and Automation, Workshop on Architectures for Intelligent Control Systems*, pp. 46–54, 1992.

[35] H. T. Dinh, M. H. C. Torres, and T. Holvoet, "Sound and complete reactive uav behavior using constraint programming," in *ICAPS Workshop on Planning and Robotics*, ICAPS Workshop on Planning and Robotics, 2018.

[36] A. Brunete, M. Hernando, E. Gambao, and J. E. Torres, "A behaviour-based control architecture for heterogeneous modular, multiconfigurable, chained micro-robots," *Robotics and Autonomous Systems*, vol. 60, no. 12, pp. 1607–1624, 2012.

[37] M. Carreras, J. Batlle, P. Ridao, and G. Roberts, "An Overview on Behaviour-Based Methods for AUV Control," *IFAC Proceedings Volumes*, vol. 33, pp. 141–146, Aug. 2000.

[38] L. Iocchi, D. Nardi, and M. Salerno, "Reactivity and deliberation: a survey on multi-robot systems," in *Workshop on Balancing Reactivity and Social Deliberation in Multi-Agent Systems*, pp. 9–32, Springer, 2000.

[39] J. Liu, H. Hu, and D. Gu, "A Hybrid Control Architecture for Autonomous Robotic Fish," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 312–317, Oct. 2006.

[40] J.-D. Liu and H. Hu, "Biologically inspired behaviour design for autonomous robotic fish," *International Journal of Automation and Computing*, vol. 3, no. 4, pp. 336–347, 2006.

[41] M. Mtshali and A. Engelbrecht, "Robotic architectures," *Defence Science Journal*, vol. 60, no. 1, p. 15, 2010.

[42] G. Butler, A. Gantchev, and P. Grogono, "Object-oriented design of the subsumption architecture," *Software: Practice and Experience*, vol. 31, no. 9, pp. 911–923, 2001.

[43] R. C. Arkin, *An Behavior-based Robotics*. Cambridge, MA, USA: MIT Press, 1st ed., 1998.

[44] A. A. Zheltoukhov and L. A. Stankevich, "A survey of control architectures for autonomous mobile robots," in *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 1094–1099, IEEE, 2017.

[45] C. Sampedro, H. Bavle, J. L. Sanchez-Lopez, R. A. S. Fernández, A. Rodríguez-Ramos, M. Molina, and P. Campoy, "A flexible and dynamic mission planning architecture for uav swarm coordination," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 355–363, IEEE, 2016.

[46] A. Viquerat, L. Blackhall, A. Reid, S. Sukkarieh, and G. Brooker, "Reactive collision avoidance for unmanned aerial vehicles using doppler radar," in *Field and Service Robotics*, pp. 245–254, Springer, 2008.

[47] R. Volpe, I. Nesnas, T. Estlin, D. Mutz, R. Petras, and H. Das, "The CLARAty architecture for robotic autonomy," in *Aerospace Conference, 2001, IEEE Proceedings.*, vol. 1, pp. 1–121, IEEE, 2001.

[48] M. Proetzsch, T. Luksch, and K. Berns, "The behaviour-based control architecture ib2c for complex robotic systems," in *Annual Conference on Artificial Intelligence*, pp. 494–497, Springer, 2007.

[49] I. A. Nesnas, A. Wright, M. Bajracharya, R. Simmons, T. Estlin, and W. S. Kim, "Claraty: An architecture for reusable robotic software," in *Unmanned Ground Vehicle Technology V*, vol. 5083, pp. 253–264, International Society for Optics and Photonics, 2003.

[50] S. Joyeux, R. Alami, S. Lacroix, and A. Lampe, "Simulation in the laas architecture," in *International Conference on Robotics and Automation-Workshop on Software Development in Robotics*, 2005.

Effective Combination of Iris-based Cancelable Biometrics and Biometric Cryptosystems

Osama Ouda[1]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Al-Jawf, Saudi Arabia

Norimichi Tsumura[2]
Department of Information and Computer Sciences, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba, 263, JAPAN

Toshiya Nakaguchi[3]
Center for Frontier Medical Engineering, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba, 263, JAPAN

***Abstract*—The fuzzy commitment scheme (FCS) is one of the most effective biometric cryptosystems (BCs) that provide secure management of cryptographic keys using biometric templates. In this scheme, error correcting codes (ECCs) are firstly employed to encode a cryptographic key into a codeword which is then secured via linking (committing) it with a biometric template of the same length. Unfortunately, the key length is constrained by the size of the adopted biometric template as well as the employed ECC(s). In this paper, we propose a secure iris template protection scheme that combines cancelable biometrics with the FCS in order to secure long cryptographic keys without sacrificing the recognition accuracy. First, we utilize cancelable biometrics to derive revocable templates of large sizes from the most reliable bits in iris codes. Then, the FCS is applied to the obtained cancelable iris templates to secure cryptographic keys of the desired length. The revocability of cryptographic keys as well as true iris templates is guaranteed due to the hybridization of both techniques. Experimental results show that the proposed hybrid system can achieve high recognition accuracy regardless of the key size.**

***Keywords*—Biometric template protection; cancelable biometrics; biometric cryptosystems; BioEncoding; fuzzy commitment**

I. Introduction

In the last few years, the marriage between biometrics and cryptography has been proven to be an effective approach to address several issues inherent to both technologies [1]. In spite of the usability advantages exhibited by biometrics, several security and privacy concerns have been raised about employing biometric characteristics in identity verification/identification. First, biometric traits are limited and therefore revoking and replacing a compromised biometric trait is not as easy as canceling and replacing a compromised password or token in traditional authentication/identification systems. Second, biometric traits are permanently associated with individuals and hence some of their private information could be revealed to adversaries if these traits are disclosed [2].

On the other hand, achieving secure management of cryptographic keys is one of the most difficult problems in cryptography. Whereas short keys are not secure, it is difficult to remember and manage several long cryptographic keys. Although user-specific tokens or cards can be employed to store long cryptographic keys, keeping such user-specific tokens secure is not assured.

Fortunately, it turned out that integrating some cryptographic concepts into biometrics systems provides practical solutions to the above-mentioned issues. This integration can take one of two main forms; namely, cancelable biometrics (CB) [3] and biometric cryptosystems (BCs) [4]. CB schemes, such as BioHashing [5], BioEncoding [6] and distorting transforms [7], aim to derive several protected templates from the original (unprotected) biometric signal employing one-way transforms. The derived templates should be revocable, renewable, noninvertible and preserve the discriminability of original templates. Basically, the CB construct is inspired from the concept of one-way hash functions in classical cryptography. However, unlike cryptographic hash functions, CB can derive similar protected templates from original biometric signals that belong to the same user. In fact, due to the intra-user variations, one should not expect that biometric samples acquired from the same user to be identical. That is why most of the existing CB techniques cannot satisfy all the requirements of the CB construct simultaneously without integrating other user-specific authentication factors in the transformation process. Moreover, although CB systems can exhibit a practical solution to the problems of template protection and privacy invasion, they are still vulnerable to some attacks that are inherent to conventional biometric systems. For example, current CB systems are neither resilient to Trojan horse attacks nor to overriding Yes/No response attack.

On the other hand, BCs, such as fuzzy extractors [8], the fuzzy vault scheme (FVS) [9] and the fuzzy commitment scheme (FCS) [10], bind/extract user-specific keys to/from biometric templates such that the key is released only if a genuine biometric sample is presented at the time of verification. In fact, this construct can be employed to protect biometric templates as well as to provide a practical approach to manage cryptographic keys. Unfortunately, although different cryptographic keys can be extracted from or linked to the same biometric template in a way that allows users to log on many systems using only their irises or fingerprints, BCs are not designed to be revocable with respect to biometric templates [11]. In other words, in BCs, it is possible to replace a compromised key and bind another key to the same biometric; however, if the biometric itself is compromised, it would not be possible to revoke it.

In order to benefit from the advantages of both approaches, several attempts of combining both CB and BCs have been proposed recently [12]-[22]. Most of these attempts employ the FCS as the BC of choice due to its simplicity and efficiency. The FCS utilizes single or concatenated ECCs to correct errors

existing in different biometric samples acquired from the same user. Hence, the recognition accuracy of such hybrid systems relies primarily on the correction capability of the employed ECC(s) as well as the error rates of the adopted biometric characteristic. Moreover, the length of the key to be linked with a biometric template is constrained by both the size of that template and the used ECC(s). Therefore, to develop an effective hybrid template protection system using the FCS, all of the above issues need to be addressed.

In this paper, we propose a novel hybrid template protection system that combines both CB and BCs effectively to protect iris templates as well as cryptographic keys at the same time. A cancelable iris template (BioCode), of any desired length, is firstly generated from the most reliable bits in a true iris template using a new variant of our previously proposed cancelable transformation scheme [6]. Then, a cryptographic key is linked to the derived cancelable template employing the FCS. Iris is one of the most accurate and reliable biometric characteristics that has been successfully implemented in many real world applications with very low false rejection and acceptance rates [23]. Moreover, thanks to the proposed BioEncoding-based cancelable transformation method, the suggested hybrid system exhibits the following advantages over other existing hybrid template protection techniques: (1) no user-specific data need to be used with the proposed hybrid system (i.e., the proposed system is tokenless), (2) no restrictions are imposed on the size of the key to be secured using the proposed system, (3) both keys and iris templates could be revoked and replaced easily in case of compromise, and (4) a perfect recognition accuracy (0% ERR) can be achieved regardless of the key size.

The remainder of this paper is organized as follows. A review of the related works is presented in Section II. The FCS is reviewed in Section III. In Section IV, base BioEncoding is revisited and the proposed variant of BioEncoding is described. The proposed hybrid template protection scheme is presented in Section V. Experimental results and security analysis are presented in Sections VI and VII, respectively. Finally, conclusions are drawn in Section VIII.

II. Related Work

Several hybrid template protection schemes have been proposed in the last few years. Most of these systems utilize either the FVS or the FCS in the key binding step. On the other hand, the main difference between the cancelable transforms employed in these techniques lies in their commitment to the invertibility property.

Liu et al. [12] suggested to combine random projection based cancelable biometrics and the fuzzy vault technique to improve the security of palm-prints recognition. Sandhya and Prasad [13] utilized the fuzzy commitment scheme to protect cancelable templates constructed using Delaunay triangles from fingerprint minutiae. Kanade et al. [14]-[17] proposed several variants of a hybrid template protection scheme that depends on the FCS for binding cancelable iris and/or face templates to cryptographic keys. For obtaining cancelable templates, they suggested to shuffle bits in the original biometric templates using user-specific shuffling keys. The problem with these techniques lies in their utilization of invertible cancelable transforms rather than non-invertible ones. Thus, once these transforms are disclosed to adversaries, the original templates would be obtained easily.

Wang and Plataniotis [18] proposed a two-stage hybrid scheme for face biometrics. At the first sage, cancelable face templates are generated by quantizing the distance vectors between the extracted Principal Component Analysis (PCA) feature vector and pairs of user-dependent random vectors. Then, the FVS is utilized to bind a randomly generated key with the obtained cancelable face template. The same approach is followed by Nandakumar et al. [19] for protecting fingerprint templates. At first, a user-specific password is utilized to derive a random transformation function that is applied to the fingerprint template. Then, the transformed template is secured using the FV framework. Teoh et al. [20] proposed a two-step technique to derive personalized cryptographic keys from the face biometric. In the first step, a cancelable template (FaceHash) is generated from original face template using BioHashing (i.e., random projection followed by simple thresholding). Then, in the second step, a cryptographic key is derived from the generated FaceHash via Shamir's secret-sharing approach. A similar technique is presented by Song et al. [21] but for fingerprint templates. A cancelable fingerprint template, referred to as FingerHash, is firstly generated using BioHashing and then linked via the FCS to a cryptographic key that is encoded using a Reed-Solomon code. More recently, Feng et al [22] proposed a three-stage hybrid algorithm for face templates. Cancelable templates are generated at the first stage using random projection. At the second stage, a discriminability-preserving (DP) transform is applied to cancelable templates in order to enhance the discriminability of the original feature templates as well as converting the resulting cancelable template into a binary string. Finally, the binarized cancelable template is linked to a randomly generated binary string (the key) using the FCS. Although all the above-mentioned techniques employ non-invertible cancelable transforms, they utilize user-specific random numbers in the cancelable transformation process. That is, they suffer from the same problems associated with traditional-based authentication systems. In other words, if these random keys are compromised, the False Acceptance Rate (FAR) would increase significantly [24]. Moreover, although the employed non-invertible transforms are much harder to be reversed compared to invertible transforms, it would be much simpler for a skilled attacker to invert the transform and disclose the original features if he/she could gain access to these user-specific factors.

Unlike other CB methods, BioEncoding satisfies all the requirements of the CB construct yet does not require any user-specific passwords/keys to be employed in the cancelable transformation process. On the other hand, unlike the FCS, the FVS suffers from a number of security vulnerabilities [25] as well as some implementation difficulties. Therefore, we believe that combining both BioEncoding and the FCS into a hybrid template protection system would address most of the issues associated to other existing hybrid techniques. However, as mentioned earlier, some issues inherent to the FCS need to be dealt with. Addressing these issues is the main goal of the new BioEncoding-based cancelable transformation method proposed in this work.

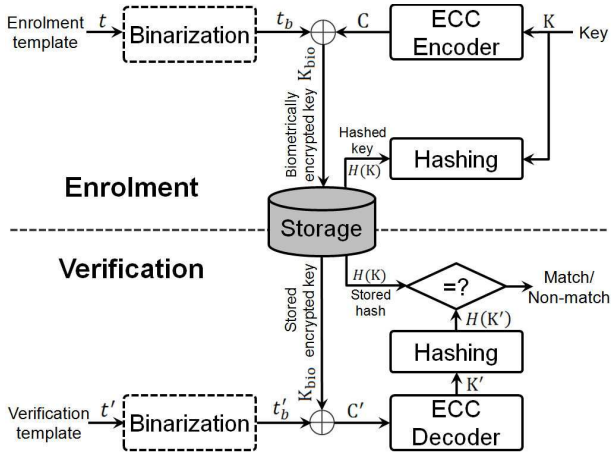


Fig. 1. Illustration of the Fuzzy Commitment Scheme.

III. Fuzzy Commitment Scheme

The FCS is one of the most attractive BCs that have been proposed so far. Due to its simplicity and efficiency, it has been applied to several biometric traits by many scientific groups [26]-[33]. As depicted in Fig. 1, the FCS requires a biometric template, t, to be represented as a binary string. Hence, a non-binary templates need to be binarized via an optional binarization module into a binary template, t_b, before the FCS can be applied. For iris biometric, since iris features are commonly represented as an ordered binary string, known as an iris code, the binarization step is not required (and hence t_b would be identical to t in Fig. 1). The FCS works as follows. On enrollment, a random binary key K is generated and encoded using an appropriate ECC(s) into a codeword C of length $n = \|t_b\|$. Both the binary template and the encoded key are then XORed to produce a biometrically encrypted key K_{bio}, also called a biometric key, as follows:

$$K_{bio} = C \oplus t_b \tag{1}$$

Furthermore, the hash value of the random key $\mathrm{H}(K)$ is computed and stored with the biometric key in a central storage or a user-specific token. At the time of verification, a binary template $t_b^{'}$ is extracted from a live biometric sample captured from the person being verified and XORed with the stored biometric key to obtain a possibly corrupted codeword C':

$$C' = t'_b \oplus K_{bio} \tag{2}$$

The obtained codeword is decoded using the ECC(s) employed on enrollment to get the verification key K'. Finally, the hash value of the recovered key, $\mathrm{H}(K')$ is computed using the same hashing function employed on enrollment and compared to the stored hash value, $\mathrm{H}(K)$. Only if the two hash values are identical, the key is released; otherwise, the authentication process fails.

IV. Generating Cancelable Templates

To bind long cryptographic keys with biometric templates, the size of such templates should be as large as possible. If (cancelable) biometric templates of any size could be generated, keys of unconstrained lengths would be employed in the key-binding phase. Unfortunately, cancelable biometrics techniques, such as BioHashing and BioEncoding, derive compact revocable templates from original biometrics templates. In BioHashing, if the generated cancelable template is not shorter than the original one, the transform could be inverted easily [34]. Also, the protected BioCodes generated using BioEncoding, as proposed in [6], are shorter than their corresponding original templates.

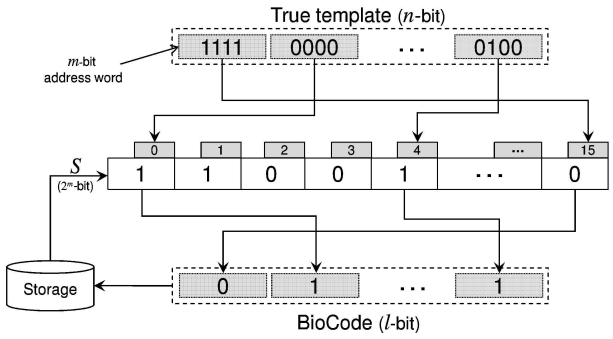


Fig. 2. Cancelable transformation of base BioEncoding.

In this paper, we propose a cancelable transformation method, based on BioEncoding, which can derive protected templates of any desired lengths from binary biometric data. In this section, we first give a brief overview of the base BioEncoding cancelable transformation scheme and then we describe the proposed variant.

A. Base BioEncoding

In BioEncoding, as illustrated in Fig. 2, bits in the true template are grouped into n/m m-bit words, where n is the bit-length of the original (unprotected) template. At the same time, a binary string S of length 2^m is randomly generated. Each word in the true iris code is mapped to a bit value in S located at the position addressed by the value of that word. For example, the first word '1111' in the true template, shown in Fig. 3, is mapped to the bit value located at position 15 $(= 1111_b)$ in S, i.e. '0'. The l $(= n/m)$ addressed bit values constitute the cancelable template (BioCode).

The most important advantage of BioEncoding over other CB methods is that it does not require the random sequence S to be neither unique nor secret. In other words, S need not be user specific as in BioHashing [5] and distorting transforms [7], for example. This is because even if the same sequence is employed with all users, different BioCodes will be generated due to the randomness that exists between iris codes generated from different eyes. In [34], BioEncoding was compared experimentally to BioHashing and the obtained results showed that BioEncoding, unlike BioHashing, does not deteriorate the recognition accuracy of the original biometric system even under the stolen-token scenario. Furthermore, the transform is non-invertible due to its many-to-one nature. Therefore, attackers would not be able to invert the transform, even if both a protected BioCode and S are known [6], [34]. In addition, BioEncoding offers significantly high renewability capacity. In fact, there are 2^{2^m} different binary strings that can be addressed using address words of length m. Therefore, a compromised

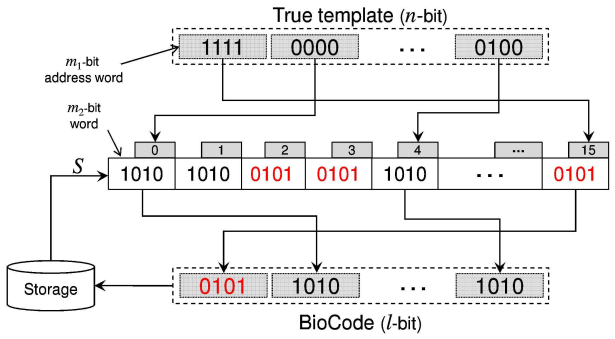


Fig. 3. Improved BioEncoding: bits replaced by words in S.

BioCode can be revoked and replaced easily by a new BioCode generated using a different string. Moreover, different random strings can be employed in different applications in order to guarantee diversity that hinders attackers from attacking different databases using cross-matching. Finally, with respect to the recognition accuracy, the performance of the original iris recognition system could be preserved if the random string S satisfies the following two conditions [6]: 1) has equal number of zeros and ones, i.e. balanced, and 2) totally different words in the true template, such as the words '0000' and '1111' shown in Fig. 2, address different bit values in S. In order to ensure that a randomly generated string satisfies these conditions, it should be tested before adopting it using the following formula:

$$f(S) = \sum_{i=0}^{2^m-2} \sum_{j=i+1}^{2^m-1} [1 \oplus (S_i \oplus S_j)].d_h(w_i, w_j) \qquad (3)$$

where m is the length of any address word in the true template, S_i and S_j are the ith and jth bits in S, w_i and w_j are their corresponding address words, respectively; $\oplus$ is the XOR Boolean operator, and d_H stands for the hamming distance between the address words w_i and w_j.

Strings that have many similar bits addressed by different words in the true templates would have large f values and therefore should be avoided since they may decrease the discriminability between the generated BioCodes. However, it should be noted that although satisfying these conditions is necessary to preserve the recognition accuracy of the original unprotected system, it would restrict the renewability capacity since only $2m$ binary strings of length 2^m can strictly satisfy the above-mentioned conditions.

B. Modified BioEncoding

The basic idea behind BioEncoding lies in mapping multiple address words in the true template to one of two values; namely '0' or '1'. This many-to-one transformation is necessary to achieve the non-invertibility property required by CB. However, this does not necessarily mean that the random string S must consist of a sequence of 0's and 1's. Rather, any two distinct values could be used instead. For example, adopting a suitable matching technique in the transform domain, S might consist of two different words such as 'black' and 'white', two different symbols such as '+' and '-', or even two different binary words such as '1010' and '0101'.

In this paper, we propose to map each address word of size m_1 in the true template to one out of two binary words of size m_2, where $m_2 \geq 1$. Fig. 3 shows an example where $m_1 = m_2 = 4$. This approach exhibits two major advantages over base BioEncoding: 1) increased renewability capacity since different variations of word lengths and forms can be employed, and 2) more flexibility in choosing the length of the resultant BioCode since the length of the protected BioCode could be decided at will. The length l of the resultant BioCode could be decided according to the following formula:

$$l = (n \times m_2)/m_1 \qquad (4)$$

Specifically, the second advantage is very important to our proposed hybrid template protection scheme since long BioCodes will allow for linking long enough keys as described in the next section.

V. Proposed Hybrid System

This section presents the proposed hybrid template protection system for securing cryptographic keys using arbitrary length protected iris codes. Fig. 4 illustrates the steps involved in both the enrollment and verification modules of the proposed system. The two modules are described in detail in the next subsections.

A. Enrollment

The enrollment module of the proposed system consists of two concurrent procedures. As shown in Fig. 4, the first procedure is responsible for preparing a revocable, non-invertible and protected iris template of arbitrary length n_p from a number of true iris codes of length n_o. On the other hand, the goal of the second procedure is to employ ECC(s) to encode an l-bit cryptographic key into an n_p-bit encoded string. Both processes are described below.

1) Protected templates generation: A practical biometric cryptosystem should be able to achieve high recognition accuracy, secure long-enough cryptographic keys and be robust against disclosure attacks of biometric templates. Here, we show how our proposed approach of integrating BioEncoding with the FCS can fulfill the above three requirements.

a) Extracting the most consistent bits for perfect recognition accuracy: Obviously, the recognition accuracy of the original biometric system has a significant impact on the performance of any biometric cryptosystem. Therefore, if the genuine and imposter distributions of the unprotected iris recognition system are not separated, the performance of the proposed hybrid system would be far from perfect. Generally, the only way for a biometric cryptosystem to achieve perfect accuracy is to have a clear separation between the genuine and imposter distributions of the original biometric system. Although iris is considered one of the most accurate biometric traits, such separation cannot be obtained using available public iris datasets. In fact, the near-to-perfect performance reported in some literature, such as the results reported by Hao et al. in [26] for example, is mainly due to the high quality iris images of the *private* dataset employed in their experiments (only 3 out of 630 authentic samples, employed in [26], have a relatively high bit-error rates compared to the correction capability of the employed coding mechanism).

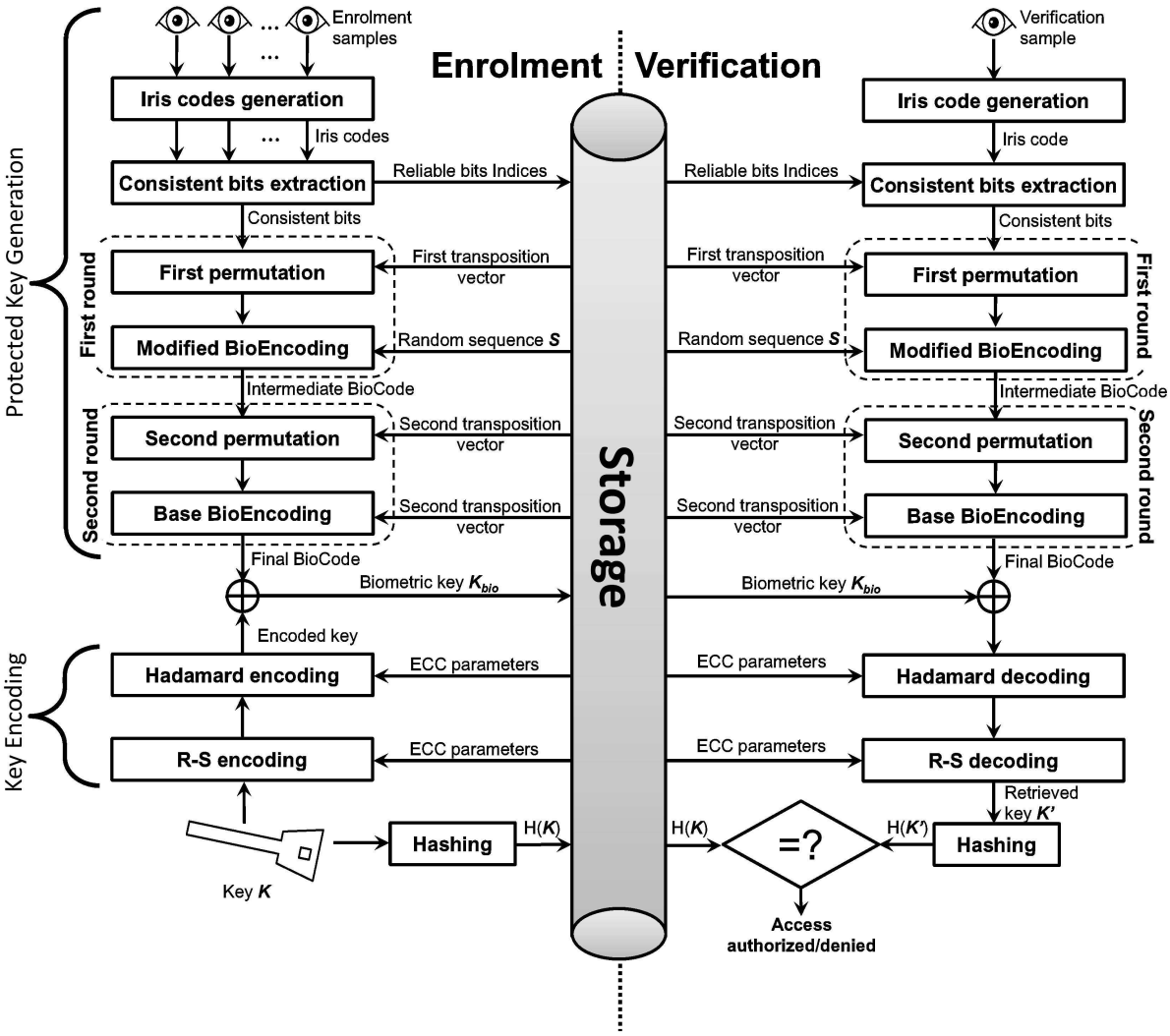


Fig. 4. Proposed hybrid template protection system.

Employing public iris datasets, a clear separation between the genuine and imposter distributions can be obtained by considering only the most consistent bits in iris codes. Consistent bits are bits that do not tend to change their values across iris codes generated from different images captured from the same eye. Hollingsworth et al. [35] showed empirically that considering only consistent bits in iris codes can improve the recognition accuracy significantly.

Therefore, in this work, a number p of iris images are acquired from a user's eye at enrollment and iris codes are generated from these images using the well known Daugman's algorithms for iris recognition [36]. Then, for all enrolled eyes in the dataset, different percentages of consistent bits (e.g. the most 50%, 40%, etc. consistent bits) are identified and extracted for each class. The genuine and imposter distributions are obtained for each percentage and the separability between the two distributions is measured using the decidability metric d' [36]:

$$d' = \frac{|\mu_i - \mu_g|}{\sqrt{(\sigma_i^2 + \sigma_g^2)/2}} \tag{5}$$

where μ_i and μ_g are the means and σ_i^2 and σ_g^2 are the variances of the imposter and genuine distributions, respectively.

Based on the obtained d' values, we adopt the largest percentage of bits that achieves a clear separation between the genuine and imposter distributions. That is, rather than applying BioEncoding on the entire iris codes, protected BioCodes are derived from the subset of bits that gives a clear separation between the two distributions and hence a perfect recognition accuracy in the original domain can be obtained.

Consistent bits are extracted by first aligning the p enrollment iris codes and then searching for bits that do not change their values across these codes. This can be done be summing up the corresponding bits and then sorting them according to the sum results. Bits that sum to 0 or p are referred to as perfectly consistent bits. For a specific number of bits $n_c < n_o$, if the number of perfectly consistent bits is

larger than n_c, we randomly select only n_c bits from them; otherwise, we proceed with bits that sum to 1 or $p-1$, 2 or $p-2$, and so on, until we get a number of bits that is equal to or greater than the specified number n_c. The separation between the resulting genuine and imposter distributions is then investigated. If the two distributions are separated, we adopt this number or percentage of consistent bits; otherwise a smaller number of consistent bits is tested until the required separation is reached. The positions of these consistent bits in the iris code of user i are collected in a position vector P_i and stored in the centralized storage to be used at the time of verification.

b) Generating arbitrary length BioCodes: In order to generate non-invertible BioCodes of any desired length, two consecutive permutation and BioEncoding rounds are applied to the extracted consistent bits. The role of the first round is to obtain protected templates, which we refer to as intermediate BioCodes, of lengths that are much larger than the length of the original templates in order to make it possible for the final BioCodes (obtained after the second round) to be linked with long cryptographic keys.

For security reasons, which are described below, these intermediate BioCodes cannot be linked directly to cryptographic keys using the FCS construction. Therefore, to make the proposed hybrid system robust against attacks that exploit employing the ECCs in biometric cryptosystems to retrieve the bound/generated key [37], a second round is required to randomize any information that might be inferred from the intermediate BioCode by deriving a random and non-invertible template, called the final BioCode, from the intermediate BioCode.

In the first round, the extracted consistent bits are firstly permuted using a random permutation key and then BioEncoded using the modified BioEncoding scheme proposed in Section III. This permutation step is necessary for two reasons. The first is to diminish the local correlation that may exist among the extracted consistent bits. Experiments conducted by Daugman [36] on a large iris dataset showed that only 249 degrees of freedom exist in 2048-bit iris codes. Therefore, permuting the extracted consistent bits is necessary for randomizing those bits and increasing their entropy. The second reason is to secure the protected BioCodes against correlation attacks [38]. One important drawback of BioEncoding is that although it is noninvertible for a single protected template, it might be possible to recover the original iris template by correlating several protected templates created from the same iris. Therefore, it is important to change the value of address words in original templates before applying BioEncoding in every new application to hinder attackers from exploiting this type of attacks. Permuting bits in the true iris codes employing different permutations in different applications is a simple yet efficient way to achieve this objective [38].

After obtaining the permuted bits, the modified BioEncoding is applied to extract the intermediate BioEncoded template. As described in Section III, using this version of BioEncoding, protected BioCodes of an arbitrary length n_i can be generated from the n_c consistent bits. Assuming that the true consistent bits are grouped into address words of size m_1, the size r of any word in the random string S can be calculated as follows:

$$r = n_i \times m_1 / n_c \tag{6}$$

As will be described later, in this work, the Hadamard ECC is employed to encode the randomly generated cryptographic key. Therefore, unfortunately, the obtained intermediate template cannot be linked directly with the encoded key. This is because it is possible to fully recover a Hadamard codeword if only few bits are known. Specifically, knowing only 7 bits, he could completely break the biometric cryptosystem of Kanade et al. [14] in which the (32, 6, 8) Hadamard ECC is employed. To boost the correction power of the Hadamard ECC, authors in [14] insert two zeros after every three bits in the original iris template. Hence, there are at least 12 known bits in every 32-bit Hadamard codeword. This is more than enough to recover all the codewords and hence the entire key.

Although no bits are explicitly known in the intermediate template obtained from the first round, a similar attack could be applied to our system if the encoded key is committed directly using this intermediate BioCode. Consider, for example, that words in the true iris template are mapped to one of the following two values in S: S_1 = "10110010" or S_2 = "01001101" where the length of words in S is 8. For the first Hadamard codeword, the attacker may assume that it starts with S_1 and hence the first 8 bits will be known. Although this is not sufficient for recovering the entire codeword, the search space will be reduced dramatically. If, on the other hand, the first Hadamard codeword starts with S_2, the result will be complementary to the S_1 case. The same procedure can be applied to the remaining parts of the Hadamard codeword at hand to reduce the search space further.

In order to hinder this type of attacks, an extra round of random permutation and (base) BioEncoding is applied. In this round, the resulting intermediate BioCode is first permuted using a second random permutation key in order to increase the system robustness against invertibiliy attacks. To realize the importance of this permutation step, let use consider the above example again assuming further that the resulting intermediate BioCode is divided into 4-bit address words in the subsequent BioEncoding step. If no permutation is applied, there would be only two possible address words for every bit in S. That is, every odd-numbered bit in the resulting BioCode could be addressed by either "1011" or "0100". Likewise, every even-numbered bit could be addressed by either "0010" or "1101". It is worth noting that although the permutation key need not be user specific, and hence the same key can be employed for all users enrolled at the same application, different permutation keys should be employed in different applications to prevent correlation attacks [38].

Using base BioEncoding, bits in the permuted template are grouped into m_2-bit words and each word is mapped to a single bit in a binary string S of length 2^{m_2}. The length n_p of the resulting (final) BioCode will be:

$$n_p = n_i / m_2 \tag{7}$$

$$= r \times n_c / (m_1 \times m_2) \tag{8}$$

According to the length of the consistent bit vector, extracted from the enrollment iris codes, as well as the required length of the protected BioCode, different values of the BioEncoding's (base and modified) parameters can be chosen. For example,

TABLE I. EXAMPLES OF DIFFERENT PARAMETER SETTINGS OF BIOENCODING WHEN $n_c = 1024$ AND $n_p = 2048$

	n_c	m_1	r	n_i	m_2	n_p
1	1024	4	32	8192	4	2048
2	1024	8	64	8192	4	2048
3	1024	4	64	16384	8	2048

if it is required to generate a 2048-bit protected BioCode from an 1024-bit consistent bit vector, there would be more than one choice for values of m_1, r, and m_2. Table I shows some examples of these values.

2) Cryptographic Key Encoding and Linking: In this module, an l-bit cryptographic key K is randomly generated and encoded into an n_p-bit codeword C using the two-layer error correcting (EC) scheme described in [26]. This concatenated EC scheme combines Hadamard and Reed-Solomon ECCs to deal with background and burst errors in iris codes, respectively. In the first layer, K is encoded using a Reed-Solomon ECC as follows. Bits in K are divided into k_{RS} blocks of k-bit each. This set of blocks is then represented as a message of k_{RS} symbols over F_{2^k} and encoded into a codeword of n_{RS} symbols using a (n_{RS}, k_{RS}, t_{RS}) Reed-Solomon code that has a correction capacity $t_{RS} = (n_{RS} - k_{RS})/2$. In the second layer, each of the resulting n_{RS} symbols is represented as a k-bit word and encoded into a 2^{k-1} bit codeword using a (2^{k-1}, k, 2^{k-2}) Hadamard code. Such Hadamard code is generated from a Hadamard matrix of order $k - 1$ and can correct $2^{k-3} - 1$ erroneous bits in each codeword. The length n_p of the final encoded key can be calculated as follows:

$$n_p = n_{RS} \times 2^{k-1} \tag{9}$$

The correction capability of this concatenated EC scheme depends primarily on the values of both n_{RS} and k_{RS}. The lowest correction capacity is obtained when $n_{RS} = k_{RS}$ (i.e., only Hadamard encoding is employed). That is, the described two-layer EC scheme can correct at least up to approximately 25% of the encoded codeword, since the correction capability of sole Hadamard coeds is up to 25% [39]. On the other hand, the correction capability of the two-layer scheme is increased by increasing the difference between n_{RS} and k_{RS}.

In this paper, the (128, 8, 64) Hadamard code is adopted and different Reed-Solomon codes are employed based on the required key length as well as the intra- and inter-user distributions of the generated BioCodes.

Finally, the resulting codeword is XORed with the generated protected BioCode to get the biometric key K_{bio}. At the same time, the hash value of the key $H(K)$ is computed using any secure hash function. Finally, the original iris template as well as the protected one is discarded safely and only the biometric key along with its hash value are stored in the centralized storage for further processing during verification.

B. Verification

At the time of verification, as illustrated in Fig. 4, a single iris image is captured from the eye being verified and its iris code is generated using the same procedure applied on enrollment. Using the stored position vector, the most n_c consistent bits are extracted from the generated code. It should be noted that due to the misalignment that may be found between the enrolled images and the image captured at verification, the generated iris code is shifted eight times in the left and right directions, as suggested by Daugman [36], and the process of extracting the most consistent bits is repeated after each shift. The first permutation vector, stored on enrollment, is then applied to the extracted consistent bits and the modified BioEncoding cancelable transformation is employed to derive the intermediate protected BioCode from the permuted original bits. Then, the second permutation followed by the base BioEncoding transformation process are applied to the intermediate BioCode to obtain the final protected iris code. To retrieve the secured key, the final protected code is XORed with the stored biometric key, K_{bio}, and the resulting bit string, is decoded using the concatenated scheme used on enrollment. Finally, the hash of the retrieved key, K', is computed and compared to the stored hash. Only if the two hash values $H(K)$ and $H(K')$, coincide, the key is released; otherwise, the authentication process fails.

VI. EXPERIMENTAL RESULTS

The publicly available CASIA-IrisV3-Interval iris images dataset [40] was used to evaluate the proposed system. This database contains 2639 8-bit gray scale images, with a resolution of 320 × 280 pixels, captured from 396 different classes (eyes). However, many classes in this dataset have just a small number of images. Therefore, a subset contains 700 images of 70 different classes, classes that have 10 images in the database, was used.

The open source MATLAB implementation for iris recognition provided in [41] was employed to generate 9600-bit iris codes, together with their corresponding 9600-bit noise masks, for all images in the selected subset. The normalized Hamming distances between all possible iris code pairs, considering all bits in the generated iris codes and taking the noise masks into account, were calculated using the following formula [36]:

$$d_H = \frac{\|(CodeA \oplus CodeB) \cap MaskA \cap MaskB\|}{\|MaskA \cap MaskB\|} \tag{10}$$

where $CodeA$ and $CodeB$ are the two iris templates being matched, $MaskA$ and $MaskB$ are the noise masks corresponding to $CodeA$ and $CodeB$ respectively and $\cap$ represents the bitwise AND operation. Fig. 5(a) shows the genuine and imposter normalized Hamming distances distributions for the adopted iris images. A clear overlap between the two distributions can be seen in the figure. The separation between the two distributions measured using the decidability metric (d'), defined in Eq. (5), is 4.061.

The separation between the genuine and imposter distributions was checked for different numbers of consistent bits. Fig. 6 shows the decidability values that result from comparing only consistent bits that represent different percentages (from 5% to 25% step 5%) of the entire length of an iris code (that is, 9600). We used bits that are perfectly consistent in fewer images, i.e. 5 images, then 4 images, etc. for eyes whose perfectly consistent bits are smaller than the tested percentage. It is clear from Fig. 6 that the decidability value decreases apparently when the number of the tested consistent bits exceeds 10%(960 bits) of

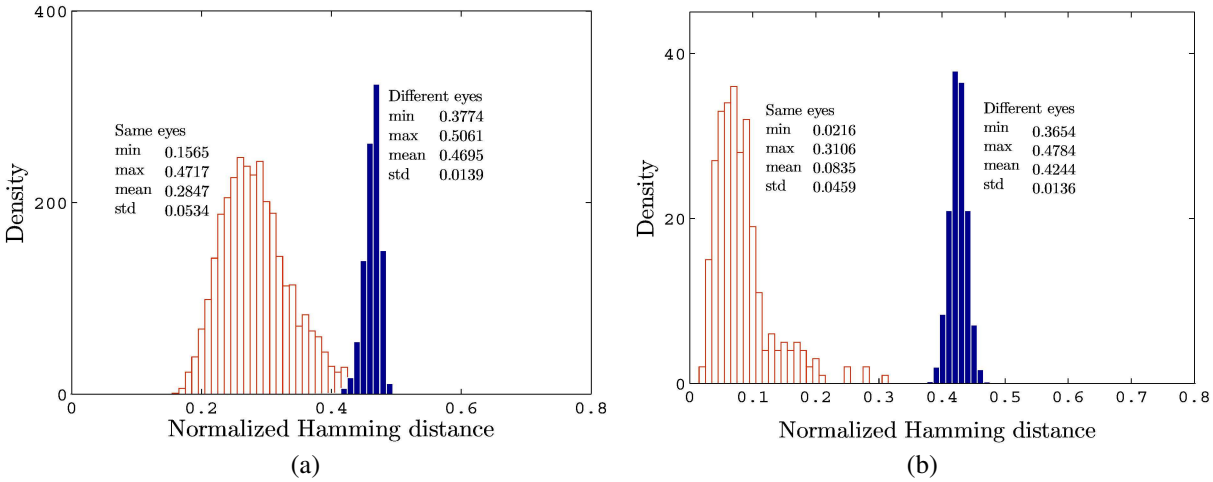


Fig. 5. Genuine and imposter distributions for (a) true iris codes, (b) the most 1024 consistent bit vectors extracted from the tested iris codes.

the entire length of the true iris code. Therefore, we decided to extract the most consistent 1024 (nearest power of two) bits from all iris codes used in our experiments. For each class in the dataset, the extracted 1024 consistent bits were collected in a consistent bit vector that is assigned to that class. Fig. 5(b) shows the genuine and imposter distributions resulted from comparing those consistent bit vectors. As shown in this figure, the statistics of both distributions imply that they are separable enough for achieving perfect recognition accuracy.

The size of the cryptographic key to be committed securely using a protected template of length n_p relies on the parameters of the employed ECCs. Hence, it is not known in advance how long the protected template should be to reliably secure a 128-bit key, for example. We began by deriving 2048-bit protected templates from the extracted 1024-bit consistent vectors and evaluating the performance of the proposed system for all possible key lengths that can be linked with these protected templates using the employed EC scheme. Recall that, as indicated from the examples in Table I, BioEncoding can be configured in many different ways to obtain 2048-bit protected templates from 1024-bit true templates.

As shown in Table II, employing the $(64, 7, 32)$ Hadamard code, it is possible to employ the derived protected templates to commit keys of lengths up to 224-bit (N.B. $224 \div 7 \times 64 = 2048$). In this case, the RS code is not used and hence the matching accuracy will be affected since no block errors are corrected. Results in Table II shows that with the introduction of RS codes, better error rates can be obtained at the expense of the key length. Perfect accuracy was achieved for $|K| = 42$.

To achieve this perfect performance for longer keys, longer protected templates should be generated. We repeated the above experiment, employing the same HC, using 4096-bit protected keys. Protected templates of such size can be obtained simply by increasing the length r of words in the random string employed in the first round of the protected template generation module of the proposed hybrid system. Based on the results obtained from the previous experiment, shown in Table II, we began by evaluating the system performance for key length $|K| = 56$ and we continued checking the

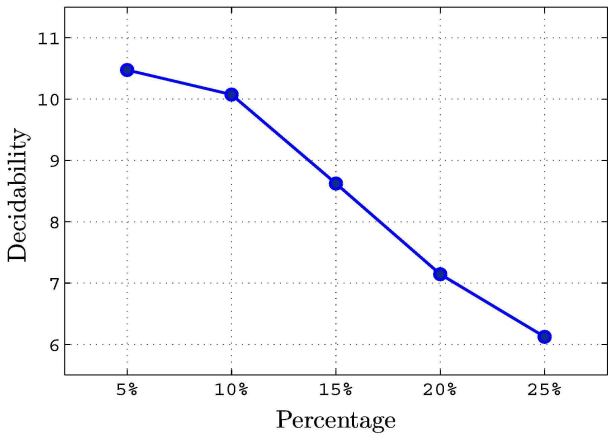


Fig. 6. Decidability values that result from comparing consistent bit vectors of different lengths (lengths represent different percentages of the entire length of an iris code).

accuracy for longer keys until a non-perfect performance was observed. Table III shows that a cryptographic key of size up to 154-bit can be committed and released perfectly (with 0% error rates) using the proposed hybrid system when 4096-bit protected templates were employed.

The above experiment was repeated again using protected templates of length $n_p = 6144$. Employing the same EC scheme, we checked the system performance for different key sizes (starting from 168-bit keys). The obtained results, shown in Table IV, indicates that employing protected templates of the mentioned size can secure cryptographic keys of long-enough sizes (up to 280-bit).

In general, since the proposed hybrid system can generate protected iris templates of an arbitrary length, the obtained results indicate that the proposed hybrid system exhibits perfect recognition accuracy regardless of the size of the key that is secured using the generated revocable iris template. A comparison between the proposed system and a number of recent iris cryptosystems is shown in Table V.

TABLE II. SYSTEM PERFORMANCE EMPLOYING $HC(64, 7, 32)$ WHEN $n_p = 2048$

$\|K\|$	t_{RS}	FRR(%)	FAR(%)
224	0	4.64	0
210	1	2.86	0
196	2	2.86	0
182	3	2.14	0
168	4	2.14	0
154	5	1.78	0
140	6	1.78	0
126	7	1.07	0
112	8	1.07	0
98	9	1.07	0
84	10	0.71	0
70	11	0.71	0
56	12	0.36	0
42	13	0	0

TABLE III. SYSTEM PERFORMANCE EMPLOYING HC$(64, 7, 32)$ WHEN $n_p = 4096$

$\|K\|$	t_{RS}	FRR(%)	FAR(%)
56	28	0	0
70	27	0	0
84	26	0	0
98	25	0	0
112	24	0	0
126	23	0	0
140	22	0	0
154	21	0	0
168	20	0.36	0

TABLE IV. SYSTEM PERFORMANCE EMPLOYING HC$(64, 7, 32)$ WHEN $n_p = 6144$

$\|K\|$	t_{RS}	FRR(%)	FAR(%)
168	36	0	0
182	35	0	0
196	34	0	0
210	33	0	0
224	32	0	0
238	31	0	0
252	30	0	0
266	29	0	0
280	28	0	0
294	27	0.36	0

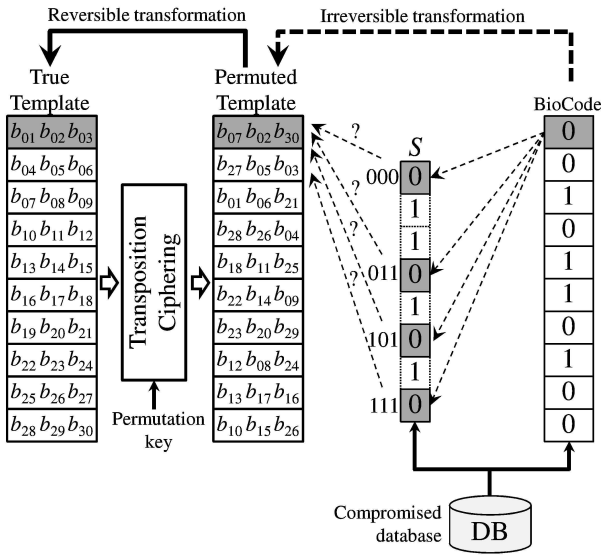


Fig. 7. Obtaining the true iris template from the protected one is infeasible due to the many-to-one nature of the transform.

VII. SECURITY ANALYSIS

The advantage of the proposed hybrid system over other hybrid template protection schemes is that it can protect cryptographic keys as well as true iris templates at the same time. Unfortunately, because other hybrid template protection techniques employ invertible cancelable transformations for deriving the protected biometric templates, it would be very simple to obtain the true templates from the protected ones if the keys, linked with the cancelable templates, are disclosed for any reason. In the proposed system, on the other hand, if a linked key is compromised, obtaining the true iris template from the protected one will be computationally very hard due to the many-to-one nature of the BioEncoding cancelable transformation process.

As shown in Fig. 7, even if both the random string S and the permutation key are known. Every bit in the protected BioCode could be originated from 2^{m-1} different address words in the true iris template where m is the size of any address word (in the example shown in Fig. 7, every BioCode bit could be originated from four different address words). Therefore, recovering all bits in the BioCode requires $2^{l(m-1)}$ trials, where l is the BioCode length. Since $l = n/m$, where n is the length of the true iris template, recovering all BioCode bits would require $2^{n(m-1)/m} \approx 2^n$, if m is large. That is, recovering the true template from the protected BioCode is approximately as difficult as guessing all bits in the true template [38]. It is important to note that although the permutation process is a reversible process, knowing the permutation key is useless unless the (irreversible) BioEncoding process is reversed. That is, the two permutation keys as well as the random strings employed in our proposed system need not be user-specific; rather, they can be treated as public data without affecting the security of the system.

On the other hand, it is important to evaluate the system security with respect to the committed cryptographic key. As shown in Fig. 5(b), the maximum intra-user fractional Hamming distance is 0.3106 and therefore a perfect accuracy can be achieved as long as the correction capability of the employed EC scheme is larger than 31.06%. As a result, in order to recover the committed key successfully, it is enough for an attacker to find a 1024-bit string which is $\leq$ 319-bit Hamming distance from the consistent iris bit vector. Accordingly, the key strength can be measured in terms of the entropy of the key (E) as follows [26]:

$$E = log_2 \frac{2^{1024}}{\binom{1024}{318}} = 114 \qquad (11)$$

That is, the attacker needs at least 2^{114} computations to recover the key successfully. It is worth noting that, however, every single computation involves two rounds of BioEncoding

TABLE V. COMPARISON WITH RECENT IRIS CRYPTOSYSTEMS

Work	Iris templates revocable?	Dataset/subset (# samples/# classes)	Key size	FRR (%)	FAR (%)
Hao et al. [26]	no	private (700/70)	140	0.47	0
Yang et al. [31]	no	CASIA ver.1 (756/108)	92	0.8	0
Lee et al. [42]	no	BERC ver.1 (990/99)	128	0.775	0
Bringer et al. [27]	no	CASIA Ver.1 (756/108) ICE 2005 (2953/244)	42	0.0665 0.0562	0 $< 10^{-5}$
Kanade et al. [14]	no	ICE 2005 (2953/244)	198	1.04	0.055
Zhang et al. [33]	no	private (348/128)	938	0.52	0
Ziauddin et al. [28]	no	Bath (500/25)	260	0	0
Chai et al. [43]	no	CASIA ver3-Interval (868/124)	200	3.63	0
Proposed	yes	CASIA ver3-Interval subset(700/70)	unlimited	0	0

and transposition processes followed by and XORing operation with the biometric key. This implies that it would be computationally very expensive, and might be infeasible, to obtain the cryptographic key committed using our proposed hybrid template protection scheme.

VIII. CONCLUSION AND DISCUSSION

This paper presented a new hybrid template protection scheme for protecting iris codes as well as securing cryptographic keys. A novel cancelable transformation method, based on BioEncoding, has been proposed to derive cancelable iris templates of any desired size form the most consistent bits in original iris codes. The derived cancelable templates were employed to secure cryptographic keys using the fuzzy commitment scheme. The proposed iris cryptosystem exhibits four major advantages over other existing systems. First, all parameters and variables employed either in the cancelable transformation process or in the binding process need not be user-specific and therefore the proposed scheme is tokenless. Second, thanks to the proposed cancelable transformation method, the presented iris cryptosystem can secure cryptographic keys of an arbitrary length. Third, since cryptographic keys are secured using cancelable iris templates rather than original ones, the revocability requirement is satisfied for both cryptographic keys and iris templates. Finally, thanks to extracting the most consistent bits from the true iris codes, experimental results showed that the proposed system achieves perfect recognition accuracy (0% ERR) regardless of the key size. This is achieved at the expense of storing the positions of the consistent bits in the application database in order to successfully match the probe sample with the gallery sample during authentication. If the adversary could gain access of these position indices, he might try to cross-match them across different applications. As a future work, we intend to deal with this issue via utilizing the challenge-response protocol in order to recover these indices without explicitly storing them in the application database.

IX. ACKNOWLEDGMENT

We would like to thank the Deanship of Scientific Research at Jouf University for providing the financial support for publishing this work.

REFERENCES

[1] D. Sadhya, S. K. Singh and B. Chakraborty, "Review of key-binding-based biometric data protection schemes." IET Biometrics, vol. 5, no. 4, pp. 263-275, 2016.

[2] Nita, Stefania, Marius Mihailescu and Valentin Pau, "Security and cryptographic challenges for authentication based on biometrics data." Cryptography vol. 2, no. 4, pp. 1-22, 2018.

[3] Nitin Kumar, "Cancelable Biometrics: a comprehensive survey", Artificial Intelligence Review, vol. 2019, pp. 1-44, 2019.

[4] Ravi Das, Biometric technology: authentication, biocryptography, and cloud-based architecture. CRC press, 2014.

[5] Y. Zheng, Y. Cao and C. Chang, "Facial biohashing based user-device physical unclonable function for bring your own device security," IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, pp. 1-6, 2018.

[6] O. Ouda, N. Tsumura and T. Nakaguchi, "Tokenless Cancelable Biometrics Scheme for Protecting Iris Codes," in Proc. IAPR Conference on Pattern Recognition (ICPR 2010), August, 2010.

[7] H. Kaur and K. Pritee, "Non-invertible biometric encryption to generate cancelable biometric templates." In Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 1-4. 2017.

[8] A. Schaller, T. Stanko, B. Škorić and S. Katzenbeisser, "Eliminating Leakage in Reverse Fuzzy Extractors," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 4, pp. 954-964, April 2018.

[9] A. Panwar, P. Singla and M. Kaur. "Techniques for enhancing the security of fuzzy vault: a review." In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 205-213. Springer, Singapore, 2018.

[10] S. Chauhan, A. Sharma, "Improved fuzzy commitment scheme." International Journal of Information Technology, pp. 1-11, 2019.

[11] R.K. Bharathi, S. D. Mohana, "A Review on Biometric Template Security." In: V. Sridhar, M. Padma, K. Rao (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol. 545, pp. 589-596, Springer, Singapore, 2018.

[12] H. Liu, D. Sun, K. Xiong, Z. Qiu, "A hybrid approach to protect palmprint templates." The Scientific World Journal, 2014.

[13] M. Sandhya and M. V. Prasad, "Cancelable fingerprint cryptosystem based on convolution coding." In Advances in Signal Processing and Intelligent Recognition Systems, pp. 145-157, Springer, 2016.

[14] S. Kanade, D. Camara, E. Krichen, D. Petrovska-Delacr'etaz and B. Dorizzi, "Three factor scheme for biometric-based cryptographic key regeneration using iris," In The 6th Biometrics Symposium 2008 (BSYM2008), September 2008.

[15] S. Kanade, D. Petrovska-Delacr´etaz and B. Dorizzi, "Multi-Biometrics Based Cryptographic Key Regeneration Scheme," In IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), September 2009.

[16] S. Kanade, D. Camara , D. Petrovska-Delacr´etaz and B. Dorizzi, ”Application of Biometrics to Obtain High Entropy Cryptographic Keys,” in Proceedings of World Academy of Science, Engeneering and Technology, vol. 39, 2009.

[17] S. Kanade, D. Petrovska-Delacr´etaz and B. Dorizzi, ”btaining cryptographic keys using feature level fusion of iris and face biometrics for secure user authentication,” In IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), June 2010, 138-145.

[18] Y. Wang and K. N. Plataniotis, ”Fuzzy vault for face based cryptographic key generation,” in Biometric Consortium Conference, Baltimore, September 2007.

[19] K. Nandakumar, A. Nagar and A. K. Jain, ”Hardening Fingerprint Fuzzy Vault Using Password,” in Proceedings of Second International Conference on Biometrics, Seoul, South Korea, pp. 927-937, 2007.

[20] A. B. J. Teoh, D. C. L. Ngo and A. Goh, ”Personalised cryptographic key generation based on FaceHashing,” Comput. Secur. 23:606-614, 2004.

[21] O. T. Song, A. B. J. Teoh and D. C. L. Ngo, ”Application-Specific Key Release Scheme from Biometrics,” International Journal of Network Security, vol. 6, no. 2, pp. 127-133, March 2008.

[22] Yi C. Feng, Pong C. Yuen and Anil K. Jain, ”A Hybrid Approach for Generating Secure and Discriminating Face Template,” IEEE Transactions on Information Forensics and Security, 5(1):103-117, 2010.

[23] J. Daugman, ”Information theory and the iriscode.” IEEE transactions on information forensics and security, vol. 11, no. 2, pp. 400-409, 2015.

[24] B. Topcu, C. Karabat, M. Azadmanesh and H. Erdogan, ”Practical security and privacy attacks against biometric hashing using sparse recovery.” EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, 2016.

[25] M. Lafkih, P. Lacharme, C. Rosenberger, M. Mikram, S. Ghouzali, M. El Haziti, D. Aboutajdine, ”Vulnerabilities of fuzzy vault schemes using biometric data with traces.” In International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 822-827, 2015.

[26] F. Hao, R. Anderson and J. Daugman, ”Combining crypto with biometrics effectively,” IEEE Transactions on Computers, vol. 55, no. 9, pp. 1081-1088, 2006.

[27] J. Bringer, H. Chabanne, G. Cohen, B. Kindarji and G. Z´emor, ”Theoretical and practical boundaries of binary secure sketches,” IEEE Transactions on Information Forensics and Security, 3(4):673-683, 2008.

[28] S. Ziauddin, M. N. Dailey, ”Robust iris verification for key management,” Pattern Recognition Letters, vol. 31, no. 9, pp. 926-935, 2010.

[29] T. A. T. Nguyen, D. T. Nguyen and T. K. Dang, ”A multi-factor biometric based remote authentication using fuzzy commitment and non-invertible transformation.” In Information and Communication Technology-EurAsia Conference, pp. 77-88, 2015.

[30] M. Yasuda, T. Shimoyama, N. Abe, S. Yamada, T. Shinzaki, and T. Koshiba, ”Privacy-preserving fuzzy commitment for biometrics via layered error-correcting codes.” In International Symposium on Foundations and Practice of Security, pp. 117-133, 2015.

[31] S. Yang, I. Verbauwhede, ”Secure Iris Verification,” Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. II-15-20, Honolulu, HI, April 2007.

[32] S. Adamovic, M. Milosavljevic, M. Veinovic, M. Sarac and A. Jevremovic, ”Fuzzy commitment scheme for generation of cryptographic keys based on iris biometrics.” IET Biometrics, vol. 6, no. 2, pp. 89-96, 2016.

[33] L. Zhang, Z. Sun, T. Tan and S. Hu, ”Robust biometric key extraction based on iris cryptosystem,” In Proceedings of the 3rd International Conference on Biometrics 2009 (ICB09) LNCS: 5558, pp. 1060-1070, 2009.

[34] O. Ouda, N. Tsumura and T. Nakaguchi, ”BioEncoding: a reliable tokenless cancelable biometrics scheme for protecting IrisCodes,” IEICE Trans. Inf & Syst., vol.E93-D, no.7, July 2010.

[35] K. P. Hollingsworth, K. W. Bowyer and P. J. Flynn, ”The best bits in an iris code,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pp. 964-973, 2009.

[36] J. Daugman, ”Evolving methods in iris recognition.” In IEEE International Conference on Biometrics: Theory, Applications, and Systems,(BTAS07), 2016.

[37] D. Akdoğan, D. K. Altop and A. Levi, ”‘Secure key agreement using pure biometrics.” In 2015 IEEE Conference on Communications and Network Security (CNS), pp. 191-199, 2015.

[38] O. Ouda, N. Tsumura and T. Nakaguchi, ”Security enhanced BioEncoding for protecting iris codes,” in Biometric Technology for Human Identification VIII, Proc. SPIE 8029, 80291U, 2011.

[39] C. Francisco, T. A. Oliveira, A. Oliveira and L. Grilo, ”Hadamard matrices and links to information theory.” In AIP Conference Proceedings, vol. 1978, No. 1, p. 460008, AIP Publishing, 2018.

[40] CASIA iris image database, Available: http://www.cbsr.ia.ac.cn/Databases.htm.

[41] Libor Masek, Peter Kovesi. MATLAB Source Code for a Biometric Identification System Based on Iris Patterns. The School of Computer Science and Software Engineering, The University of Western Australia. 2003.

[42] Y.J. Lee, K. Bae, S.J. Lee, K.R. Park, J. Kim, ”Biometric key binding: Fuzzy vault based on iris images,” in: Springer LNCS 4642: International Conference on Biometrics, August 2007, pp. 800-808.

[43] T. Y. Chai, B. M. Goi, Y. H. Tay, and Z. Jin, ”A New Design for Alignment-Free Chaffed Cancelable Iris Key Binding Scheme.” Symmetry, vol. 11, no. 2, p. 164, 2019.

Learning Management System Personalization based on Multi-Attribute Decision Making Techniques and Intuitionistic Fuzzy Numbers

Jorge Luna-Urquizo
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—The personalization of Learning Management Systems is a fundamental task in the current context of e-Learning and the WWW. However, there are many controversies around the criteria used to make the selection and presentation of the most appropriate content for each user. The most used approaches in the last decade were the identification of learning styles, the analysis of the history and navigational behavior, and the classification of user profiles, without finding conclusive evidence to determine a method that can be adopted universally, considering the complexity of the cognitive processes involved. This paper proposes an approach based on multi-attribute decision making techniques, which allows considering and combining the criteria most effectively used in the area, according to particular contexts, as a new approach to the content personalization and appropriate learning objects selection. The application of this approach aims to maximize the effectiveness and efficiency of the teaching process and enrich the user experience.**

Keywords*—*Learning Management Systems (LMS); e-Learning; multi-attribute decision making; learning styles; content personalization; learning objects selection

I. Introduction

Over the last decade, online learning environments, such as e-Learning platforms, Massive Open Online Courses (MOOC's), and educational mobile applications, as well as the amount of resources available on the WWW, have grown exponentially, facilitating access to education at all levels and employability skills training. However, this has also made it increasingly difficult to find, select and customize the contents, activities and experiences most appropriate for the needs of each user, according to their individual characteristics, which constitutes an essential requirement to achieve meaningful learning effectively.

It is for this reason that researchers in the area have been developing various techniques, tools, and models focused on the personalization of learning management systems (LMS), the most widely used being the identification of learning styles (LS), the analysis of the history and navigational behavior, the classification according to user profiles, among others. However, there are still many controversies about the effectiveness of the criteria used, and little conclusive evidence that allows the generalization of these approaches, or the convergence towards some method that can be universally adopted, considering the complexity of cognitive processes and diversity typical of human nature.

In accordance with this, new trends in education, such as project-based learning, problem-based learning, or the competency-based approach; the different approaches and theories presented in recent years; and the evidence of the most recent case studies, seem to point to the need to consider other strategies for the appropriate selection of contents, or combinations thereof, as well as the attributes of each learning object to be used or task to solve. According to [1]: "while there is no one perfect way, teachers across the globe strive to identify a right mix of pedagogies to effectively capitalize on the learning capacity and styles of the new generation of college students".

This paper proposes a novel approach to improve the implementation and customization of LMS, which allows combining the most important and relevant educational approaches and strategies in the area, through an evaluation scheme that can be adapted to each particular context and educational level, together with the profile of each student (individual characteristics), as well as a mechanism to correlate these attributes with the learning objects and main activities available on the most used LMS platforms, in order to achieve a high level of customization.

The importance of this approach is that it will facilitate the implementation of personalized learning environments that enrich the user experience and maximize the efficiency of the teaching / learning process. One of the limitations of the proposed approach is that, by combining different approaches and techniques, it requires the participation of experts who master all the techniques considered and can give an opinion regarding them, as well as a specialized repository of learning objects with a great diversity of resources, which could delay its implementation.

The rest of the paper is organized as follows. Section 2 presents a literature review and related works; Section 3 details the proposed model, including the definition of the user profile and the proposed method for the selection of learning objects, as well as other resources and activities, and Section 4 presents the conclusions of this research.

II. Related Work

The identification of LS has been one of the most widely used approaches for the design and implementation of adaptive systems and LMS. According to Nafea et.al. [2], several studies indicate that "adaptive e-learning environments based on specific LS are not only more productive, but also create higher student satisfaction levels, decrease learning times, and increase students academic achievement", while in the opposite

case: "students with a strong preference for a specific LS have difficulty learning when it is not supported by the teaching environment" [2]. For Ali et al. [3], LS "can be considered as a valuable factor for enhancing learning process by adopting an effective learning technology", so various authors agree that "personalized learning is a key in the field of e-Learning" [3].

In this regard, there are different classifications of LS, as well as different techniques to determine the particular LS of a student. Theoretically, the predominant LS is the one that would allow him to maximize the efficiency of the learning process. These techniques range from the most traditional approaches as the psychological questionnaires or tests, up to the methods of automatic detection or prediction, based on Artificial Intelligence and data mining techniques, such as the proposals by [2], [3], [4], [5], [6]. Within the review of related works, carried out for the present work, the methods proposed by Alfaro et al. [7], Aguilar et al. [8], and Seyal et al. [9], have been particularly considered, and will be briefly described in the following paragraphs.

Alfaro et al. [7], presents an approach based on neural networks and fuzzy logic, used to automatically identify the LS of each student, analyzing their interactions within the platform through the log-file, and to go reviewing and updating the identification made over time. To do this, the authors identified the relationship between the different LS, based on the Honey-Alonso model, and the main categories of resources and activities available on the Moodle platform, which are similar in the most used LMS around the world. This tool serves as an instrument for analyzing the user's behavior and choices within the platform. The authors also considering the cases in which the students are more adaptable to any LS, and the cases in which erratic behaviors occur in a single learning session, which do not affect the overall behavior, obtaining acceptable results compared to other techniques.

In Aguilar et al. [8], Social Learning Analytics (SLA) is used, wich "focuses mainly on the analysis of social networks (SNA) and the WWW, to obtain hidden information in large amounts of data (Big Data), and discover patterns of interaction and behavior of educational social actors", with the objective of determining the particular LS of each student. The importance of this approach is that it allows obtaining a point of view outside the learning environment, considering the interactions of the subject in "situations with greater freedom of action and greater diversity of resources" [8]. This work involves concepts such as Big Data, Semantic Mining, Text Mining, Data Mining, among others domains. Similarly, the approach proposed by Hamdaoui et al. [10] could be applied, where a game is used to collect "certain metrics and information susceptible of monitoring the player's interactions in the game", with which it is possible to make a correlation of students behaviour with LS, in a context totally different from LMS or other conventional learning tools.

Seyal et al. [9] proposes to examine "the relationship between students' personality traits and LS". For this purpose, the authors applied a quantitative survey, based on a random sampling approach, and the results were analyzed using statistical techniques, such as Chi-square test, the Big-five (Ocean) personality theory, to investigate the student's personality traits, and the VARK LS model, to investigate the student's LS preferences, establishing certain relationships between both domains. This study aims to helping educators to design an develop "an effective teaching/learning style as a rewarding one other than focusing on the traditional classroom environment". According to [9], "when the educator's teaching methods match the learning styles of the students, the chance for them to learn easily and understand quickly can be advantageous to their university ranking due to high rate of degree achievement".

While there are a large number of case studies and published articles about the application and efficiency of LS in the learning process, such as those presented in [2], [3], [4], [5], [6], [7], [8], [10], [9], [11], [12], [13], [14], their efficiency has been highly questioned. The most recent studies seem to indicate that there is insufficient evidence to consider that the LS concept, by themselves, can have a significant impact on the learning process or in the knowledge retention process, as mentioned in [15], [16], [17], [18], [19], [20], [21]. There are even discussions on the validity of the methods used as a reference, as well as the controversy over whether LS they are "fixed, biologically determined and inflexible dispositions, or if they are dynamic, adaptable and flexible characteristics" [22]. However, since its use is very widespread, so you should not lose sight of this approach.

Another approach to consider, whose use has been gaining strength in recent years, is the Cognitive Load Theory (CLT), which uses an information processing approach to cognition, based on the human cognitive architecture. According to Moussa-Inaty et al. [23], CLT "was a better predictor of student performance than student LS", but at the same time maintains a partial relationship with said theory, especially evidenced in multimedia learning environments, as stated in [24]. According to [25], CLT " it is a major theory providing a framework for investigations into cognitive processes and instructional design". One of the advantages of the CLT approach is that it considers simultaneously the structure of information and the cognitive architecture that allows learners to process that information.

This approach has been successfully applied in the achievement of complex learning, taking into account student experience, and allowing differentiate and prioritize between several learning objectives in progress [26], in simulation-based education applications such as [27], [28], multimedia learning with different objective measures [29], game-based learning approaches [30], among others. However, this approach, as well as the techniques commonly associated with it, also has some shortcomings, for example, according to [31]: "recent developments in CLT suggest that the human motor system plays an important role in cognition and learning; however, it is unclear whether models of working memory (WM) that are typically espoused by CLT researchers can reconcile these novel findings".

The research in affective computing seeks to understand and manage the influence of the affective or emotional factors of the person during their interaction with the technology. LMS are not out of this approach, the most important factors being the influence of the emotions and personality of the student. For Jarvenoja [32], "emotions usually manifest themselves in collaborative scenarios or discussion groups, and can influence learning". According to Matthews [33]: "Personality traits are dispositions towards action, belief and attitude formation, dif-

fer across individuals and influence behavior" and in practice, "they can modulate the way in which the student participates in a given situation" [34]. For Santos [35], affective computing could help to "enrich the personalized support provided in online educational settings by taking into account the influence that emotions and personality have in the learning process", and delivery of the "appropriate affective support in diverse educational settings".

Nafea et al. [2], presents the concept of Learning Behavior Patterns, which are defined using three variables: (i) the time spent on each LO; (ii) the number of messages sent; (iii) the format of the LOs accessed during a session. These variables are correlated by a set of equations, and used as inputs for an algorithm called "learning style adaptation algorithm" [2], taking as reference the Felder and Silverman LS model. Such research proposes an architecture based on four stages: information retrieval (system log-file analysis), dynamic checking sessions (profile construction), profile adaptation phase, and updating student profile after each learning session, in order to establish a user profile based on LS and user behavior within an LMS, the which is updated after each learning session, and is evaluated using two metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Another approach used in this study is the Learning Progressions (LP), which is defined by Plummer [36], as "the meaningful sequencing of teaching and student learning expectations accounted for across disciplines, student developmental stages, and grades". LP provide a scope and a work plan for teachers to develop student knowledge and skills. According to [37], LP are characterized by two traits: "(i) Standards intended to address student abilities, social, emotional, and physiological needs; (ii) Sequence of standards to meet necessary expectancy and actualization". The LP approach uses the concepts of nodes and subnodes, where "each learning node in the progression depicts knowledge and/or skills a student must attain before moving along the learning spectrum", and the concept of construct maps, defined by [38] as: "representations of models of cognition by which the results of the assessment can be interpreted".

In an integrated framework, many of the approaches mentioned above could be considered as criteria or attributes, and then, the problem of combining and evaluating multiple alternatives based on these approaches, to find the best options for a student in particular within an LMS, lies in the theory of decision making with multiple attributes (MADM). This theory is one of the most important techniques used in the business administration and financial field in recent years, as can be seen in [39], [40], [41], [42]. This approach has also recently been used in other areas such as the petrochemical industry, heterogeneous wireless networks, sustainable renewable energy development, satellite layout, among others.

For example, in Budiharjo et.al. [43], MADM is applied in selecting the best elementary school, considering criterias such as teacher information, student information, number of courses, infrastructure, among others. Purnomo and Rozi [44], uses weighted product (wp) and simple additive weighting (saw) methods in the best graduation selection system. In Perez et al. [45] the diffusion intuitionist dimensional analysis is applied to the selection of suppliers within the supply chain, considering various criteria such as delivery time, load capacity, cost, repeatability, quality of service, and programming flexibility, with the objective of "establishing competitive advantages for the company" [45].

According to Keshavarz et al. [46], MADM problems usually involve discrete decision variables and a limited number of alternatives for evaluation, so that "uncertainty is an inevitable part of information when the evaluation process is performed by human judgement". For this reason, this type of problems usually also involves the use of fuzzy logic to capture the uncertainty of evaluation processes. However, according to Perez et al. [45], in recent years the intuitionistic fuzzy sets (IFS) have gained ground replacing the classic fuzzy sets, "due to their greater ability to deal with vague information and in environments of uncertainty", especially in combination with other decision making support techniques, such as MADM.

According to [47], a IFS is "characterized by a membership function and a non-membership function, which is a generalization of the concept of fuzzy set whose basic component is only a membership function", and represents a a powerful tool to deal with vagueness. There are currently several variants of the IFS, such as: Interval-Valued IFS (IVIFS), Neutrosophic Sets (NS), Interval type-2 Fuzzy Sets, Intuitionistic fuzzy parameterized soft set, Intuitionistic Hesitant Fuzzy Sets (HFS), Tolerance-based intuitionistic fuzzy-rough set Intuitionistic Fuzzy Topological Spaces, etc. In general, an IFS can be defined according to the Equation 1.

$$A = \{\langle x, \mu_A(x), \nu_A(x)\rangle \,|x \in X\} \tag{1}$$

where:

$$\mu_A : X \to [0, 1] \tag{2}$$

$$\nu_A : X \to [0, 1] \tag{3}$$

with the condition:

$$0 \leqslant \mu_A(x) + \nu_A(x) \leqslant 1 \tag{4}$$

The numbers $\mu_A(x)$ and $\nu_A(x)$ represent, respectively, the membership degree and non-membership degree of the element x to the set A.

Kurilovas [48], presents a similar approach to the proposed, using techniques such as learning analytics and multiple criteria decision analysis, "testing with different weights of evaluation criteria on empirical data" [48]. For this purpose, the author uses an ad-hoc methodology based on the assessment of suitability, acceptance and use of personalized learning units (LU), defined as "methodological sequences of components or learning objects" [48], taking as reference the Educational Technology Acceptance & Satisfaction Model (ETAS-M), and the Felder-Silverman LS Model (FSLSM), both approaches widely accepted. The approach presented in this paper differs from the previous one in terms of the attributes considered as part of the student's profile, the selection criteria, and the operators used for decision making.

ETAS-M Model is used as a decision-making methodology aimed at assessing the applicability, acceptance and use of custom LOs, and is an extension of the Unified Theory

on Acceptance and Use of Technology (UTAUT) model, considering the premise that "usability aspects of e-learning systems cannot be treated independently from their impact on learning behavior and the pedagogical setting in which they are implemented" [49]. In UTAUT there are 7 artefacts designed to determine the intention to use one or more components of a model, of which, Kurilovas [48] considers that "four of them have a significant role for the analysis of LOs and LMS content personalization":

1) Performance expectancy (PE)
2) Effort expectancy (EE)
3) Social influence (SI)
4) Facilitating conditions (FC).

The first three factors are direct determinants of the intention to use the technology (LMS), and the last one is a direct determinant of user behavior. For the application of the model, factors of gender, age, experience, among others, are considered to moderate the impact of these four artifacts. In ETAS-M [49], influence of pedagogical paradigm (IPP) is proposed instead of social influence (SI) criteria in UTAUT. For [48], one of the advantages of this kind of model is that "not require specific high-level technological expertise from experts evaluators".

Regarding the collection and analysis of input data, regardless of the models and techniques to be used, one of the most commonly used approaches is Learning Analytics, which according to [48], "are known as the measurement, collection, analysis, and reporting of data about learners and their contexts to understand and optimize learning and environments in which it occurs". In addition to being focused on learning, it is a discipline that interacts with others of great relevance in this decade such as educational data mining (EDM), business intelligence (BI), social network analysis (SNA) and related Machine Learning (ML).

For the application in education, the basic tasks of LA should be at least:

1) Clasification: to classify each student in a classroom into one of predefined set of students group.
2) Clustering: "to determine studentes group that need special course profiling" [48].
3) Association rules: "to discover interesting relations between course elements which were used by particular students" [48].
4) Prediction: "to predict dependencies of using learning environment's activities/tools and final student's learning outcomes" [48].
5) Decision tree of students' actions. According to Kurilovas [48], "the decision tree algorithm automates and integrates the entire hypothesis generation process and evaluates the result, and are able to handle primary data with a little or no preprocessing data action". For example, in Kurilovas [48], the analysis was performed choosing student's ID as a target subject.

The approach presented in this paper is initially oriented towards traditional LMS, however, there is the possibility of applying it in future works with other technologies, such as: cloud-based student-centric context-aware systems [50], integrated gamification models in E-environments Learning (E-MIGA) [51], interactive e-learning with integrated virtual reality (VR) [52], STEM e-Learning in an Immersive VR Environment [53], and even in learning environments or tools based on virtual reality using wearable head-mounted displays (HMDs) [54], among others.

III. Proposed Model

The proposed model comprises two main components: (i) the user's profile, which is adaptive and evolves over time, (ii) the mechanism for evaluation and selection of Learning Objects and Learning activities (alternatives), which is based on MADM, using the attributes of the user profile, in order to select the most appropriate learning contents.

The model uses the open source LMS Moodle platform for the presentation of the selected contents and the interaction with the users, since it is one of the most used platforms around the world, while the feedback for the model is obtained from the analysis of user behavior (interactions with the system), through log-file analysis.

A. Student Profile

The student's profile is defined by a set of attributes (criteria), which must be applicable to evaluate all of the proposed alternatives (LOs), and at the same time they must be differentiators, since if an attribute or criterion is similar for several alternatives will not serve as a discard mechanism. Table I details the attributes defined for the construction of the student's profile.

B. Method of Evaluation and Selection of Learning Objects

The mechanism of selection and evaluation of LO's is based on diffuse multi-attribute intuitionist dimensional analysis techniques, for which it is necessary to define the following sets:

- $O = \{O_1, O_2, ..., O_n\}$, as the set of learning objects and learning activities defined for a given learning unit or skill to develop, which are stored in a specialized repository, and represent the alternatives to evaluate.
- $A = \{A_1, ..., A_n\}$, as the vector that contains the values or linguistic labels defined for each of the A_x attributes of the student's profile (Table I), which represent the criteria to evaluate.
- $DM = \{DM_1, ...DM_l\}$, which represents the set of decision makers, who can be experts in education, teachers in the area, educational psychologists or other related stakeholders.
- $W = \{w_1, ...w_m\}^T$, which represents the vector of the relative weights for each of the attributes of the user profile, according to the opinion of the experts (decision makers).

For the application of the MADM approach, as a first step, it is necessary to determine the importance of each decision

TABLE I. ATTRIBUTES DEFINED FOR THE CONSTRUCTION OF THE STUDENT PROFILE

Attributes	Description
ID	Unique identifier for each student, which allows to relate to their personal data record, and reuse the student profile with other models and educational tools.
Age	Stores the student's date of birth, from which the age of student can be calculated easily and quickly whenever necessary. This is important because certain LOs could be more appropriate according to the age range of the learner, and even more significant, when understanding historical situations or contexts more in line with the experiences, and at the other extreme, some LOs could be difficult to understand due to jargon or examples unknown to the apprentice.
LMS Learning Style	It stores, and keeps updated, the preferred LS of each student, identified from the analysis of their interactions within the LMS (user behaviour), using the method proposed by Alfaro et al. [7], based on neural networks and fuzzy logic, which takes as reference the Honey-Alonso model.
Social Learning Style	Stores the LS of each student, identified from their development in social networks and websites, obtained through Social Learning Analytics techniques according to the proposal of Aguilar et al. [8], which is done on large amounts of data, taking as reference the Felder and Silverman model.
Learning behavior patterns	Stores the LS of each student, obtained from the identification of learning behavior patterns, based on the time spent on LO's, the number of messages exchanged, and the format of the LO's accessed by the student, according to the [2] proposal, and its equivalence with the Felder and Silverman model.
Personality	Stores the student's personality type according to the Five Factor Model (FFM), obtained using the method proposed by Seyal et al. [9]. Personality traits are dispositions towards action, belief and attitude formation, differ across individuals and influence behaviour, and is much more stable than emotions. This attribute is necessary for delivery the appropriate affective support in diverse educational settings [35].
Learning Preference	Indicates if the student has a preference for some learning approach such as: individual, group, cooperative, collaborative, problem-based, project-based, among other, or is easily adaptable to any type of approach. This is important because it will allow a better planning of the resources and activities available within the course.

makers DM_i, using a linguistic term associated with a intuitionist fuzzy number (IFN), selected from the Table II. This qualification can be done taking as reference the professional experience, institutional position, level of specialization, or other factors deemed appropriate.

Then, be $DM_k = \{\mu_k, \upsilon_k, \pi_k\}$ the diffuse number of the rating of the k^{th} decisor, the weight corresponding to the

TABLE II. LINGUISTIC TERMS TO DETERMINE THE IMPORTANCE OF DMS [45]

Linguistic term	IFN (μ, υ, π)
Beginner	(0.1,0.9,0)
Practitioner	(0.35,0.6,0.05)
Proficient	(0.5,0.45,0.05)
Expert	(0.75,0.2,0.05)
Master	(0.9,0.1,0)

opinions of DM_k is calculated using the equation 5, proposed by [55].

$$\lambda_k = \frac{(\mu_k + \pi_k(\frac{\mu_k}{\mu_k+\upsilon_k}))}{\sum_{k=1}^{l}(\mu_k + \pi_k(\frac{\mu_k}{\mu_k+\upsilon_k}))} \tag{5}$$

Where:

$$\lambda_k \geq 0 \tag{6}$$

$$\sum_{k=1}^{l} \lambda_k = 1 \tag{7}$$

In second place, we proceed to determine the importance of each criterion or attribute of the user profile through the individual evaluation by each of the DM, using the linguistic terms shown in Table III. In general, all criteria cannot assume the same importance, and decision makers may have different opinions regarding the same criteria, which is one of the strengths of the proposed model.

TABLE III. LINGUISTIC TERMS TO DETERMINE THE IMPORTANCE OF ATTRIBUTES [45]

Linguistic term	IFN (μ, υ, π)
Very Unimportant	(0.1,0.9,0)
Unimportant	(0.35,0.6,0.05)
Medium	(0.5,0.45,0.05)
Important	(0.75,0.2,0.05)
Very Important	(0.9,0.1,0)

Once the evaluation has been carried out, the opinions of all DMs for each specific criterion are added using the IFWA (Intuitionistic Fuzzy Weighted Average) operator, proposed by [56]. In this way, be $w_j^k = \{\mu_j^k, \upsilon_j^k, \pi_j^k\}$ the Intuitionistic Fuzzy Number assigned to the A_j criterion for the k^{th} DM, then the weight vector of the j criterion, called w_j is calculated using the equation 8.

$$\begin{aligned} w_j &= IFWA(w_j^1, w_j^2, ..., w_j^l) \\ &= \lambda_1.w_j^1 \oplus \lambda_2.w_j^2 \oplus ... \oplus \lambda_l.w_j^l \\ &= [1 - \prod_{k=1}^{l}(1-\mu_j^k)^{\lambda_k}, \prod_{k=1}^{l}(\upsilon_j^k)^{\lambda_k}, \\ &\quad \prod_{k=1}^{l}(1-\mu_j^k)^{\lambda_k} - \prod_{k=1}^{l}(\upsilon_j^k)^{\lambda_k}] \end{aligned} \tag{8}$$

Where w_j is a intuitionistic fuzzy number, and then, the specific weight corresponding to each criterion A_i is calculated using the equation 5, previously used. Later, the vector of weights obtained $W = \{w_1, w_2, ..., w_m\}$, must satisfy the following equations:

$$w_j > 0 \quad (j = 1, 2, ..., m) \tag{9}$$

$$\sum_{j=1}^{m} w_j = 1 \tag{10}$$

Third, each of the alternatives is rated individually by each of the DMs, using the linguistic terms and diffuse intuitionist numbers shown in Table IV.

TABLE IV. LINGUISTIC TERMS TO EVALUATE THE ALTERNATIVES (LEARNING RESOURCES) [45]

Linguistic terms	IFN (μ, υ, π)
Extremely Bad (EB)	(0.1,0.9,0)
Very Bad (VB)	(0.1,0.75,0.15)
Bad (B)	(0.25,0.6,0.15)
Medium Bad (MB)	(0.4,0.5,0.1)
Fair (F)	(0.5,0.4,0.1)
Medium Good (MG)	(0.6, 0.3, 0.1)
Good (G)	(0.7, 0.2, 0.1)
Very Good (VG)	(0.8,0.1,0.1)
Excellent (E)	(1, 0, 0)

Then, using the sets defined above, the intuitionist fuzzy decision matrix of each DM (IFDM) is constructed, denoted as $R^k = [r_{ij}^k]_{nxm}$, and subsequently all opinions of DMs are agregated using the IFWA operator (Equation 11), giving rise to the Aggregate Weighted Intuitionistic Fuzzy decision matrix (AWIFDM), which represents the evaluations, based on the opinions of the DMs, for each of the alternatives (resources) to evaluate O_n.

$$\begin{aligned} R &= (r_{ij})_{nxm} \\ r_{ij} &= IFWA(r_{ij}^1, r_{ij}^2, ..., r_{ij}^l) \\ &= \lambda_1.r_{ij}^1 \oplus \lambda_2.r_{ij}^2 \oplus ... \oplus \lambda_l.r_{ij}^l \\ &= [1 - \prod_{k=1}^{l}(1 - \mu_{ij}^k)^{\lambda_k}, \prod_{k=1}^{l}(\upsilon_{ij}^k)^{\lambda_k}, \\ &\quad \prod_{k=1}^{l}(1 - \mu_{ij}^k)^{\lambda_k} - \prod_{k=1}^{l}(\upsilon_{ij}^k)^{\lambda_k}] \end{aligned} \tag{11}$$

After which, the intuitionistic final decision matrix is defined as:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{bmatrix} \tag{12}$$

Or more specifically according to the form shown in the equation 13.

$$R = \begin{bmatrix} (\mu_{c_1}, \upsilon_{c_1}, \pi_{c_1})_(x_1) & \cdots & (\mu_{c_1}, \upsilon_{c_1}, \pi_{c_1})_(x_m) \\ \vdots & \ddots & \vdots \\ (\mu_{c_n}, \upsilon_{c_n}, \pi_{c_n})_(x_1) & \cdots & (\mu_{c_n}, \upsilon_{c_n}, \pi_{c_n})_(x_m)) \end{bmatrix} \tag{13}$$

Then the ideal solution $S = (x_1^*, x_2^*, ..., x_m^*)$ for the proposed problem is calculated, considering that, according to the nature of the problem, all the attributes belong to the category commonly referred to as benefits, and there is no cost category, so:

$$\mu_s(x_j) = (max\ \upsilon_{c_i}(x_j)) \tag{14}$$

$$\upsilon_s(x_j) = (min\ \upsilon_{c_i}(x_j)) \tag{15}$$

$$\pi_s(x_j) = 1 - (\mu_s(x_j) + \upsilon_s(x_j)) \tag{16}$$

Then, the Intuitionistic Fuzzy Index of Similarity (IFIS) of the alternative C_i is calculated with respect to the ideal solution S, using the equation 17.

$$\begin{aligned} IFIS_i &= \sqrt[W]{\prod_{j=1}^{m}\left\{\frac{x_j^i}{s_j}\right\}^{W_j}} \\ &= \prod_{j=1}^{m}\left\{\frac{\mu_{c_i(x_j)}, \upsilon_{c_i(x_j)}, \pi_{c_i(x_j)}}{\mu_{s(x_j)}, \upsilon_{s(x_j)}, \pi_{s(x_j)}}\right\}^{W_j} \end{aligned} \tag{17}$$

Where:

$IFIS_i$ = Intuitionistic Fuzzy Index of Similarity of the alternative i.

s_j = Value of the ideal solution for criterion j

x_j^i = Value of criterion j for alternative i

w_j = The weight corresponding to criterion j

Finally, the similarity index (IS) of a candidate c_i with respect to the IFIS is calculated, using the equation 18 proposed by [57].

$$IS_i = 1 - \frac{1 - \mu_{IFIS_i}}{1 + \pi_{IFIS_i}} \tag{18}$$

Where:

IS_i = Similarity index of the candidate i

μ_{IFIS_i} = Truth value of the candidate i with respect to his IFIS

$\pi_{[}IFIS_i]$ = Uncertainty value of the candidate i regarding his IFIS

The highest value of IS (closer to 1), indicates the learning object that best suits the preferences and needs of a specific student at a given time, within a given topic or unit of learning, so it would be enough for the LMS to select and display the learning objects with the highest value of IS, being able to be limited to a maximum number of resources shown or through a threshold for the value of the IS.

IV. Conclusion

The proposed definition of the student profile, which considers different pedagogical approaches, cognitive strategies, and other attributes, makes possible the application of different intelligent user analysis techniques, towards the personalization of contents within LMS or similar environments.

The application of MADM techniques, mostly used in the field of administration and finance, in the evaluation and selection of learning objects, is a great help for the customization of content within LMS platforms, as long as there is enough information (variety of criteria) for the correct use of these techniques.

Intuitionistic Fuzzy Numbers represent a mechanism with a lot of potential in solving problems with a high degree of uncertainty and multiple criterias that influence the decision, such as the analysis of the user profiles and the application proposed, due to the complexity of human nature and of the related cognitive processes.

Finally, it is considered that, the use of the student profile defined in this investigation, together with MADM and IFN techniques, under the proposed method, constitute a good alternative for the selection and customization of contents within any LMS, which will allow contribute to improve the efficiency of the learning process and enrich the user experience.

The approach presented in this paper is initially oriented towards traditional LMS, however, there is the possibility of applying it in future works with other technologies, such as: gamification models in e-Learning environments, interactive e-learning with virtual reality, STEM e-Learning in an Immersive VR environments, among others.

Acknowledgment

The author would like to thank the Universidad Nacional de San Agustín de Arequipa - Perú (UNSA), for supporting this research.

References

[1] T. Harraf, "Effects of four varied teaching pedagogies on freshmen student success," *Proceedings of International Academic Conferences*, jun 2019.

[2] S. M. Nafea, F. Siewe, and Y. He, "A novel algorithm for dynamic student profile adaptation based on learning styles," in *Advances in Intelligent Systems and Computing*, pp. 29–51, Springer International Publishing, aug 2019.

[3] N. A. Ali, F. Eassa, and E. Hamed, "Adaptive e-learning system based on personalized learning style," *Journal of Fundamental and Applied Sciences*, vol. 10, pp. 246–251, Mar. 2018.

[4] I. Azzi, A. Jeghal, A. Radouane, A. Yahyaouy, and H. Tairi, "A robust classification to predict learning styles in adaptive e-learning systems," *Education and Information Technologies*, aug 2019.

[5] J. Bernard, T.-W. Chang, E. Popescu, and S. Graf, "Learning style identifier: Improving the precision of learning style identification through computational intelligence algorithms," *Expert Systems with Applications*, vol. 75, pp. 94–108, jun 2017.

[6] Özcan Özyurt and H. Özyurt, "Learning style based individualized adaptive e-learning environments: Content analysis of the articles published from 2005 to 2014," *Computers in Human Behavior*, vol. 52, pp. 349–358, nov 2015.

[7] L. Alfaro, C. Rivera, J. Luna-Urquizo, E. Castaneda, and F. Fialho, "Utilization of a neuro fuzzy model for the online detection of learning styles in adaptive e-learning systems," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 12, 2018.

[8] J. Aguilar, O. Buendia, A. Pinto, and J. Gutiérrez, "Social learning analytics for determining learning styles in a smart classroom," *Interactive Learning Environments*, pp. 1–17, aug 2019.

[9] A. Seyal, N. Z. Siau, and W. S. H. Suhali, "Evaluating students' personality and learning styles in higher education: Pedagogical considerations," *International Journal of Learning, Teaching and Educational Research*, vol. 18, pp. 145–164, jun 2019.

[10] N. Hamdaoui, M. K. Idrissi, and S. Bennani, "Modeling learners in educational games: Relationship between playing and learning styles," *Simulation & Gaming*, vol. 49, pp. 675–699, jul 2018.

[11] N. Jalaledin, M. A. MAHMOOD, and F. Somayeh, "Classification of learning styles using behavioral features and twin support vector machine," *JOURNAL OF TECHNOLOGY OF EDUCATION*, vol. 13, no. 3, pp. 459–469, 2019.

[12] K. Crockett, A. Latham, and N. Whitton, "On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees," *International Journal of Human-Computer Studies*, vol. 97, pp. 98–115, jan 2017.

[13] M. A. Palomino, M. Strefezza, and L. Contreras, "Sistema difuso para la detección automática de estilos de aprendizaje en ambientes de formación web," *Ciencia, Docencia y Tecnología*, vol. 27, pp. 269–294, May 2016.

[14] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," *Computers in Human Behavior*, vol. 55, pp. 1185–1193, feb 2016.

[15] A. M. Brown and A. N. Kaminske, *Five Teaching and Learning Myths—Debunked*. Routledge, apr 2018.

[16] P. A. Kirschner, "Stop propagating the learning styles myth," *Computers & Education*, vol. 106, pp. 166–171, 2017.

[17] S. Moser and J. Zumbach, "Exploring the development and impact of Learning Styles: An empirical investigation based on explicit and implicit measures," *Computers & Education*, vol. 125, pp. 146–157, oct 2018.

[18] P. Newton, "Learning styles and evidence-based approaches to education," *FASEB Journal*, vol. 33, apr 2019.

[19] M. Papadatou-Pastou, M. Gritzali, and A. Barrable, "The learning styles educational neuromyth: Lack of agreement between teachers judgments, self-assessment, and students intelligence," *Frontiers in Education*, vol. 3, nov 2018.

[20] R. Wang and C. Liu, "The relation of dental students' learning styles to their satisfaction with traditional and inverted classroom models," *BMC Medical Education*, vol. 19, aug 2019.

[21] D. T. Willingham, "Ask the cognitive scientist: Does tailoring instruction to "learning styles" help students learn?," *American Educator*, vol. 42, no. 2, pp. 28–32, 2018.

[22] S. Cassidy, "Learning styles: An overview of theories, models, and measures," *Educational Psychology*, vol. 24, pp. 419–444, aug 2004.

[23] J. Moussa-Inaty, , F. Atallah, M. Causapin, and and, "Instructional mode: A better predictor of performance than student preferred learning styles," *International Journal of Instruction*, vol. 12, pp. 17–34, jul 2019.

[24] Ünal Çakiroğlu, M. Güler, M. Atabay, and M. Güler, "Connections between learning styles and perceived cognitive load in multimedia learning: An experimental study," *Journal of Educational Technology Systems*, p. 004723951984450, apr 2019.

[25] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[26] S. Kalyuga and A.-M. Singh, "Rethinking the boundaries of cognitive load theory in complex learning," *Educational Psychology Review*, vol. 28, pp. 831–852, dec 2015.

[27] K. L. Fraser, P. Ayres, and J. Sweller, "Cognitive load theory for the design of medical simulations," *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, vol. 10, pp. 295–307, oct 2015.

[28] L. M. Naismith and R. B. Cavalcanti, "Validity of cognitive load measures in simulation-based training: a systematic review," *Academic Medicine*, vol. 90, no. 11, pp. S24–S35, 2015.

[29] A. Korbach, R. Brunken, and B. Park, "Measurement of cognitive load in multimedia learning: a comparison of different objective measures," *Instructional Science*, vol. 45, pp. 515–536, apr 2017.

[30] C.-C. Chang, C. Liang, P.-N. Chou, and G.-Y. Lin, "Is Game-based learning better in flow experience and various types of Cognitive Load than non-game-based learning? Perspective from Multimedia and Media richness," *Computers in Human Behavior*, vol. 71, pp. 218–227, 2017.

[31] S. Sepp, S. J. Howard, S. Tindall-Ford, S. Agostinho, and F. Paas, "Cognitive load theory and human movement: Towards an integrated model of working memory," *Educational Psychology Review*, vol. 31, pp. 293–317, feb 2019.

[32] H. Järvenoja and S. Järvelä, "Emotion control in collaborative learning situations: Do students regulate emotions evoked by social challenges," *British Journal of Educational Psychology*, vol. 79, no. 3, pp. 463–481, 2009.

[33] G. Matthews and S. E. Campbell, "Sustained performance under overload: Personality and individual differences in stress and coping," *Theoretical Issues in Ergonomics Science*, vol. 10, no. 5, pp. 417–442, 2009.

[34] A. Solimeno, M. E. Mebane, M. Tomai, and D. Francescato, "The influence of students and teachers characteristics on the efficacy of face-to-face and computer supported collaborative learning," *Computers & Education*, vol. 51, no. 1, pp. 109–128, 2008.

[35] O. C. Santos, "Emotions and personality in adaptive e-learning systems: An affective computing perspective," in *Human–Computer Interaction Series*, pp. 263–285, Springer International Publishing, 2016.

[36] J. D. Plummer and J. Krajcik, "Building a learning progression for celestial motion: Elementary levels from an earth-based perspective," *Journal of Research in Science Teaching*, vol. 47, no. 7, pp. 768–787, 2010.

[37] L. A. Annetta, R. Lamb, D. Vallett, and M. Shapiro, "Project-based learning progressions: Identifying the nodes of learning in a project-based environment," in *Contemporary Technologies in Education*, pp. 163–181, Springer, 2019.

[38] N. J. Brown and M. Wilson, "A model of cognition: The missing cornerstone of assessment," *Educational Psychology Review*, vol. 23, no. 2, p. 221, 2011.

[39] Y. Melia, "Multi attribute decision making using simple additive weighting and weighted product in investment," *International Academic Journal of Business Management*, vol. 3, no. 7, pp. 1–55, 2016.

[40] S. Fenny, A. Nursikuwagus, and T. Hartono, "Rank computation model for distribution product in fuzzy multiple attribute decision making," *TELKOMNIKA*, vol. 16, no. 1, pp. 274–281, 2018.

[41] A. Rucitra, "Application of multi attribute failure mode analysis of milk production using analytical hierarchy process method," in *IOP Conference Series: Earth and Environmental Science*, vol. 131, p. 012022, IOP Publishing, 2018.

[42] N. Kumar, T. Soota, N. Gupta, and S. K. Rajput, "Multi attribute outranking approach for supplier selection," in *IOP Conference Series: Materials Science and Engineering*, vol. 404, p. 012008, IOP Publishing, 2018.

[43] Budiharjo, A. P. Windarto, and A. Muhammad, "Comparison of weighted sum model and multi attribute decision making weighted product methods in selecting the best elementary school in indonesia," *International Journal of Software Engineering and Its Applications*, vol. 11, no. 4, pp. 69–90, 2017.

[44] A. S. Purnomo and A. F. Rozi, "Comparative analysis uses weighted product (WP) and simple additive weighting (SAW) methods in the best graduation selection system," *Sebatik Journal*, vol. 23, no. 1, pp. 44–52, 2019.

[45] L. Pérez, A. Alvarado, J. L. García, and O. O. Vergara, "Selección de proveedores con análisis dimensional difuso intuicionista," *Cultura Cientifica y Tecnológica - CULCyT*, vol. 12, pp. 136–145, Dec. 2015.

[46] M. Keshavarz Ghorabaee, M. Amiri, E. K. Zavadskas, and J. Antucheviciene, "Supplier evaluation and selection in fuzzy environments: A review of madm approaches," *Economic research-Ekonomska istraživanja*, vol. 30, no. 1, pp. 1073–1118, 2017.

[47] G. Wei and J. M. Merigó, "Methods for strategic decision-making problems with immediate probabilities in intuitionistic fuzzy setting," *Scientia Iranica*, vol. 19, no. 6, pp. 1936–1946, 2012.

[48] E. Kurilovas, "Advanced machine learning approaches to personalise learning: learning analytics and decision making," *Behaviour & Information Technology*, vol. 38, pp. 410–421, nov 2018.

[49] P. Wessa, S. Poelmans, K. Milis, and E. Van Stee, "Modeling educational technology acceptance and satisfaction," *Proceedings of EDULEARN09 (on cd-rom)*, 2010.

[50] P. Moore, Z. Zhao, and H. V. Pham, "Towards cloud-based personalised student-centric context-aware e-learning pedagogic systems," in *Advances in Intelligent Systems and Computing*, pp. 331–342, Springer International Publishing, jun 2019.

[51] Á. Torres Toukoumidis, L. M. Romero Rodríguez, and M. A. Pérez Rodríguez, "Modelo teórico integrado de gamificación en ambientes e-learning (e-miga).," *Revista Complutense de Educación*, vol. 29, no. 1, pp. 129–145, 2018.

[52] N. P. Doe, R. John, M. Sitti, and N. Thirunavukkarasu, "Interactive e-learning with integrated virtual reality," *Journal of Innovation in Computer Science and Engineering (JICSE)*, vol. 8, no. 1, pp. 1–5, 2018.

[53] C. B. Rogers, H. El-Mounayri, T. Wasfy, and J. Satterwhite, "Assessment of stem e-learning in an immersive virtual reality (vr) environment," ASEE, 2018.

[54] C. wen Shen, J. tsung Ho, P. T. M. Ly, and T. chang Kuo, "Behavioural intentions of using virtual reality in learning: perspectives of acceptance of information technology and learning style," *Virtual Reality*, vol. 23, pp. 313–324, may 2018.

[55] F. E. Boran, S. Genç, M. Kurt, and D. Akay, "A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method," *Expert Systems with Applications*, vol. 36, pp. 11363–11368, oct 2009.

[56] Z. Xu, "Intuitionistic fuzzy aggregation operators," *IEEE Transactions on Fuzzy Systems*, vol. 15, pp. 1179–1187, dec 2007.

[57] S.-f. Zhang and S.-y. Liu, "A gra-based intuitionistic fuzzy multi-criteria group decision making method for personnel selection," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11401–11405, 2011.

A Closer Look at Arabic Text Classification

Mohammad A.R. Abdeen[1], Sami AlBouq[2]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah, Saudi Arabia

Ahmed Elmahalawy[3]
Faculty of Engineering
Menoufia University
Menoufia, Egypt

Sara Shehata[4]
Faculty of Computer Science
Ain-Shams University
Cairo, Egypt

***Abstract*—The world has witnessed an information explosion in the past two decades. Electronic devices are now available in many varieties such as PCs, Laptops, book readers, mobile devices and with relatively affordable prices. This and the ubiquitous use of software applications such as social media and cloud applications, and the increasing trend towards digitalization, the amount of information on the global cloud has surged to an unprecedented level. Therefore, a dire need exists in order to mine this massively large amount of data and produce meaningful information. Text Classification is one of the known and well established data mining techniques that has been used and reported in the literature. Text classification methods include statistical and machine learning algorithms such as Naive Baysian, Support Vector Machines and others have widely been used. Many works have been reported regarding text classification of various languages including English, Chinese, Russian, and many others. Arabic is the fifth most spoken language in the world. There has been many works in the literature for Arabic text classification. However, and to the best of our knowledge, there is no recent work that presents a good, critical and comprehensive survey of the Arabic text classification for the past two decades. The aim of this paper is to present a concise and yet comprehensive review of the Arabic text classification. We have covered over 50 research papers covering the past two decades (2000 - 2019). The main focus of this paper is to address the following issues: 1) The techniques reported in the literature including. 2) New Techniques. 3) Most claimed efficient technique. 4) Datasets used and which ones are most popular. 5) Which feature selection techniques are used? 6) Popular classes/categories used. 7) Effect of stemming techniques on classification results.**

***Keywords*—Arabic text classification; support vector machines; k-NN; Naive Bayesian; decision trees; C4.5; maximum entropy; feature selection; Arabic dataset**

I. Introduction

The 21st century is truly the information age. With the introduction of ubiquitous computing via electronic/mobile devices in an unprecedented way in just about all aspects of live resulted in the production of data in Zetta scales. Today, one can move across the four corners of the earth without carrying a dime of cash. Also we can communicate live with video with just about anyone that has an average quality mobile phone. People are exchanging messages, documents, recipes, life experiences, medical advice, all through the now fully connected world.

The resulting mass of data needs to be analyzed and classified so that useful information is extracted. Data mining is an area of computer science that deals with developing ways to extract information out of existing masses of data. One of data mining applications is text classification (TC). Text classification is a data mining technique that is used to assign a given document in a set of documents to a given class or category.

There has been many works in the literature that addresses the topic of text classification in various languages such as English [1], Chinese [2], Russian [3] and many other languages.

Arabic language is one of the most widely spoken languages in the world. It is the fifth most spoken language in the word and the fifth most used language on the Internet. More than 6.0% of the world's population speaks Arabic language (more than 422 million speakers) [4]. Arabic language has rich morphology and an intricate orthography. The span of highlight feature vectors may increment and that make the undertaking of highlight choice progressively imperative to maintain a strategic distance from the insignificant data [5] that produces different words with different meanings. On contrary of Arabic text, there are many benchmarking corpus that can be used for Latin base language, Japanese and Chinese text classification [6] [7]. This language has specific letters known as Arabic vowels (waw, yaa, alf) that require a special system of morphology and grammars. The others are called consonant letters [8]. What also distinguishes Arabic is the huge amount of vocabularies and concepts [9].

There are 28 letters in Arabic language, in addition to the Arabic hamza (ء) which is considered as a letter by some Arabic linguistics and it is written from right to left. It has two genders: feminine and masculine. Numerical are singular, dual, and plural numbers. Grammatical are three cases: nominative, accusative, and genitive. A noun has three linguistic cases: nominative case when it is subject; accusative case when it is the object of a verb; and the genitive case when it is the object of a preposition [10] [11].

Arabic doesn't employ an upper/lower case. It also employs diacritics which represent a small vowel letters such as "fatha, kasra, damma, sukun, shadda, and tanween" [9].

The process of Text-Classification is concerned with assigning a given document or a group of documents to a given class or category. As an example, the task of classifying various news of a news website to several categories (Weather, Politics, Sport...). The explosion of Internet contents in various languages including Arabic presented a pressing need for effective classification. Classification algorithms are developed for this sole purpose which are called Classifier) [12] [13].

The remainder of the paper is organized as follows: Section II presents a short outline of Text Classification. Section III presents a brief description of the process of Arabic text classification stages. Section IV represents an overview of the previous work. Section V presents a tabulated summery of survey results. Section VI provides conclusion and recommendation.

II. Overview of Text Classification

Text Classification (TC) is the process of assigning given text in a document to preset categories [14]. Another definition describes the classification process as an assignment of category labels to natural language documents with the possibility that a one document be included in more than one category [15].

Manual TC is the process of classifying documents by some trained individuals. This process is time consuming and is prone to human errors. In addition, with the latest surge in data sizes and variety of classes of documents, the manual process is certainly not scalable and impractical [16] [17].

Automatic TC is defined as the assignment of documents in a collection to a predefined class or category. Therefore, text classification can be achieved through machine learning techniques by training the algorithms with a training dataset [18] [17]. Automatic TC approaches have been reported in the literature during the past three decades. Numerous algorithms have been introduced and evaluated in some surveys [19] that address the most common TC algorithms and a lot of works have been achieved to evaluate and compare between these algorithms.

Text classification can be performed with two main techniques/methodologies; the statistical techniques, and the machine learning techniques. The following two sections presents some details about those techniques.

A. Statistical Techniques

Statistical text classification techniques are based on mathematical foundations. These techniques have been developed relatively earlier than the machine learning techniques and are more suitable for relatively small datasets. Some of them are also more suitable for binary classifications rather than multiclass classification. Examples of those techniques are: the Frequentist procedures, Bayesian procedures, and the Binary and multiclass procedures.

B. Machine Learning Techniques

Due to the surge in the size of data for the past two decades, automation process is required to achieve the goals of information extraction and classification/clustering of data for a variety of purposes. Those include email filtering and routing; news observing; Spam filtering and search engines [20]; newsgroups classification, and survey data grouping [17]. Depending on the nature of the available data, machine learning can be classified to three main categories [10] [21].

1) Supervised learning where training of the model is required a priori using a previously labeled data. Data labeling can be a difficult task when data size is significantly large as it is done by humans. Some of the known supervised machine learning techniques include Naive Bayes, Logistic Regression, Support Vector Machine, Artificial Neural Networks and Decision Trees.

2) Unsupervised learning where initial training data is not required. The model rather groups the data in clusters (clustering) without labeling based on some feature similarity. Principal component analysis and Self-Organizing Maps (SOM) are some of the popular unsupervised learning algorithms.

3) Semi-Supervised machine learning: is a combination of the above mentioned two techniques. It employs a relatively small amount of labeled data with a significantly larger one of unlabeled data. Examples of the semi-supervised learning approaches are generative and the graph-based models.

There are two issues of Arabic language are the high dimensionality of the feature space and the rate of the precision is approximately low. The complexity of many learning algorithms increases in parallel with the increase in data dimension.

There are several reports in the literature that claim to present a review study on Arabic text classification. As an example, the work presented in [9] claims to be survey of the Arabic text classification. The author however provided a very broad background and definition of the term and gave a brief background of the popular four techniques, i.e. the Naïve Bayesian, the K-Nearest Neighbor, the Support Vector Machine and the Artificial Neural Network. The author did not provide a detailed or comprehensive treatment of the subject nor of the various techniques or algorithms used nor the results of those algorithms.

The nature of the data source may influence the execution of a classification algorithm; the insignificant and repetitive highlights of data may lessen the nature of the outcome [22]. The span of feature vectors may increment by wealth of the language that make task of feature selection vital to stay away from the immaterial component [5].

For Arabic language there is a lack of the studies on the classification of Arabic text documents with limitation of free benchmarking dataset [6] [23]. On the other hand, the richness in morphology of Arabic language significantly increases the length of the feature vector and that significantly influenced research and studies in the field of text classification [24].

The research deanship of Islamic University of Madinah.

There are three principle stages for the text classification as shown in Fig. 1 [24] [15]:

- Data pre-processing
- Text classification
- Evaluation

Classifying Arabic documents requires accomplishing some preprocessing steps for the documents through stemming the words; this process is quite a major issue in terms of reducing the number of related words in a document.

The text classification can be divided into other sub problems that have been investigated in the literature, for example, the document indexing, the weighting assignment, the document grouping, the dimensionality reduction, the threshold determination and the types of classifiers.

Document indexing is related with the method for extracting document's keywords. There are two fundamental ways to deal with the document indexing: the first methodology considers list terms as bag-of-words [25] and the second sees the file terms as phrases [7] [17] [26] [27].

A drawback of the first approach is that it complicates the extraction process of index term by increasing the number of words that must be dealt with in the document as well as dealing with irrelevant words (unrelated to any category). Classifying Arabic documents requires achieving some preprocessing steps for the documents through stemming the words, this process is quite significant in terms of eliminating the number related words in a document. The preprocessing of our dataset was an important step as it increased the accuracy of the classification and reduced the required memory size for the classification process.

Several techniques have been introduced to perform preprocessing tasks such as stemming, root extraction and thesaurus. Weight assignment procedures associate a real number that ranges from 0-1 for every term in the collection of documents [19], the loads will be required to classify new arrived documents. Different information retrieval models use different methodologies to compute these weights, for example the Boolean model assigns either 0 or 1 for each index term.

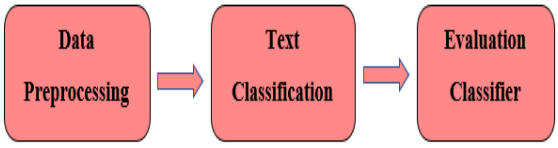


Fig. 1. Text Classification Stages.

Historically, the most widely famous Naïve Bayesian model was known as the binary independent classifier [26]. One of the first attempts to address clustering techniques for TC problem was stated through introducing a comparison between the k-means algorithms and hierarchical clustering algorithms [25]. The results showed better performance for the hierarchical algorithms although they were slower than the k-means algorithm. Distinctive TC strategies have been raised to categorize documents, for example, K-nearest neighbor KNN [28], the K-NN different models compute the separations between the document index terms and the known terms of every category by applying distance functions, for example, cosine, dice similarity or Euclidean functions, the returned classes are the kth classes with elevated scores.

III. Arabic Text Classification Stages

Text classification in general has five main stages: Data collection, Data preprocessing, Feature extraction/selection, Text classification, and Classifier evaluation [7]. Fig. 2 depicts those stages [29] [30].

Those main stages can be further broken to more fine grain tasks as given in the below list [23]:

- Data gathering;
- Data labeling;
- Data cleaning (removal of stop word, stemming);
- Feature extraction;
- Feature Selection;
- Classification Model Training;
- Classification algorithm testing.

1) Data Gathering: There are numerous standard data sets for English text classification that freely accessible. Unfortunately for Arabic language, we are not aware of a open access standard dataset [7]. The Open Source Arabic Corpus [31] is freely available but is not standardized. The majority of the researchers in the field of Arabic text classification built their test corpus from online Arabic news sites [32]. The scope of the chosen documents that introduced in the dataset varied from as low as 240 documents split into six categories [23] up to 17,658 documents partitioned into seven genres [33].

2) Data Pre-processing (labeling and cleaning): The purpose of this stage is to remove words that do not contribute to the semantics of the document such as stop words. Example of these words are the pronouns. This stage also removes suffixes and prefixes. It also combines words of the same root/origin. The main objective of this stage is to reduce the feature set of a given document and provide better classification accuracy.

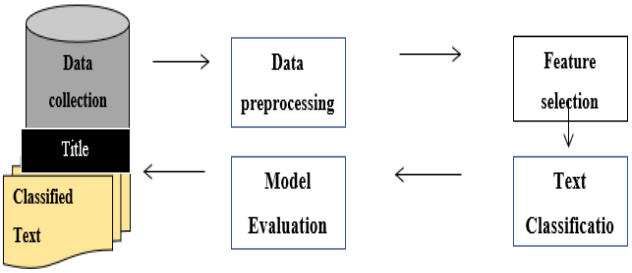


Fig. 2. Detailed Text Classification Stages.

To accommodate various writing styles, some normalization is also performed. This includes Hamza " ء " Taa Marboutah " ة", Alif " ا" and the Yaa "ى" . Combining words of similar root/origin is performed by stemming algorithms. There are three main types of stemming algorithms: the root-based stemmer; the light stemmer; and the statistical N-gram algorithm. These types of stemmers are summarized below.

1) Light Stemmer: In this method, the commonly used affixes (prefixes, infixes and suffixes). The commonly used affixes and stored in a separate file and are used in stem extraction [34]. This method is relatively easy and contributes positively to the efficiency of the classification algorithm.

2) Root Stemmer: In this method, the actual root of the word is found by applying some morphological rules. These rules are largely language dependent. A variant of this approach requires to build and search a lexicon, which is very time consuming and inefficient [35]. Another approach used by Al-Serhan [36] did not require a lexicon or to store roots in a separate file.

3) N-gram Stemmer: This stemming technique is classified as a statistical and is language independent. The algorithm forms a gram of two (bigram), or more, consecutive letters. The common di-gram between words is used to calculate a metric of similarity referred to as the Dice's coefficient [37].

One significant difference between the text classification of Arabic language and other commonly used languages such as English is the fact that stemming could significantly affect the result of the classification [38] [11] [39] [40]. Arabic is a highly inflectional and morphologically rich language and Arabic wards can come from a stem of three, four, five, and six letter words. Almost 80% of the Arabic words come from a three-letter root [38].

The work by [38] studied the effect of various stemming techniques on the accuracy (precision and recall) of the K-Nearest Neighbor (KNN) classifiers. The study concluded that light stemmer produces the best results as compared to root stemmer and statistical stemmers.

A. Feature Extraction and Selection

Two important states of text preprocessing prior to classification is feature extraction and feature selection. The process pf feature extraction is concerned with transforming the unstructured text into a structured representation and for removing redundancy which facilitates further processing and the application of machine learning techniques. Feature selection is an even further preprocessing step to exclude irrelevant features and reduce the high dimensionality of the result of the feature extraction prior step. There are three main categories of feature selection algorithms: the Wrapper, the Filter, and the Embedded [41]. Some of the most commonly used feature extraction is Chi-square, correlation coefficient scores, information gain, recursive feature elimination, and the Least Absolute Shrinkage and Selection Operator. Another approach for feature selection is using Rough Sets (RS) in various languages including Arabic [5] [42] [43] with satisfactory results.

B. Text Classification Methods and Algorithms

Text classification algorithm are the following: Decision Trees: which are used to classify documents through building a tree by computing the entropy function of the selected index terms such as ID3 and C4.5. Naive Bayesian models which have portrayed great outcomes in the text classification field. Historically, the most widely famous Naïve Bayesian model was known as the binary independent classifier. In the Clustering Techniques, the issue was stated through introducing a comparison between the k-means algorithms and hierarchical clustering algorithms. The conducted results showed better performance for the hierarchical algorithms although they were slower than the k-means algorithm. In the Distinctive text classification, strategies have been developed to sort documents, for example, K-Nearest Neighbor KNN, the K-NN different models compute the distances between the document index terms and the known terms of each category by applying distance functions, for example, cosine, dice similarity or Euclidian functions, the returned classes are the kth classes with most noteworthy scores. Support Vector Machine Support vector machines (SVMs) are considered one of the most well-known text classifiers. SVMs are one of the supervised machine learning techniques. In SVMs a training algorithm is used to build a model that will be used to assign a new unknown document to one category from a set of predefined categories. SVMs can be used to perform a linear and a non-linear classification. Fig. 3 explained Linear SVM vs Nonlinear SVM [44].

C. Model Evaluation

Text classification is assessed based on the efficiency and effectiveness of categorization. There are some techniques that have been utilized to quantify the advancement of the classifier. One of those techniques is F1, precision and recall that are utilized in the field of information retrieve and machine learning. There are different types of measurements to test the classifiers and this may not justify the result such as:

- F1-precision measure.
- Fallout and error rate as accuracy measure.
- K-fold cross-validation technique utilized to test the precision.

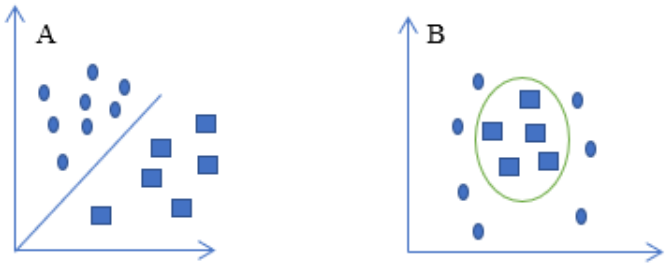


Fig. 3. Linear SVM (Left) vs. Nonlinear SVM (Right).

IV. Previous Work

There are some reports in the literature claiming to provide a review of the Arabic text classification work to date. To the best of our knowledge, none so far is reviewing the subject from the viewpoints considered in this research nor the works reported present a comprehensive treatment of the subject.

As an example, in [45], despite the fact that the author claims to present a review of the TC topic, they reviewed only four papers [46] [17] [32]. These works are relatively old and there has been numerous other works in the past two decades that the author did not include in their review.

The work presented by [29] was a relatively short overview of previous work and most of the cited work is a decade old. Despite the fact that authors cited various techniques used in classification process, they missed some significant recent works and did not consider and did not consider the effect of various stemming techniques on the accuracy of the result.

In [44] despite the fact that the title is claiming to review the topic, in fact the author presented a comparative study of Arabic text classification using various techniques such as NB, KNN, SVM and Artificial Neural Networks. The authors used their own dataset of size 4000 documents collected from Arabic websites.

There are many researches in classifying English documents. In addition, there are numerous researches in European languages, for example, German, Italian and Spanish, Asian languages as well, for example, Chinese and Japanese. In the past decade, the work on Arabic text classification started with four primary researches [47]:

1) El-Kourdi et. al. [17], used naïve bayes (NB) algorithm to automatic Arabic document classification. The average accuracy revealed was approximately 69%.

2) An Named Entity Recognition (NER) system is called Siraj from Sakhr [48]. The framework has no specialized documentation to clarify the strategy utilized in the framework and the exactness of the framework.

3) Sawaf et. al. [46] introduced a framework for statistical classification methods strategies, such as maximum entropy to classify and cluster news articles. The best classification accuracy obtained was approximately 63% with a precision of 50% which is a low reported precision in the research area.

4) El-Halees [49] depicted a technique that depends on affiliation principles to classify Arabic documents. The classification precision revealed was 74.41%.

The majority of authors used an in-house built Arabic dataset with various sizes and contents, with the Internet websites as their major source of data [50] [40]. The examinations in the field of Arabic text classification utilized Naïve Bayes [17] [51], Support vector machine [52], Decision Trees [24] as classifier algorithm. In [24], the author gathered in-house Arabic corpus that comprises of 1000 documents, were separated into 10 categories and utilized it to compare the performance of three Arabic text classifiers: Naïve Bayes, K-Nearest Neighbor and Distance Based. The recall, precision, error rate and fallout are utilizing to compare the accuracy of the classifiers. The data was preprocessed by expelled the stop words and extricated the root of the words. The results of the experimentation demonstrate that the Naïve Bayes classifier beats the other two (over 95%).

In [53] the authors' main objective was to classify the Internet content of the Arabic text. They used Internet data of size 40 Gigabytes and classified them into five categories, namely, religion, politics, culture, sports, and economics. They used the NB and K-NN machine learning algorithms.

In [16], the authors utilized in house corpus comprises of 242 documents which have a place with six categories to assess the execution of two classifiers K-NN and Naive Bayes to classify the Arabic content. The k-fold cross-validation strategy is utilized to test the accuracy. They extricated feature set of keywords to improve the execution. The outcome demonstrated that K-NN classifier would do well to execution.

In [26], the authors examined Naïve Bayesian and Support Vector Machine on various Arabic data indexes. The Exploratory outcomes against various Arabic text categorization data sets uncover that SVM algorithm outflanks the NB. While in [15] which thought about the Support vector machine and C5.0 classifier and C5.0 classifier gives better precision. By [23], Support vector machine had demonstrated the predominance in highlight determinations, weighting techniques, and classification algorithms, trailed by the decision tree algorithm (C4.5) and Naive Bayes. The best classification precision was 97% for the Islamic Themes dataset, and the least exact was 61% for the Arabic Poems dataset. Duwairi et al. [38], the examination makes a comparison between (stemming, light stemming, and word cluster). For training purposes, they pick K Nearest Neighbor KNN method, to demonstrate that light stemming accomplishes best performance and least time of model development. Another study [18] looked at 3 Feature Subset Selection (FSS) measurements. They did a relative report to look at the impact of the component choice measurements as far as accuracy. The results in general revealed that Odd Ratio (OR) worked superior to the others. A few examinations concentrated on different procedures like N-gram and distinctive distance measures and demonstrated their impacts on Arabic TC. El-Kourdi et al. [17] classified Arabic text documents automatically utilizing NB. The normal precision revealed was about 68.78%, and the best accuracy reported was about 92.8%. El-Kourdi utilized a corpus of 1500 text documents having a place with five categories; each category contains 300 text documents. All words in the documents are changed over to their foundations. The vocabulary size of

resultant corpus is 2,000 terms/roots. Cross-validation was utilized for assessment.

Sawaf et al. [46] utilized Maximum entropy (ME) to make a grouping to News articles. The investigation gives precision about 62.7%. Al-Zoghby [32] utilized Association Rules for Arabic text classification, furthermore, he utilized CHARM algorithm with soft-matching over hard big O exact matching. Data indexes comprise of 5524 records. Each record is a scrap of emails having the subject - nuclear. The vocabulary size is 103,253 words. Harrag et al. [54] utilized the feature selection dependent on hybrid approach for Arabic text classification. He utilized direct tree algorithm and the accuracy was of 93% for scientific data set, and 90% for literary data-set. Harrag gathered two data indexes; the first one is from the scientific encyclopedia. In [29], the authors addressed the issues of lacking free open Arabic corpora. In [30] centered around the distributed researches in the field of Arabic text classification and illustrates a logical view its procedure and camper the assessment of text classification strategies that were utilized. The authors in [12] gave a novel framework for text classification dependent on BPSO/REP-Tree hybrid. The first term refers to the "Binary Particle Swarm Optimization" that we use it for the feature selection process and the second term refers the classifier we used "Reduced Error Pruning Tree". We will show the results of the experiments on a data-set collected from the BBCArabic website using the Weka tool which specific for data classification. In [16] actualized the K-NN and Naïve Bayes algorithm so as to make a commonsense comparison among them and past studies. The algorithms are considered as probably the most renowned algorithms in the field of text classification. In [17] utilized a Naive Bayes (NB) algorithm which is a statistical machine learning algorithm. It is utilized to classify non-vocalized Arabic web documents (after their words have been changed to the corresponding canonical form, i.e., roots) to one of five predefined categories. In [4] utilized term Frequency-Inverse Document Frequency (TF-IDF), with the Convolutional Neural Network (CNN) on different sizes of the datasets (data_27k, data_55k, data_83k, data_111k). The accuracy was over 92%. In [55] Master-Slaves technique (MST) was updated and implemented on Arabic text classification. About 16757 Arabic documents are used 90% as training data and 10% as testing data. The accuracy was 86.495%. In [56] presented a parallel classification approach based on the Naïve Bayes algorithm for large volume of Arabic text using MapReduce with enhanced speedup, and preserved accuracy. The accuracy was close to 97%.

Master-Slaves technique (MST) consists of one classifier as a master and several other classifiers as slaves. The master classifier modifies its probability according to the results of slaves by multiplying each probability by a factor. This factor reflects the weight of those slaves. Multinomial Logistic Regression (MLR) is also known as Log-linear Models, and a conditional exponential classifier or logistic regression classifier.

Multinomial logistic regression modelling is a general and an intuitive way for estimating a probability from the data and it has been applied successfully in various natural language processing tasks [55]. Voting Technique building multiple models (typically of different types), and simple statistics (like calculating the mean) are used to combine predictions This work depends on NB, MW, MLR, and KNN in its results.

Maximum Weight (MW) is a new classifier suggested. It is a very simple method for text classification, which works by selecting the highest weight of the term among the categories and only these values are used to predict the best class for any input example [55]. Rocchio classifier is an information retrieval algorithm. Rocchio classifier is a linear classifier, and it is based on relevance feedback. Rocchio classifier developed based on Vector Space Model [44]. Artificial Neural Networks (ANNs) are one of the main tools used in machine learning. The concept of ANNs is inspired from biological human brains. In ANNs the system learns how to perform tasks after training stage. ANNs are available in different forms and shapes such as supervised and unsupervised learning. When ANNs are introduced the aim is to solve problems in the same way the human brain solves it. ANNs can be found in different shapes such as single layer perceptron, radial basis network (RBN), multi-layer perceptron (MLP) [44]. Arabic Text Classification using deep learning Technics: Term Frequency-Inverse Document Frequency (TF-IDF) with the Convolutional Neural Network (CNN) one of the most famous deep learning algorithms used especially in image processing and pattern recognition fields. CNNs is simple and efficient method to classify Arabic text from large dataset [4]. Fig. 4 shows a comparison between the traditional and deep learning techniques for various size dataset.

The accuracy is tested using the K-fold cross-validation strategy. In this strategy, the original sample is grouped into K sub samples. Of the K sub-samples, a solitary sub sample is held as the validation data for testing the model, and the rest of the K-1 sub samples are utilized as training data. The cross-validation process is then rehashed K times (the folds), with every one of the K sub samples utilized precisely once as the validation data. The K results from the folds at that point can be arrived at the midpoint of (or otherwise combined) to create a solitary estimation.

The advantage of this strategy over then rehashed arbitrary sub-sampling method is that all perceptions are utilized for both training and validation, and every perception is utilized for approval precisely once. The 10-fold cross-validation is generally utilized.

In stratified K-fold cross-validation, chosen with the goal that the mean reaction esteem is roughly equivalent in every one of the folds. Because of a dichotomous classification, this implies each fold contains generally similar extents of the two types of class labels.

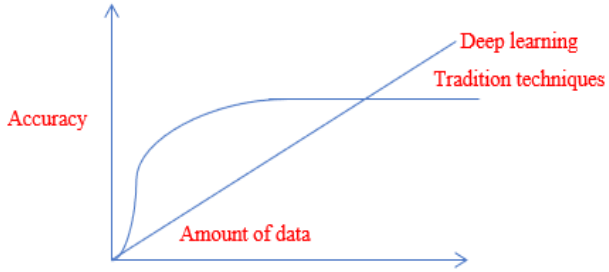


Fig. 4. Deep Learning vs Traditional Techniques.

V. TABULATED SUMMARY OF THE SURVEY WORK

The section shows a summary of the survey work in a tabular form. We present three tables, one showing the feature selection methods, the second shows the classification algorithms used, and the third shows the dataset chosen by various authors for testing.

Table I gives a comparison of various feature selection algorithms covered in this work and the best performing.

Table II gives a comparison of various text classification techniques used and the advantages and performance results of them.

Table III below shows the various datasets used by the referenced papers in this research. The table also shows the source of the dataset whether is in-house or open source. It also shows the size and categories of the datasets.

TABLE. I. FEATURE SELECTION ALGORITHM

Authors	**FS Technique**	**Conclusions**
[44]	Information gain, gain ration ,Chi square	Experiments show that chi-square is a little bit better than gain ratio and information gain
[27] [12]	Binary Particle Swarm Optimization (BPSO-KNN)	Results show that the BPSO-KNN produces classification results that compares well when other feature selection techniques are used
[57]	Term collocation	Results show that the classification outperforms the full and summery corpus techniques.
[55]	TF-IDF	Only one feature selection is presented by author due to its simplicity
[8]	None specified	–
[58]	Root and light stemmers	Light10 stemmer produced better results than the roo-based stemmer.
[33]	Authors used 10 feature selection methods that are included with the RapidMiner software. These include IG, TF, DF, CHI square, and Mutual Information (MI)	The CHI square showed best average accuracy when used with the SVM classifter.
[31]	They used the root-based and the light stemmer techniques.	The number of features for the light stemmer were larger than those of the root-based stemmer (15,000 as compared to 12,000).
[26]	Not specified	.
[16]	The TF-IDF term weighting	
[47]	The TF-IDF term weighting	.
[15]	The Chi-Square technique	The Chi-square is applied on document frequency and the top 30 terms in each class of documents.
[28]	TF, TF-IDF, WIDF	WIDF scheme provides the best performance on k-NN, while TF-IDF shows the best performance on Rocchio.
[10]	They have used six feature selection methods.	The ones used are Information Gain (IG), Chi-Square, Mutual Information (MI), NGL, GSS, and Odds Ratio (OR) feature selection. .
[52]	TF and IG	They used a combination of Term frequency to eliminate rare terms and Information Gain to leave the most valuable terms. They also used a combination of local an global features.
[54]	A combination of TF and DF is selected.	Authors tested various values of TF and analyzed the effect of the TF threshold on the efficiency of the classifier. An improvement of as high as 25% can be achieved by selecting the right value of the TF threshold.
[6]	They used the TF	Authors did not report any other feature selection methods.
[17]	They used TF-IDF	Authors tested their technique with and without root extraction selection. They concluded that the efficiency of the classifier is not sensitive to root extraction.

TABLE. II. TEXT CLASSIFICATION ALGORITHMS

Authors	TC Technique	Conclusion
[44]	Information square gain, gain ration,	Experiments show that chi-square is little bit better than gain ratio and information gain.
[27] [12]	Binary Particle Swarm Optimization (BPSO-KNN)	Results show that the BPSO-KNN produces classification results that compare well when other feature selection techniques are used.
[57]	SVM, NB, J45 and KNN	Results show that their AMLT technique based on term collocation (full-corpus bigram) outperforms other full-corpus and summery corpus.
[55]	Master-Slave technique where the NB is the Master classifier and the KNN, MLR and Maximum Weight (MW) are slave classifiers	Results show that using weighted voting of the slave classifiers will produce a better accuracy than the accuracy of each individual classifier.
[8]	NB, SVM, LSVC	It was found that the LSVC showed the best values of precision and recall
[58]	NB, SVM, KNN, Decision Trees (J48), and Decision Tables. Authors utilized machine learning software Weka and RapidMiner	SVM classifier showed the better results. Also the use of the light10 stemmer produced better results than the root based stemmer.
[33]	Conventional Neural Nets (CNN)	Experiments show with dataset of this size, the CNN outperforms traditional techniques such as SVM.
[31]	KNN, C4.5, NB, MLP, SVM	The results showed that the SVM classifier showed the best accuracy as compared to the other four classifiers with achieved average accuracy of 72% especially when using the Chi-Square term selection method. The NB classifier came very close to the SVM with an accuracy of 68%.
[26]	Artificial Neural Nets (ANN), SVM and the Hybrid Bee Swarm Optimization (BSO) BSO-Chi2-SVM	The authors compared the three techniques while using various stemming techniques such as light and root stemmers. Their results showed that the BSO-Chi2-SVM with light stemmer slightly out performs the other two techniques. The approximate accuracy for the three techniques was around 94% for the three techniques. They also showed that the model training time for the hybrid technique is more than the others [59].
[16]	SVM and NB	Authors used F1, Precision and Recall measures to evaluate the techniques. In most categories, the SVM technique outperformed the NB.
[47]	k-NN and NB	The authors compared their work to other work for the same algorithms that is reported in the literature and their results are in agreement. They also showed that the K-NN classifier outperform the NB. In addition, the performance results of the k-NN algorithm depend on the selection of k value. They showed that the effectiveness of the model declines after a value of k = 15.
[15]	They developed a tool called Arabic Text Classifier (ATC) based on NB and k-NN	The tool computes the accuracy of the two algorithms and selects the average accuracy.
[28]	SVM and Decision Trees C5.0	RapidMiner and Clementine software tools have been used to test the two algorithms. The results showed that the accuracy of the C5.0 algorithm shows a better accuracy by over 10% as compared to the SVM.
[10]	Authors used the k-NN, Rocchio and the NB	NB gave best measure results with Micro-F1 followed by k-NN and the Rocchio.
[52]	Authors used only the SVM algorithm	They used a publically available software called TinySVM. They tested the classifier for with and without feature selection. They concluded that the CHI, GSS and NGL showed best performance with the SVM classifier.
[54]	Tested SVM, NB, k-NN, and the Roc- chio classifiers	Their results showed that the SVM classifter outperform the other in high dimensional feature space.
[6]	Authors used the Decision Tree classifier (C4.5).	They compared their classifier with others such as the NB, and Vector Space Model (VSM). The F1 results of the comparison showed that the C4.5 classifier produces better values than other classifiers in the comparison.
[17]	The Maximum Entropy Classifier	Authors built their own software (ArabCat) using Java programming and based on an existing Arabic morphological analyzer. They compared their system to other existing systems such as those built by Sakhr, Al-Halees, Sawaf and El-Kourdi. The overall performance of their system was better than the others with some exception in the precision value.
[44]	NB algorithm	The average accuracy of all five categories were 68.8% which is comparable to other work by [60] that showed a result of 75%.

TABLE. III. DATASETS

Authors	Dataset	Conclusion
[44]	Authors built their own dataset.	Dataset consisted of 4000 documents collected from news websites including Al-Jazeerah, Al-Hayat and the Saudi press agency. The dataset is split into 8 categories including Economics, Politics, Arts, Culture, TEchnology, Science and Education.
[27] [12]	They have used three readily available datasets.	the Akhbar-Alkhaleej Arabic dataset that consists of 5690 documents, the Al-jazeerah dataset (Alj-News) that consists of 1500 documents, and the Alwatan Arabic dataset of 20,291 documents. These datasets are split into 4-5 categories including Sports, Arts, Economy and Religion.
[57]	800 Arabic text documents divided into four categories Economy, Politics, Religion, and Science, 200 documents for each class	Results show that the classification is better when using the proposed technique of term collocation.
[55]	16757 Arabic documents collected from Al-Sabah newspaper that are manually organized into five different categories	Authors used 90/10 percentage for the training and testing data and used a ten-fold cross-validation method.
[8]	Web content dataset but not mentioned specifically	Due to the limited size of the dataset the authors suspect the result could have a probability of inaccuracy due to insufficient training.
[58]	2700 online Arabic articles collected by Diab Abu Aiadh equally spread over nine categories.	Due to their hardware limitations, they managed to perform classification of 1000 documents spread over 5 categories.
[33]	Used three Arabic websites (Assabah, Hespress, Akhbarona) website to obtain around 111,000 documents	Authors claim to have collected over 300 million words covering five categories including sports, politics, culture, and economics. Experiments show that the training time for such a large dataset can be as high as 10 hours.
[31]	Authors built their own corpus that consisted of seven genres including Saudi press, Saudi newspapers, Writers, websites, Forums, and Islamic topics.	The total number of documents collected was over 17,000 with the number of wards of 11 million in total.
[26]	Authors used the Open Source Arabic Corpus (OSAC).	The corpus consists of 22,429 textual documents representing ten categories such as Economics, Religion, Health and Education.
[16]	The Saudi Newspapers (SNP)	Dataset consists of 5121 documents of various lengths that are categorized into seven categories (Culture, Economics, General, Sports, Information Technology, Politics and Social).
[47]	Used their own corpus	Corpus consisted of 242 documents grouped into 6 categories (which is a relatively small dataset).
[15]	Built their own corpus out of various newspaper websites including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1562 documents in 6 categories.
[28]	They used the Saudi Press Agency (SPA) and The Saudi News Papers (SNP) and other genres including writers, Islamic topics and others.	The corpus consisted of 17,658 documents and over 11 million words. A 70/30 dataset split is used for training and testing results.
[10]	In house built from online available Arabic newspaper archives including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1445 documents of various lengths that are spread over 9 categories including Medicine, Sports, Religion, and Politics.
[52]	in house built from online available Arabic newspaper archives including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1445 documents of various length that are spread over 9 categories including Medicine, Sports, Religion, and Politics.
[54]	In house built collection of documents.	Authors collected 1,132 documents that contained about 95,000 words (22000 unique words). These documents were collected from the three news websites, the Ahram, Akhbar, and El-Gumhuria.
[6]	Authors build their own corpora but based on two existing ones	The two exiting datasets are the Arabic Scientiftc Encyclopedia ”Haal Taalam” or “Do you know” that contained 373 documents and 8 categories. The second is based on “Prophetic Hadeeth” and consists of 453 documents and 14 categories.
[17]	Built their own corpus from the Al-Jazeerah news website	The corpus contained six categories including arts, science and technology, politics, sports, culture, economics and health.
[44]	Authors built their own dataset collected from the Al-Jazeerah news website	They collected 1500 documents spread equally over five categories including sports, business, culture-art, science, and health. Each category had 300 web pages.

VI. Conclusions

In this work we reviewed over 50 papers reported on the literature regarding Arabic text classification. The majority of those papers were reporting technical content regarding various algorithms, datasets, and feature selection approaches while few other claimed to present a survey on the subject. We claim that our work presented herein provides a true comprehensive survey with a critique view. The following is a summary of our finding based on this survey.

- The majority of work on classifying the Arabic language is done in the past decade (2000 – 2010) with the exception of few incidences. They applied the mainstream classification techniques such as SVM, NB, k-NN, Decision Trees, and ANN. Few incidences proposed combined classifiers such as Maximum Entropy, master-slave and BSO-SVM.
- The datasets used in the work presented is mainly in-house built from news Arabic websites. Few works used datasets built by other researchers such as the Open Source Arabic Corpus. There are no popular Arabic corpses that is used by the majority of researchers.
- The size of the corpora used is relatively small as compared to the largest available English corpus (400 million words). The largest reported Arabic collected corpus in the literature contained about 310 million words collected from three websites only and covering five categories including a general one. This is only recently reported and is not standardized. Also, the number of categories are relatively small as compared to the nature and the richness of the Arabic language to other language.
- Some works used root-based stemming while others used light stemmers. It was shown that the use of light stemmer produces better performance measure than the root-based stemmers.
- The majority of work reported better performance measures of the SVM technique over other such as NB, k-NN, ANN, and Rocchio.
- Numerous feature selection techniques have been used including the TF, DF, TF-IDF, and Chi-Square. The majority of works used either TF or the TF-IDF due to their simplicity. The Chi-Square technique, however, proved to produce better results than other feature selection mechanisms.
- Some hybrid techniques have been used such as Master-Slave and BSO-SVM which showed better performance over the individual techniques.

VII. Future Work

We propose the importance of building a professional and diverse Arabic corpus to encourage further research on the Arabic language and to help generate benchmarks. We also recommend the adoption of one of the readily available light stemmer as a standard. Further research could be done on using the semi-supervised machine learning techniques to avoid the need for a large training dataset that is prepared with some human intervention that is usually prone to errors.

Acknowledgment

This research is funded by the Islamic University of Madinah Tamayoz research Fund number 1440/53. The authors would like to thank the Dean of Research, Dr. Saleh Al-Amri and the Vice-Dean, Dr. Tareq AlFraidy for their continuous help and support.

References

[1] R. Jindal, R. Malhotra, and A. Jain, "Techniques for text classification: Literature review and current trends." *Webology*, vol. 12, no. 2, 2015.

[2] X. Luo, W. Ohyama, T. Wakabayashi, and F. Kimura, "A study on automatic Chinese text classification," in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 920–924.

[3] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and Moloshnikov, "Machine learning models of text categorization by author gender using topic-independent features," Procedia Computer Science, vol. 101, pp. 135–142, 2016.

[4] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," International Journal of Grid and Distributed Computing, vol. 11, no. 9, pp. 103–114, 2018.

[5] Y. Caballero, R. Bello, D. Alvarez, and M. M. Garcia, "Two new feature selection algorithms with rough sets theory," in IFIP International Conference on Artificial Intelligence in Theory and Practice. Springer, 2006, pp. 209–216.

[6] A. M. El-Halees, "Arabic text classification using maximum entropy," IUG Journal of Natural Studies, vol. 15, no. 1, 2007.

[7] M. S. Khorsheed and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Language resources and evaluation, vol. 47, no. 2, pp. 513–538, 2013.

[8] M. A. Ahmed, R. A. Hasan, A. H. Ali, and M. A. Mohammed, "The classification of the modern Arabic poetry using machine learning." Telkomnika, vol. 17, no. 5, 2019.

[9] A. M. F. A. Sbou, "A survey of Arabic text classification models," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 6, pp. 4352–4355, 2018.

[10] A. M. A. Mesleh, "Support vector machines based arabic language text classification system: feature selection comparative study," in Advances in Computer and Information Sciences and Engineering. Springer, 2008, pp. 11–16.

[11] M. M. Syiam, Z. T. Fayed, and M. B. Habib, "An intelligent system for Arabic text categorization," International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, pp. 1–19, 2006.

[12] H. Naji, W. Ashour, M. Al Hanjouri, and P. Gaza, "Text classification for arabic words using bpso/rep-tree," In: International Journal of Computational Linguistics Research, vol. 9, no. 1, pp. 1–9, 2018.

[13] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.

[14] Internet World Stats, "Internet world users by language-top 10 languages," 2015. [Online]. Available: http://www.internetworldstats.com/stats7.htm

[15] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh, "Automatic arabic text classification," Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, Lyon-France, 2008.

[16] M. J. Bawaneh, M. S. Alkoffash, and A. Al Rabea, "Arabic text classification using k-nn and naive Bayes," Journal of Computer Science, vol. 4, no. 7, pp. 600–605, 2008.

[17] M. El Kourdi, A. Bensaid, and T.-e. Rachidi, "Automatic Arabic document categorization based on the naïve Bayes algorithm." The 20th International Conference on Computational Linguistics, Geneva, 2004.

[18] A. Moh'd Mesleh, "Feature sub-set selection metrics for Arabic text classification," Pattern Recognition Letters, vol. 32, no. 14, pp. 1922–1929, 2011.

[19] D. Said, N. M. Wanas, N. M. Darwish, and N. Hegazy, "A study of text preprocessing tools for Arabic text categorization," in The second international conference on Arabic language, 2009, pp. 230–236.

[20] S. C. Dharmadhikari, M. Ingle, and P. Kulkarni, "Empirical studies on machine learning based text classification algorithms," Advanced Computing, vol. 2, no. 6, p. 161, 2011.

[21] F. Sebastiani, "Text classification," 2005. [Online]. Available: http://nmis.isti.cnr.it/sebastiani/Publications/EDTA05.pdf.

[22] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," Journal of advances in information technology, vol. 1, no. 1, pp. 4–20, 2010.

[23] A. Karima, E. Zakaria, T. G. Yamina, A. Mohammed, R. Selvam, V. VENKATAKRISHNAN et al., "Arabic text categorization: a comparative study of different representation modes," Journal of Theoretical and Applied Information Technology, vol. 38, no. 1, pp. 1–5, 2012.

[24] R. Duwairi, "Arabic text categorization," Int. Arab J. Inf. Technol., vol. 4, pp. 125–132, 2007.

[25] S. Al-Saleem, "Associative classification to categorize Arabic data sets," International Journal of ACM Jordan, vol. 1, no. 3, pp. 118–127, 2010.

[26] S. Alsaleem et al., "Automated arabic text categorization using svm and nb." Int. Arab J. e-Technol., vol. 2, no. 2, pp. 124–128, 2011.

[27] H. K. Chantar and D. W. Corne, "Feature subset selection for Arabic document categorization using bpso-knn," in 2011 Third World Congress on Nature and Biologically Inspired Computing. IEEE, 2011, pp. 546–551.

[28] G. Kanaan, R. Al-Shalabi, S. Ghwanmeh, and H. Al-Ma'adeed, "A comparison of text-classification techniques applied to Arabic text," Journal of the American society for information science and technology, vol. 60, no. 9, pp. 1836–1844, 2009.

[29] M. Ahmed and R. Elhassan, "Arabic text classification review," International Journal of Computer Science and Software Engineering, vol. 4, no. 1, pp. 1–5, 2015.

[30] M. Ali and R. Elhassan, "Arabic text classification process," International Journal of Computer Science and Software Engineering, vol. 6, no. 11, pp. 258–265, 2017.

[31] M. K. Saad and W. M. Ashour, "Osac: Open source Arabic corpora," Osac: Open source arabic corpora, vol. 10, 2010.

[32] A. Al-Zoghby, A. S. Eldin, N. A. Ismail, and T. Hamza, "Mining Arabic text using soft-matching association rules," in 2007 International Conference on Computer Engineering & Systems. IEEE, 2007, pp. 421–426.

[33] B. Al-Shargabi, W. Al-Romimah, and F. Olayah, "A comparative study for arabic text classification algorithms based on stop words elimination," in Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications. ACM, 2011, p. 11.

[34] M. Aljlayl and O. Frieder, "On arabic search: improving the retrieval effectiveness via a light stemming approach," in Proceedings of the eleventh international conference on Informa- tion and knowledge management. ACM, 2002, pp. 340–347.

[35] M. Gheith and T. El-Sadany, "Arabic morphological analyzer on a personal computer," in Arabic Morphology Workshop, Linguistic Summer Institute, Stanford, CA, 1987.

[36] H. M. Al-Serhan, R. Al Shalabi, and G. Kannan, "New approach for extracting Arabic roots," 2003.

[37] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prog- nosis," Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 188–195, 2011.

[38] R. Duwairi, M. N. Al-Refai, and N. Khasawneh, "Feature reduction techniques for Arabic text categorization," Journal of the American society for information science and technology, vol. 60, no. 11, pp. 2347–2352, 2009.

[39] R. M. Duwairi, "A distance-based classifier for Arabic text categorization." in DMIN, 2005, pp. 187–192.

[40] R. Mamoun and M. Ahmed, "Arabic text stemming: Comparative analysis," in 2016 Conference of Basic Sciences and Engineering Studies (SGCAC). IEEE, 2016, pp. 88–93.

[41] L. C. Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, 2002, pp. 306–313.

[42] Q. A. Al-Radaideh and G. Y. Al-Qudah, "Application of rough set-based feature selection for arabic sentiment analysis," Cognitive Computation, vol. 9, no. 4, pp. 436–445, 2017.

[43] Q. A. Al-Radaideh and L. M. Twaiq, "Rough set theory for Arabic sentiment classification," in 2014 International Conference on Future Internet of Things and Cloud. IEEE, 2014, pp. 559–564.

[44] A. H. Mohammad, "Arabic text classification: A review," Modern Applied Science, vol. 13, no. 5, 2019.

[45] A. M. Al Sbou, "A survey of arabic text classification models." International Journal of Electrical & Computer Engineering (2088-8708), vol. 8, 2018.

[46] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for arabic news articles," Natural Language Processing in ACL2001, Toulouse, France, 2001.

[47] Z. S. Zubi, "Using some web content mining techniques for arabic text classification," in Proceedings of the 8th WSEAS international conference on Data networks, communications, computers, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society. Citeseer, 2009, pp. 73–84.

[48] A. Gelbukh, "Computational Linguistics and Intelligent Text Processing" 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings. Springer, 2006, vol. 3878.

[49] A. M. El-Halees, "A comparative study on arabic text classification," A comparative study on Arabic text classification, vol. 30, no. 2, 2008.

[50] J. Ababneh, O. Almomani, W. Hadi, N. K. T. El-Omari, and A. Al-Ibrahim, "Vector space models to classify Arabic text," International Journal of Computer Trends and Technology (IJCTT), vol. 7, no. 4, pp. 219–223, 2014.

[51] M. K. Saad and W. M. Ashour, "Arabic text classification using decision trees," Arabic text classification using decision trees, vol. 2, 2010.

[52] T. F. Gharib, M. B. Habib, and Z. T. Fayed, "Arabic text classification using support vector machines." IJ Comput. Appl., vol. 16, no. 4, pp. 192–199, 2009.

[53] M. Abdeen, A. Elsehemy, T. Nazmy, and M. C. Yagoub, "Classifying the arabic web—a pilot study," in 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2011, pp. 865–868.

[54] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving arabic text categorization using decision trees," in 2009 First International Conference on Networked Digital Technologies. IEEE, 2009, pp. 110–115.

[55] Z. A. Abutiheen, A. H. Aliwy, and K. B. Aljanabi, "Arabic text classification using master-slaves technique," in Journal of Physics: Conference Series, ser. 1032, no. 1. IOP Publishing, 2018, pp. 012–052.

[56] M. M. Abushab and R. S. Baraka, "Large-scale Arabic text classification using mapreduce," Master, Islamic University of Gaza, 2017.

[57] F. OLAYAH and W. ALROMIMA, "Automatic machine learning techniques (amlt) for Arabic text classification based on term collocations." Journal of Theoretical & Applied Information Technology, vol. 96, no. 12, 2018.

[58] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, "Automatic arabic text categorization: A comprehensive comparative study," Journal of Information Science, vol. 41, no. 1, pp. 114–124, 2015.

[59] R. Belkebir and A. Guessoum, "A hybrid bso-chi2-svm approach to arabic text categorization," in 2013 ACS Inter- national Conference on Computer Systems and Applications (AICCSA). IEEE, 2013, pp. 1–7.

[60] M. Yahyaoui, "Toward an arabic web page classifier," Master's Project, 2001.

Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study

Raghad Bogery[1], Nora Al Babtain[2], Nida Aslam[3], Nada Alkabour[4], Yara Al Hashim[5], Irfan Ullah Khan[6]
Department of Computer Science
College of Computer Science and Information Technology
Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia

***Abstract*—Due to widespread availability of Internet there are a huge of sources that produce massive amounts of daily news. Moreover, the need for information by users has been increasing unprecedently, so it is critical that the news is automatically classified to permit users to access the required news instantly and effectively. One of the major problems with online news sets is the categorization of the vast number news and articles. In order to solve this problem, the machine learning model along with the Natural Language Processing (NLP) is widely used for automatic news classification to categorize topics of untracked news and individual opinion based on the user's prior interests. However, the existing studies mostly rely on NLP but uses huge documents to train the prediction model, thus it is hard to classify a short text without using semantics. Few studies focus on exploring classifying the news headlines using the semantics. Therefore, this paper attempts to use semantics and ensemble learning to improve the short text classification. The proposed methodology starts with preprocessing stage then applying feature engineering using word2vec with TF-IDF vectorizer. Afterwards, the classification model was developed with different classifier KNN, SVM, Naïve Bayes and Gradient boosting. The experimental results verify that Multinomial Naïve Bayes shows the best performance with an accuracy of 90.12% and recall 90%.**

***Keywords*—Natural language processing; feature engineering; word embedding; text classification; ensemble learning**

I. Introduction

At present, due to low cost hand-held multimedia enabled devices along with the fast internet, huge amount of information is created and accessed daily. Internet is the main source of information and the integral part of individual's life. Due to the focus on mobility and the internet in the recent years and to reduce the paper waste, many news companies went online and changed the traditional way of printing newspapers and articles. Because of that there's a huge number of different articles in the news website databases. However, categorizing news article in its respective category manually is very difficult and time consuming. Automatic categorization of the news corpus will profit society in several ways. However, automatic categorization of the news headlines is a challenging task as the length of news varies. There is a need for an automated way to extract and access the news according to the user's interest. Hence, this paper aim to propose an automatic news headline categorization that is based on machine learning techniques.

NLP is mainly used to automatically categorize documents and speech by words count or frequency without considering the meaning behind the words. This method is useful for document or huge chunk of text categorizing. However, news headlines and descriptions are usually short, so such methods might not accurately categorize the articles in their respective category. Because of that, many researchers started exploring semantic classification instead of relying on single word meanings to achieve better accurate results.

In-order to achieve effective classification or clustering of news headlines, it is important to consider four methodologies that helps in the semantic analysis i.e. the background knowledge, word representation and feature vectorization technique, topic modeling, the similarity measure that might be used to assess the clustering algorithm. Table I summarizes these methodologies with a brief explanation and examples.

The remaining part of our study is organized as follow. Section 2 contains a literature review. Section 3 contains the description of the proposed methodology. Section 4 contains empirical studies that include dataset description, Section 5 contains the experimental setup and finally the last section contains the conclusion of our study.

TABLE. I. Methodologies to Enhance Clustering Algorithm [1] [2]

Method	Definition	Examples
Background knowledge	Background knowledge can help learning model to better understand the relationships, the context and the meaning of words.	Ontologies, WordNet, semantic networks, treasures, and taxonomies.
Topic Modeling	Models the topics of different documents.	Lantent Dirchlet Allocation algorithm (LDA), Lantent Semantic Indexing technique (LSI).
Word Representation	The words in NLP are usually represented in the vector space model, each vector represent a word against its occurrence in corpus. Vectors with single words are called Bag of Words (BOW).	BOW, term frequency, Tf-IDF, Word2Vec, GloVe.
Similarity Measures	Used to see wither two words have the same or the opposite meanings in the vector space.	Cosine similarity, Euclidean distance, Jacob similarity.

II. Literature Review

This section discusses the literature reviews of the most important studies related to our topic. These topics are feature engineering, such as word embedding and text summarization, clustering using different methods, and finally classification.

A. Feature Engineering

One of the important steps in NLP is applying feature engineering for the text dataset, first to preserve the context of text and secondly to reduce the vector's dimensionality of the text. The word embeddings and semantic network like WordNet can help in preserving the context which will be discussed in the following sections, while text summarization can be used to reduce the dimensionality of text.

a) Word embedding: As the BOW, term frequency, TF-IDF; all represents the words in vector space model for the learning algorithm. However, they don't preserve the context and the relationships of the words in the documents. Word embedding can solve this problem as it models the semantics and relationships of words in a corpus using a vector with low dimensional as compared to the dimensional size of vocabularies in a corpus [1]. Several studies have been made on comparative analysis of word embedding techniques in various domains i.e. biomedical, twitter elections [3]. Twitter is one of the biggest source of individual opinion, news media [4]. There are two main approaches for words embedding that are known for their efficiency and accuracy and they are Word2Vec and GloVe.

A comparison study has been made for bio-medical NLP using Wikipedia, biomedical publications from PubMed and Medlist [5]. In 2013 Mikolov et. al [6], model was proposed for word2vec which has two architectures to learn and represent the words in the vector space. The first architecture is the continuous bag of words (CBOW) and second the skip-gram architecture. They performed word analogy to test their model, the word analogy was based on the semantic and syntactic questions that were produced by the authors. The authors observed a higher accuracy for both semantics and syntactic questions when both dimensional size and the number of training vocabulary increased. The methods that were proposed have the advantage of low computation time, but both consider the context locally in a document without making the advantages of the occurrence in different documents.

To overcome the previously mentioned problem, Jeffery et al. [7] proposed GloVe Model which stands for Global Vectors. The word's vector in the GloVe model is represented by not only considering the word co-occurrence probability in one document, but also considers the ratio co-occurrence probability across the documents. They tested their model on different tasks and conducted a comparison between CBOW and GloVe model and other baselines. They used the same testing approach as Mikolov et al. The performance of the model of the task analogy was increasing with the number of dimensions. In their comparison, they showed that the proposed model outperforms both architectures of word2vec in word analogy in semantic and syntactic questions. A comparative study to compare various feature engineering mechanisms for news articles and twitter tweets [8] . Continuous bag of words and skip-gram word embedding method using Convolutional neural network classifier. The experiments were conducted using these two word-embedding models and without any feature engineering approach for real news article and tweets. For the news article the CBOW achieved the highest accuracy while for the tweets Skip-gram outperform. A comparative study has been made by Jang et al. [9] on news article and news on social media in Korean language. The news was downloaded from NAVER a Korean news site, while twitter API was used to download the news from twitter.

b) Text summarization: Chi et al. [10] proposed a summarization model named Sentence Selection with Semantic Representation (SSSR). Through learning semantic sentence representation and implementing appropriate selection methods. SSSR also has two main parts which are sentence selection strategy and the sentence representation learning. Sentence selection strategy is to select a sentence that can rebuild the original document with the minimum falsification.

While in the semantic representation of sentences before implementing the selection strategy. Sentences can be represented using two representation the weighted mean of words embedding (SSSR-w) and deep coding (SSSR-d). The word embeddings were based on word2vec model, each word embedding is weighted based on the TF-IDF. Their experiment was conducted on DUC2006 and DUC2007 datasets they used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to evaluate the text summary results. Both models produce good results compared to other baselines using the F-measure metric, though SSSR-d has outperformed SSSR-w.

B. Clustering

Text clustering is considered as a challenging task, in the following sections three methodologies will be discussed for text clustering, dependency graph clustering, word embeddings clustering, and WordNet and lexical chains clustering.

a) Clustering using dependency graph: Asmaa K., et al. [11] proposed a way for reducing the problem that occurs from clustering using the traditional methods and in fact increase the clustering accuracy by using a method called dependency graph. A dependency graph represents one document, where each node is associated with a word and can be used as meta-data for the document. While semantic relations between words can be captured by using edges that are between their corresponding nodes, every edge has term weight based on TF-IDF. Dependency graph will affect the clustering result despite the clustering algorithm that is being used, and to display this K-means clustering algorithm was used to cluster the dataset. Where the number of clusters was 20 and so the value of K is 20. The number of correctly cluster documents was 188 out of 200 when using the dependency graph, while it decreases to 173 without the dependency graph.

b) Clustering using word embeddings: In contrast, Juneja et al. [12], used word embeddings to improve text clustering results. They compared between the word embeddings algorithms. GloVe, CBOW, and skip-gram all of

them have high dimensional space, however, GloVe has a higher dimensional space and time complexity. They used GloVe for text clustering since they were more concerned about accuracy rather than the time complexity. Their proposed methodology used T-SNE algorithm to reduce the dimensionality of the GloVe model for better understanding and visualizing the results of their work. In addition, it also helps in reducing the curse of dimensionality where the irrelevant words mask the relevant words. They also saved the words embedding in files to reduce the time complexity of the GloVe model and used k-means as the clustering algorithm. They also saved the words embedding in files to reduce the time complexity of the GloVe model and used k-means as the clustering algorithm. They tested their methodology on two datasets, one dataset showed an increase of error rate with the decrease of the number of dimensions, and it is due to the data loss. While the second dataset doesn't have a specific number of K and thus the result was acceptable.

c) Clustering using WordNet and lexical chains: Using word embedding isn't the only way to improve the semantic analysis of a corpus. Tingting Wei, et. al [13] used WordNet and lexical chains and a modified Word Sense Disambiguation (WSD) to propose a method that meaningfully cluster texts while reducing the text dimensions. They modified the WSD similarity measure by combining two methods to create a more accurate similarity measure which is Wu–Palmer measure based on the least common subsume (LCS) and Banerjee and Pedersen's measure based on mutual words in the word's definitions. After performing WSD using the modified similarity measure, and extracting core semantics using lexical chains, Tingting Wei, et al. performed clustering using Bisecting K-means by assigning K as the number of classes was previously known. They compared it against other methods without using lexical chains by the same clustering method. The used methods were Base (WSD is not performed while performing all basic preprocessing such as removing stop words), Disambiguated Concepts (WSD is performed as well as performing all basic preprocessing), Disambiguated Core Semantic (WSD is performed using lexical analysis as well as performing all basic preprocessing).

They have shown that disambiguated core semantic method that uses lexical chains produce the highest F1-measure and purity on three groups. These results prove that the proposed method not only produces purer clusters, but also decreases the computational cost by decreasing the text dimensions using lexical chains.

In the study WordNet was used to find the semantic relation between the words. The sense of the words has been found by first selecting the key word and then finding the related words by using the WordNet hierarchical semantic relations. For finding the sense of the word, not only the selected word but also the related words were considered i.e. hypernym, holonym etc. using WordNet. The experiments were performed on SENSEVAL-2 dataset and the achieved accuracy was 32%. WordNet and Senti-WordNet [14] was used to classify the news headlines based on semantic and sentiment. In the study [15], the multilayer model was used for Forex real market data news categorization, the model used the combination of WordNet and Senti-WordNet. Senti-WordNet was used to find out the polarity of the news i.e. positive or negative news for the market prediction and achieved the accuracy of 83.33%. The Nassirtoussi et al study was recently extended by Seifollahi et al. [16] by adding the WSD in the semantic analysis module in order to exploit the impact of WSD on results. The dataset was initially divided into two categories date and time. The interval of the news was 2 hours. The headlines were analyzed by using the proposed model to monitor the exchange rate to predict the P (increase in exchange rate) and N (decrease in exchange rate). The system outperforms the previous system in terms of accuracy and time.

C. Classification

Another task in NLP is text classification, in the following section different algorithms were used to for text classification.

a) Classification using Different classifiers: Vishwanath et al. [17] proposed an improved term graph model and conducted a comparison between the KNN, term graph algorithm model and Naïve Bayes. The term graph model was used to preserve the semantics of the words in the datasets by using a weight in a directed graph for frequently co-occurring words. In this model, documents are treated as transactions and uses frequent item set mining algorithms. On the other hand, The KNN uses the vector space model which is based on TF-IDF and similarity measures to see whether one document belong to a certain class or not. The Naïve Bayes assign a probability for terms that belong to a certain class, the document, in the end, is classified to one class by the summation of each term probability for a certain class. They trained KNN, term graph model and Naïve Bayes on the dataset and then compared between the three model results. Among the three algorithms, the KNN outperformed both the term graph model and Naïve Bayes; term graph model has higher and closer accuracy to KNN, while Naïve Bayes has the worst accuracy between all the models. The proposed term graph model showed an improved result compared to the other baseline term graph model. A study has been made on 130000 news article consists of 8 categories using Naive Bayes, Support Vector machine and ANN model [2]. Features were selected using chi-square and LASSO. LASSO enhanced the accuracy of Naive Bayes while SVM achieved highest accuracy with chi-square feature selection. Another study [18] was made to identify the fake news using deep semantic structural model and improved Recurrent Neural Network on twitter dataset with 99% accuracy. Experiments were conducted using individual DSSM, LSTM and the hybrid approach the combination of DSSM and LSTM. Three experiments were conducted i.e. first by dividing the data equally among training and testing, secondly by dividing into 80-20 and finally with 75-25 division. The hybrid model achieved highest outcome with 75-25 data sampling division for training and testing. Pambudi et al. [19] classified the Indonesian news into multi-class classification using Pseudo Nearest Neighbor (PNNR). The PNNR was initially proposed for the binary classification and was later extended for multiclass as well. Several proximity functions were used, and

Cosine proximity similarity measure produced the highest results as compared to Manhattan and Euclidian.

b) Classification using WordNet: The study aimed to produce a model to predict suicidal thoughts by collecting data from Twitter using Twittr4J and used Weka as a data mining tool [13]. This paper also implements its own algorithm that calculates the semantic similarity between the collected data depending on a semantic analysis resource using WordNet. They manually constructed a vocabulary related with suicide and then collected data from Twitter. After that they applied IB1, J48, CART, SMO and Naïve Bayes algorithms to perform the classification. Then they improved their results using semantic analysis based on WordNet. The precision of the algorithms is shown in Table II based on the precision of the algorithms that were used.

TABLE. II. RESULTS FROM SUICIDE PREDICTION MODELS

Algorithm	IB1	J48	CART	SMO	Naïve Bayes
Precision (tweets with risk of suicide)	71%	81.2%	83.1%	89.5%	87.5%
Precision (tweets without risk of suicide)	63%	75.4%	66.7%	70%	61%

Finally, in word embedding GloVe and word2vec both have their own strength and weakness, GloVe is designed for preserving the context in a large corpus with multiple documents while the word2vec is designed for preserving context in one document with multiple records. SSSR-w and SSSR-d both used in word representation in the SSSR model, while SSSR-d tends to have a better performance over SSSR-w. On the other hand, all the cited literature in clustering section preserve the context of words by using either a dependency graph, a semantic graph or a word embedding. While in the classification, they used a either TF-IDF or WordNet to represent the words and to preserve its semantics. As most of the traditional classification algorithms doesn't work with the word embeddings technique, we can conclude with this gap which is how to combine between the word embeddings and the classical classification algorithms while maintain a high accuracy.

III. DESCRIPTION OF THE PROPOSED TECHNIQUES

Our methodology comprises of several stages which are preprocessing, feature engineering using word2vec, classification using K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multinomial Naïve Bayes classifiers and finally training the model using boosting classifier.

A. Preprocessing Techniques

The preprocessing stage is very important for the semantic analysis. The dataset used in our study contains two columns that might help in future news headlines prediction, those two columns were combined and cleaned. As it is important for our model to maintain the semantics, thus returning each word to its root is important e.g. play and playing is the same word. Therefore, there was a trade-off between using stemming and lemmatization to improve our model. Stemming is a technique that removes suffix, it is simple and uses less computational time however it might lead to over-stemming errors due to its simplicity. On the other hand, Lemmatization uses the relationship between words and depends on WordNet to return the root of the word. This indicates that Lemmatization ensures less error but take more computational time. Since over-stemming might give wrong results for our classifier, we preferred the lemmatization over the stemming.

B. Feature Engineering Using Word2Vec

As mentioned previously, Word2Vec is a word embedding technique that uses a shallow neural network to represent the words in the vector space based on their context. There are two approaches to this technique, continuous bag of words (CBOW) and the skip-gram, as explained previously. This feature engineering method was used because it preserves the words semantics while lowering the dimensionality by dropping words that appear less than min_count, which is a hypermeter of the Word2Vec model. It also has two more important hypermeters, the dimensionality size of the word vector and the maximum distance between the current and predicted word within a sentence. In this study, the words were mapped to its produced vectors into a dictionary which was pipelined with the classification model using a TF-IDF vectorizer. Both Word2Vec approaches were tested with the classifiers.

C. Classification KNN, SVM and Naïve Bayes classifiers

The following section discussed the different classifiers that were used which are the K-Nearest Neighbor, Support Vector Machine, and Naïve Bayes classifiers.

a) K-Nearest Neighbor (KNN) Classifier: K-Nearest Neighbor (KNN) is a supervised learning algorithm that classifies data based on the training data using a similarity measure. It is one of the simplest supervised learning algorithms. This classifier was chosen because, despite its simplicity, it performs well. Moreover, in contrast to eager learners such as Naïve Bayesian, it is a lazy learner that stores the training data for future predictions and doesn't generate rules from them, so it doesn't require prior knowledge. KNN works by searching for the nearest similar K neighbor points in the training data and count their majority voting to predict the unknown class. In other words, it simply matches the unknown class attributes with the training data attributes and looks for the closest match. Because of that, its training time is short since it simply stores the training data. On the other hand, the testing time in KNN is usually far longer than the training time because it needs to compute the K neighbor voting for every test data [1].

From the previous description, KNN relies heavily on its training data, any noise can influence the prediction. Furthermore, a huge amount of training data will take time to test. Finally, the K value is also very important since it defines how many neighbors the algorithm consider while classifying. Usually, K is an odd number to avoid evenly split voting. As mentioned, KNN computes how similar the neighbor points using a similarity measure (distance measures). This study will use the Euclidian distance measure. The Euclidian is defined in equation (1) [1].

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

Where X= (x_1, x_2,…, x_i) are the set of attributes for the first data and Y= (y_1, y_2,..., y_i) for the second data. The result of d(X,Y) coordinate is plotted and compared to its neighbors.

b) Support Vector Machine (SVM) Classifier: The Support Vector Machine classifiers is one of the best machine learning techniques and outperform in most of the cases. Support Vector Machine classifiers were first introduced by Corinna Cortes and Vladivir Vapnik in 1995 [20]. It is a learning algorithm that works for both classification and regression problems. This classifier was selected because of its high performance even though it has a huge computational time [15]. The goal of SVM is to find the optimal separating hyperplane that gives the maximum separation margin between the hyperplane and the nearest points of both classes. For the set of training data that are shown in (2):

$$\{(x^1, y^1), \dots., (x^m, y^m), \quad x \in R^n, y \in \{1, -1\} \tag{2}$$

A hyperplane can be found to separates the two classes. A hyperplane is shown in equation (3):

$$\langle w, x \rangle + b = 0 \tag{3}$$

It can be said that the hyperplane is separating the classes efficiently if the distance between the nearest point and the hyperplane is maximum.

There are some parameters that affect the result of the SVM classifier. The first one is the Regularization (C) parameter, lower value of C turns a high error rate that is given for the training set and the hyperplane margin will be large which means a smaller decision function. On the other hand, a higher value of C turns low error rate that is given for the training set and the hyperplane margin will be small. The second parameter is the gamma, it defines how far the effect of a single training point reach. If the gamma is large that means the point that is close will be used for calculation while a small value of gamma means that points that are far will be used for calculation.

c) Naïve Bayes Classifier: Multinomial Naïve Bayes classifier has a version for the textual document classification based on word count. Multinomial naïve bays classifier or known as term frequency or raw term frequency tf, is an approach for characterizing text based on number of times a term t appears in document d as shown in equation (4).

$$tf\ (t, d) \tag{4}$$

A multinomial naïve Bayes is known to be simple to implement, but very efficient since it assumes that the features are mutually independent, which is why it is one of our choices in the set of classifiers to implement. In practice, usually the term frequency tf is normalized by dividing it over the document length (or the sum of the number of terms in the document) n_d as seen in equation (5).

$$normalized\ term\ frequency = \frac{tf(t,d)}{n_d} \tag{5}$$

Using the term frequency, we can estimate the maximum-likelihood from the training data to find the class-conditional probabilities, where equation 6 shows the calculations needed to find this estimation.

$$\rho(x_i|w_j) = \frac{\sum tf(x_i, d \epsilon w_j) + \alpha}{\sum N_{d \epsilon w_j} + \alpha \cdot V} \tag{6}$$

Where

x_i represents *A word from a particular sample in the feature vector x*

$\sum tf(x_i, d \epsilon w_j)$ represents *the total sum of the term frequencies of a specific word x_i from the document in the training sample d that belongs to the class w_j.*

α represents *Smoothing parameter*.

$\sum N_{d \epsilon w_j}$ represents *the total sum of all the term frequencies N in the training dataset d that belong to the class w_j.*

V represents *the vocabulary size that is in the training set*.

Then we can use the product of the likelihoods of individual words in the document to give us the class conditional probability of encountering a word x, as shown in equation 7.

$$\rho(x|w_j) = \rho(x_1|w_j) \cdot \rho(x_2|w_j) \cdot \rho(x_3|w_j) \cdot \dots \cdot \rho(x_n|w_j) = \prod_{i=1}^{m} \rho(x_i|w_j) \tag{7}$$

d) Improving the Model Using Ensemble Learning: Ensemble learning is used to increase the classifier accuracy and reduce the variance and bias, it follows different approaches including bagging, boosting, stacking and voting. The basic idea behind the ensemble learning is using multiple classifiers to improve the model's prediction. In this study gradient boosting classifier will be used to improve the model.

e) Model Improved by Gradient Boosting Classifier: The Gradient boosting classifier combine many weak classifiers, the number of classifiers indicates how many times the model will be trained. In each training phase the misclassified instances will be given higher weight to reclassify them correctly. This can be an advantage for the imbalanced classes in our dataset and will decrease the number of misclassified classes in each training iterations.

IV. Empirical Studies

This section gives a brief description of the dataset characteristics. Also, describes the experimental setup that have been done to produce the models using the different selected classifiers K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multinomial Naïve Bayes classifiers and finally improving the model using boosting classifier. As well as the followed optimization strategy.

A. Description of Dataset

The learning of any model relies on the nature and condition of the data used. In our study News Category Dataset obtained from Kaggle website have been used. This dataset is about collected news headlines from the year 2012 to 2018 obtained from HuffPost. It contains 202,372 records and 6 attributes, namely category, headline, authors, link, short description, and date. The target is the category of the headlines, containing 41 classes as shown in the Fig. 1. After cleaning the data and dropping records with empty cells, the dataset contains 200,746 records. In our study 3 classes were used containing the most records, as shown in Table III.

TABLE. III. FILTERED DATASET CLASSES

Class	Records
Travel	9887
Style & Beauty	9649
Parenting	8677

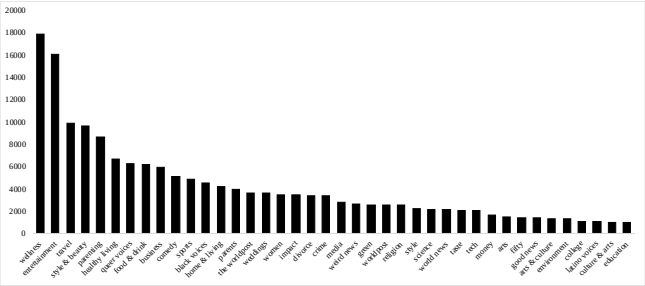


Fig. 1. News Category Dataset Class label distribution

V. EXPERIMENTAL SETUP

A. Dataset Cleaning

To clean this dataset, we started by removing the stop words, emojis and numbers. Next the missing values in the dataset were removed. The words were lemmatized using through stem.wordnet from nltk library. After that we applied word tokenization.

B. Feature Extraction using Word2Vec

After cleaning, tokenizing, and combining the data, the Word2Vecotrization was performed on the combined features using both CBOW and Skip-Gram Word2Vec methods. Both models were initialized with different measures as shown in Table IV. Based on the similarities between words using the cosine similarity measure, the CBOW3 model was chosen for the classification. For example, Fig. 2 shows the cosine similarity between the words 'health' and 'care'. Finally, the CBOW3 model's words were mapped to its vector as a dictionary to be used for classification.

C. Experimental Setup of K-Nearest Neighbor (KNN) Classifier

Firstly, a pipeline was used to combine the KNN classifier with the Tf-idf Vectorizer Word2Vec CBOW3 dictionary. Secondly, the dataset was split using the stratified technique to 70% for training data and 30% for testing data. Thirdly, a brute force grid search method of 5-fold was used on the training data to find the best odd K values between 7 and 15, resulting in 5 * 5 = 25 fits. Fourthly, the folds results were plotted as shown in Fig. 3. Lastly, the KNN classifier was validated using the unseen testing data using the best K parameter as found by the grid search previously.

D. Experimental Setup of Support Vector Machine (SVM) Classifier

For the SVM classifier the first step is to create a pipeline that combines the SVM classifier with the Tf-idf Vectorizer Word2Vec CBOW3 dictionary. The second step is to select several C and gamma parameters to be tested in the next step. The third step is to perform a 3-fold cross validation grid search on a data that was split to 70% testing a 30% training. Finally, the result of all the classifiers with the best fit parameter will computed.

E. Experimental Setup of Naïve Bayes Classifier

The multinomial Naïve Bayes has many forms such as multinomial, Gaussian, as well as Bernoulli. However, since our main goal is to preform multinomial Naïve Bayes on our dataset. The following steps have been used:

- Make sure that you have already split your data into testing and training
- Make an object from the multinomialNB() class
- Pipeline the object by the following code:
- MultiNB = Pipeline([('vect', TfidfVectorizer()), ('clf', MultinomialNB())])
- Then use It to train the data by fitting it to the object
- Define a predicted value to compare with: predicted = MultiNB.predict(X_test)
- And finally find its accuracy by finding the mean the values where the predicted = the test.

```
---------- CBOW ---------------

CBOW1 Cosine similarity between 'health' and 'care':0.938

CBOW2 Cosine similarity between 'health' and 'care':0.936

CBOW3 Cosine similarity between 'health' and 'care':0.917

---------- SkipGram (SG)---------------

SG1 Cosine similarity between 'health' and 'care': 0.835

SG2 Cosine similarity between 'health' and 'care': 0.785

SG3 Cosine similarity between 'health' and 'care': 0.795
```

Fig. 2. Cosine Similarity between the Words Example.

TABLE. IV. WORD2VEC MODELS

	Model name	Window distance	Minimum word count	Vector dimension
CBOW	CBOW1	3	50	100
	CBOW2	5	100	70
	CBOW3	7	150	50
Skip Gram	SkipGram1	3	50	100
	SkipGram2	5	100	70
	SkipGram3	7	150	50

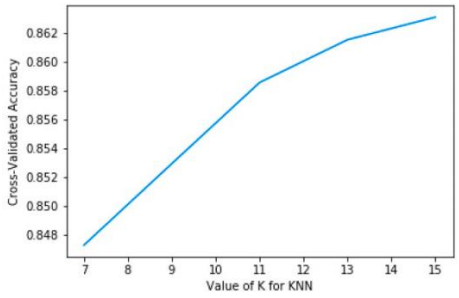


Fig. 3. 5-Fold Grid Search Results.

F. Experimental Setup of Gradient Boosting Classifier

The Gradient Boosting Classifier was used, which have different parameters that will help in tuning the boosting classifier. Those parameters are the base classifier, the number of estimators, the learning rate, the minimum sampling leaf (the number of samples to consider in the leaf) and split (the number of samples to consider when splitting the tree), the maximum depth for the classifying tree and finally the criteria of measuring the error in each iteration. All these parameters will also be tuned through grid search technique to select the best parameters that gives the optimal accuracy for the classifier. The base classifier was set on the default classifier (tree classifier) since the SVM and KNN doesn't work with classifier as a base classifier. Different parameters were tested to get the obtained results, which are listed in the following:

- Number of estimators: 50,100,200,300,400,500,600.
- Learning rate: 0.25, 0.1, 0.01,0.001.
- Maximum depth: 3,4,5,6,7,8.
- Minimum split sampling: 2,3,4,5,6,7,8.
- Minimum split leaf: 0.1,0.2,0.3,0.4.

Not all these parameters were included in the grid search, instead the best parameter from each was selected iteratively to reduce the computation time. The grid search was finally used with cross validation = 5 with the best selected parameters.

G. Optimization Strategy

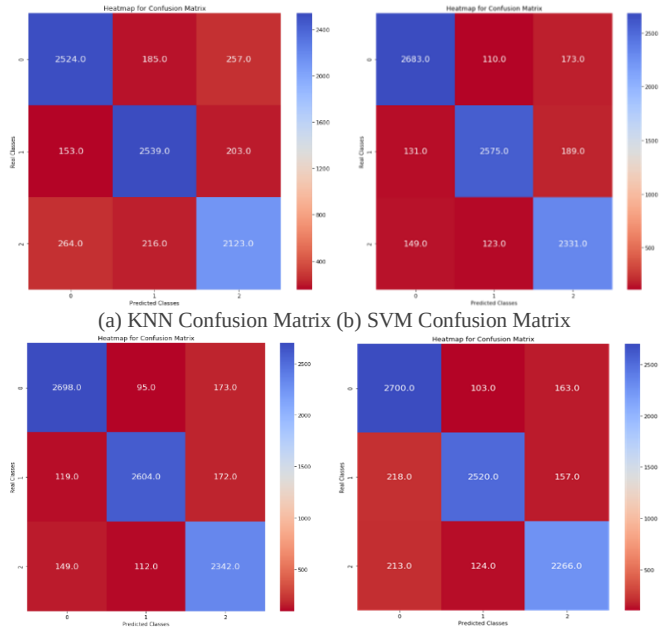
To optimize the results in each classification model, pipeline and grid search were used. In the pipeline, different parameters were selected, the grid search uses this pipeline and creates a combination from these parameters to train the model with cross validation value and then select the best results from these parameters.

VI. Result and Discussion

This section describes the results produced from our model. Based on the grid search, the best K parameter for the KNN classifier is 15 with 84.57% training accuracy. The testing accuracy of the 15-KNN model is 84.73%. Moreover, the confusion matrix is shown in Fig. 4(a), the highest false classification is in the third class 'travel' with 289 classified as 'style & beauty' and 177 as 'parenting'. The 15-KNN model recall score is 84.6%. While SVM best parameters are for gamma is 15 with C equals to 0.01 based on the grid search results, which gave a 90% training accuracy and 89.19 % testing accuracy. Additionally, the result of the confusion matrix is shown in Fig. 4. From the grid search the best parameters found was 600 for the number of estimators, 0.25 for learning rate, 7 for the tree maximum depth, 6 for minimum split sampling and 0.1 for minimum split leaf.

SVM increased the accuracy and the recall it also minimized the misclassified classes, as shown in the confusion matrix in Fig. 4(b). Finally, Naïve Bayes showed a high accuracy with 90.16% in training while the testing got a bit lower with 90.12% while the recall is 90.14%. Similarly Fig. 4(c, d) represent the confusion matrix for Multi-Nominal Naïve Bayes and Gradient Boosting. Table V shows each classifier result with precision, recall, F1 and accuracy metric for each category of news headlines.

Table VI shows the overall outcome of all the classifiers. As shown in the table most of the classifiers have low variance and low bias which indicate the proposed model doesn't have underfitting and overfitting. Also, most of the classifiers has a high recall, as recall is one of the important measures in the text classification problem. The Gradient Boosting classifier outperforms the other classifier in terms of precision and F1 score. While, the naïve Bayes gives the best accuracy and recall between them all.



(a) KNN Confusion Matrix (b) SVM Confusion Matrix

(c) MNB Confusion Matrix. (d) GB Confusion Matrix

Fig. 4. Confusion Matrix for Each Classifier.

TABLE. V. Results Comparison

Classifier	Class	Precision	Recall	F1-Score
KNN	Travel	0.82	0.82	0.82
	Style & beauty	0.86	0.87	0.86
	Parenting	0.86	0.85	0.85
SVM	Travel	0.87	0.90	0.88
	Style & beauty	0.92	0.89	0.90
	Parenting	0.91	0.90	0.90
Multinomial Naive Bayes	Travel	0.87	0.91	0.89
	Style & beauty	0.89	0.89	89
	Parenting	0.91	0.90	0.90
Gradient Boosting	Travel	0.88	0.87	0.87
	Style & beauty	0.92	0.87	0.89
	Parenting	0.86	0.91	0.88

TABLE. VI. OVERALL RESULT

Classifier	Precision	Recall	F1-Score	Testing Accuracy	Training Accuracy
KNN	84.66 %	84.66%	84.66%	84.73%	84.57%
SVM	86.33%	89.66%	86.82%	89.19%	90%
MNB	88.33%	90%	88.14%	90.12%	90.16%
GB	90%	88.33%	89.81%	88.58%	87.66%

VII. CONCLUSION

In this paper, we produced a text classification model that maintains the semantics of text to gain more accuracy and recall score. The semantics of the text was preserved by using word2vec word embeddings with TF-IDF vectorizer. We conducted a comparison between different classifiers the KNN, SVM, Naïve Bayes and the Gradient Boosting classifiers. The training was done with parameter tuning and optimization to give a better result. The best classifier out of these are Multinomial Naïve Bayes which has higher accuracy and recall compared to other classifiers. Also, compared to the other reviewed studies with the same classifier our classifier has better accuracy. The limitation of our study is it covers only three categories of the news. In the future, a possible enhancement to our work is to apply the classification on more than three targets in the dataset and improve the model using methodologies like the neural network.

REFERENCES

[1] M. P. Naik, H. B. Prajapati, and V. K. Dabhi, "A survey on semantic document clustering," 2015 IEEE Int. Conf. Electr. Comput. Commun. Technol., pp. 1–10.

[2] R. A. Sinoara, "Text mining and semantics : a systematic mapping study," 2017.

[3] X. Yang, C. Macdonald, and I. Ounis, "Using word embeddings in Twitter election classification," Inf. Retr. J., vol. 21, no. 2–3, pp. 183–207, 2018.

[4] S. M. Haewoon Kwak, Changhyun Lee, Hosung Park, "What is Twitter, a Social Network or a News Media?," WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA. ACM 978-1-60558-799-8/10/04, 2010.

[5] Y. Wang et al., "A comparison of word embeddings for the biomedical natural language processing," J. Biomed. Inform., vol. 87, pp. 12–20, 2018.

[6] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Vector Space," pp. 1–12.

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," pp. 1532–1543, 2014.

[8] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," PLoS One, vol. 14, no. 8, pp. 1–21, 2019.

[9] B. Jang and J. Yoon, "Characteristics Analysis of Data from News and Social Network Services," IEEE Access, vol. 6, pp. 18061–18073, 2018.

[10] C. Zhang, L. Zhang, C. Wang, and J. Xie, "Text Summarization Based on Sentence Selection with Semantic Representation," 2014 IEEE 26th Int. Conf. Tools with Artif. Intell., pp. 584–590, 2014.

[11] "Graph Based Text Representation for Document," Vol. 76, No. 1, 2015.

[12] K.- Means, "C Ontext - Aware C Lustering using G Lo V E and," no. July, 2017.

[13] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A Semantic Approach for Text Clustering using WordNet and Lexical Chains Expert Systems with Applications A semantic approach for text clustering using WordNet and lexical chains," Expert Syst. Appl., vol. 42, no. 4, pp. 2264–2275, 2014.

[14] S. Baccianella, A. Esuli, and F. Sebastiani, "S ENTI W ORD N ET 3. 0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," vol. 0, pp. 2200–2204, 2008.

[15] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," Expert Syst. Appl., vol. 42, no. 1, pp. 306–324, 2015.

[16] A. Khadjeh, S. Aghabozorgi, T. Ying, D. Chek, and L. Ngo, "Expert Systems with Applications Text mining of news-headlines for FOREX market prediction : A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," Expert Syst. Appl., vol. 42, no. 1, pp. 306–324, 2015.

[17] S. Seifollahi and M. Shajari, "Word sense disambiguation application in sentiment analysis of news headlines : an applied approach to FOREX market prediction Word sense disambiguation application in sentiment analysis of news headlines : an applied approach to FOREX market prediction," no. April 2018, 2019.

[18] S. S. Jadhav and S. D. Thepade, "Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier," Appl. Artif. Intell., vol. 33, no. 12, pp. 1058–1068, 2019.

[19] R. A. Pambudi, Adiwijaya, and M. S. Mubarok, "Multi-label classification of Indonesian news topics using Pseudo Nearest Neighbor Rule," J. Phys. Conf. Ser., vol. 1192, no. 1, 2019.

[20] V. V. Corinna Cortes, "Support-Vector Networks," Kluwer Acad. Publ., 1995.

Identification of Learning Styles and Automatic Assignment of Projects in an Adaptive e-Learning Environment using Project based Learning

Luis Alfaro[1], Erick Apaza[2], Jorge Luna-Urquizo[3], Claudia Rivera[4]
Universidad Nacional de San Agustín de Arequipa, Arequipa - Perú

***Abstract*—The use of the project-based learning approach is one of the emerging trends in education and Adaptive e-Learning platforms. One of the main challenges in this line of research is to identify the learning style of students, whose results are considered in this work, for the assignment of course projects which best suit the characteristics of each student in particular as they incorporate different types of learning strategies and objects in order to contribute to facilitate and simplify the teaching/learning process of the adaptive e-Learning platform, which uses the project-based learning approach. In this work, after carrying out a review of the literature for the theoretical foundation and establishment of the state of Art, a line module of automatic recognition of learning styles is proposed, which uses the information of the interaction of the student with the system and the one that is based on Neural Networks and Fuzzy Logic, whose results are considered by the module of selection and assignment of projects that uses the Case Base Reasoning, later carrying out the tests and the analysis of the results obtained.**

Keywords*—Adaptive e-Learning; project based collaborative learning; case-based reasoning; learning styles; back propagation neural networks

I. Introduction

Project-based Collaborative Learning (PBCL) is an emerging approach within academia, and "one of the most effective strategies for the teaching/learning process" [1]. According to Lee [2], "the main reason is that it is in accordance with the learning concepts of constructionism, cognitive psychology, and integrated curriculum". From within this approach, each project is defined as a complex task or real problem whose resolution involves research activities, planning, design, and development, among others.

According to [3], "it is opposed to the behaviorist pedagogy which is based on a unilateral and passive transmission of knowledge from the teacher to the student". PBCL places the emphasis on topics of everyday concern, and allows students to "explore and solve problems together with their peers" [2], and it contributes to developing in students the motivation to search for and produce knowledge [4]. However, the application of PBCL in the classroom, within the context of traditional teaching/learning, is a task which poses great challenges due to limitations of physical space, time, planning, and available resources, among others [2]. In this context e-learning platforms and environments, under the supervision of teachers or tutors, are constituted as an important alternative for the implementation of and support for the PBCL perspective, thus giving rise to the term 'NetPBL' (Internet Project-Based Learning) although, like all tools, it is also necessary to consider the use risks associated with this technology [5].

Currently, although there are some platforms which, in their most recent versions, give support to the PBLC perspective, they can only do it in a very basic way, without incorporating other elements of higher relevance, such as the use of collaborative tools and project assignments based on user profiles and learning styles (LS), either in individual or group form, in the case of which the LS, profiles, and cultures of the members of each group must be combined according to the proposed objectives, toward the end of enriching the experience and improving social skills and communication.

The present project proposes a model of selection and course-project assignment as support for the PBCL approach, which takes as input the LS and profiles of the user, previously identified from the students' online interactions [6], and utilizes Case-Based Reasoning techniques to recover and adapt the most adequate projects for a determined group of students. The tests were performed with a group of students at the Universidad Nacional de San Agustín de Arequipa (UNSA).

The rest of this paper is organized as follows, Section II presents a revision of related works. Section III describes the proposed model and its method for the identification of LS. Section IV presents the results achieved from the tests performed. Section V presents conclusions and future recommendations for this work.

II. Related Works

A. Project-based Collaborative Learning

Project Based Learning (PBL) is a student centered teaching method based on constructivism and constructionism theories developed by Gergen, Piaget, Inhelder and Vygotsky [7]. A review of literature shows that PBL has increasingly been trialed and adopted across a diversity of educational institutions worldwide [8]. A short definition of PBL establishes: "a model that organizes learning around projects". Likewise, the PBL approach is a teaching method by which students acquire knowledge and skills over a period of time, to investigate and answer a question, a problem or a real challenge which can be complex and involving [9]. According to [10], PBL can be defined in three ways: "a process of inquiry, a learning-to-learn approach, and a method for acquiring new knowledge".

Some of the elements of project design [9], include:

- Essential knowledge, comprehension and success skills: the project focuses on the student's learning

objectives, including basic content and skills such as problem solving, critical thinking, and collaboration.

- A problem or question that constitutes a challenge: the project is framed by a significant problem to be solved or a question to be answered, at an appropriate level of challenge.
- Sustainable research: students engage in a rigorous and lengthy process that consists of asking questions, seeking resources, and applying solutions that come from available information.
- Authenticity: the project presents a context, tasks and tools pattern of quality or real impact, and can also address the concerns, interests, and personal issues of students in their own lives.
- Opinion and possibility of selection of students: Students make decisions about projects, including aspects of how they work, or about what they create.
- Reflection: Students and teachers reflect on the learning and effectiveness of their research activities and projects, the quality of students' work, the emergence of obstacles and how to overcome them.
- Critical review: students give, receive and use feedback to improve their processes and products.
- Product advertising: students publicize the results of their projects, explaining, exhibiting and/or presenting works outside their field of study.

For Maldonado [4], in the PBL model lies the essence of the problem of teaching, showing the student the way to obtain concepts. The projects to be used for learning with a collaborative approach are different from others in aspects such as: They must be student-centered and student-led. They must be clearly defined. They are appropriate. They must connect the student, their personal environment, and the labor competencies. They should provide opportunities for feedback and evaluation by experts. Promote opportunities for reflection and self-assessment. The content should be meaningful to students; which can be directly observable in their environment. Must correlate with real-world problems. Be first-hand research and development. Be sensitive to local culture and culturally.

For [9], learning by means of a project is likely to increase motivation and give the students a sense of satisfaction, it is helpful for developing long-term learning skills, to develop deep, integrated understanding of content and process, it allows students learn to work together to solve problems, and it promotes responsibility and independent learning. In fact, PBL works to integrate and apply:

- Structured new knowledge covered in the course.
- Knowledge learned in other courses.
- Prior life experiential based knowledge.
- New self-taught knowledge.

The unique characteristics and potential of PBL, appropriately incorporated into Adaptive e-Learning environments, will allow the student to interact with learning objects (LO) and resources of the emerging media, which will enrich the learning experiences, making them unique, singular and lasting.

B. Learning Styles (LS)

Learning, which is a series of biological and psychological processes that occur in the cerebral cortex that, thanks to the mediatization of thought, lead the subjects to modify their attitude, ability, knowledge and information, as well as their forms of execution, through the experiences they acquire in the interaction with the external environment, in search of adequate answers. In this perspective, knowledge about different LS is a crucial element for teachers, educational institutions and especially for the design of teaching/learning systems supported by emerging technologies. All individuals have their own style for learning new facts, and there is a great diversity of such LS, which require specific instruments to be identified. In this perspective, there are several models that allow us to identify these characteristics of each individual's LS. For the development of this work it is necessary to review the works of some classic authors of this approach.

On the other hand, according to [11], two dimensions are necessary for learning to occur. The first is described as perception of the medium and the second as processing. Likewise, the author affirms that LS are relatively stable modes according to which individuals acquire and process information in order to act and solve problems. In this way, he says that in order to learn it is necessary to have four basic capacities: concrete experience (CE), reflective observation (RO), abstract conceptualization (AC) and active experimentation (AE).

Students learn in different ways; they have individual preferences and ways of how they perceive and process information. These individual preferences are called LS. In [11], defines them as "personal qualities that influence students' abilities to acquire information, interact with others, and enabling them to participate in learning experiences".

Fewster [12] relied on the experiential learning theory to classify LS. Starting from the idea that learning and transforming experience into knowledge, the author developed an Inventory of LS, which consists of a self questionnaire with a series of questions about how people learn. Thus, the experiential learning cycle derives four LS: divergent, convergent, assimilators and accommodators.

One of the models derived from Kolb's theory is that of Honey-Alonso [13]. According to this, learning and human behavior result from the interaction between the environment, the previous experience lived by the adult learner and the individually constructed knowledge. Thus, there is a strong correlation between Kolb's and Honey-Alonso's styles. In Honey-Alonso model [14], establish four types of apprentices:

- Active Style: they appreciate new experiences, have an open mind and are enthusiastic about new tasks. The adjectives applied to this style are: animators, improvisers, discoverers, open to risk, spontaneous.
- Reflective Style: they like to consider an experience from different perspectives, among other attributes. The adjectives that are applied to this style are: pondered, conscious, receptive and analytical.

- Pragmatic style: they like to put ideas into practice. The adjectives applied to this style are: experimental, practical, direct, effective and realistic.
- Theoretical Style: Facing problems in a logical way. The adjectives applied to this style are: methodical, logical, objective, critical and structured.

In the context of this research project, LS are defined as the form in which each individual selects, obtains, and processes information in the most adequate way for them, according to their individual characteristics and aptitudes. There are currently diverse models or learning-style classification, with 71 models already having been identified [15]. Therefore, it must be considered that each teaching system, according to the model followed, utilizes different methods in order to obtain and identify the LS of each student [16], those being the automatic models utilized in the authors' line of research, taking as a point of reference Honey model [17], because it is centered on how information is perceived and processed, while similiar approaches center on other aspects which are not so relevant from the point of view of e-learning platforms.

If each LS has particular characteristics such as those mentioned above, which can be observed in the way each student learns, it is logical to assume that there are methodological strategies that benefit learning more according to each student's own style. One of the greatest challenges in the design and development of adaptive e-Learning systems is to include the methodological strategies that best suit each type of learner.

C. Case-based Reasoning (CBR)

The CBR is a process that involves the reuse of past experiences, and was used in expert systems and cognitive science. In this approach, the user trying to solve a new problem recognizes similarities with previously solved problems called cases. A case is a commonly specific problem that has been identified, resolved, stored and indexed in a memory with the solution, and optionally the process for obtaining [18].

A case is basically an experience of a solved problem. This can be represented in many different ways. A case base is a collection of such cases. The term-based means that the reasoning is based on cases, that is, cases are the first source for reasoning. The term most characteristic of the approach is reasoning. It means that the approach is intended to draw conclusions using cases, given a problem to be solved. Case based reasoning is an Artificial Intelligence approach adopted by various systems in various problem domains. There are several experiences applied to [19]:

- Classification: select the class to which an entity be appropriate.
- Diagnosis: Select what the analysis of a problem is.
- Prediction: Answered what happens in the future.
- Planning: select an order of activities to a given goal.
- Configuration: Select technical features to included.

Other experiences in literature of the area report diverse applications, such as e-Learning, in which CBR is approached as a solution to an adaptation problem. The objective of the CBR approach is to solve a current objective problem called "objective case", through the use of a similar solved problem called "source problem" and its solution "source solution". The pairs $(sourceProblem, sourceSolution)$ are called "source cases" and the set of source cases is called "Base cases". This resolution process can be modeled using a five-phase static cycle, using a knowledge base of the problem domain [20]:

- Elaborate: reformulate a clear description of the target case from the application submitted.
- Recover: extract cases similar to the description of the problem to be solved.
- Reuse: select the appropriate case for the target case of the recovered cases and reuse it as a solution.
- Review: evaluate the proposed solution (this can be: accepted, corrected or rejected).
- Retain: store final solution as a new case for future use.

For Lopez [21], the process described above at the knowledge level can be instantiated in different CBR systems. In doing so, different CBR system typologies can be considered according four criteria: knowledge source, function, organization, and distributiveness. On the other hand, RS can be considered as one of the areas of CBR application with greater impact that present challenges for research and application [22], such as the determination of the similarity of a product, which can become a complex issue when conventional product descriptions are not available or when they are incomplete when extracting unstructured information and noise from available information, in the opinions of users, treated through the analysis of feelings.

For [23], RS predict users' preferences over a large number of items by pooling similar information from other users or items when observations are sparse, which are particularly useful in personalized prediction. It has become an essential part of e-commerce, with applications in movie rentals, restaurant guides, book recommendations and personalized e-news, for these considerations, it is that another perspective for the treatment of this research work may be the approach of RS.

The recommender systems (RS) take into account not only information about the users but also about the items they consume; comparison with other products, and so on and so forth. Nevertheless, there are many algorithms available to perform a RS. For instance:

- Popularity, where only the most popular items are recommended.
- Collaborative Filtering, which looks for patterns in the user activity to produce user-specific recommendations.
- Content-based Filtering, the recommendation of items with similar information the user has liked or used in the past (description, topic, among others).
- Hybrid Approaches, combines the two algorithms mentioned above.

III. Model System Development

The proposed model was developed as an extension of the Moodle open source LMS, to which two intelligent modules were added, which will be described in the following sections.

A. Automatic Learning-style Recognition Module

The automatic LS identification module, take as reference the Honey-alonso LS model [24] . This model in particular was used as opposed to others, such as the proposals described in Kolb [25], Vincent [26], etc., due to the fact that it centers on "how information is perceived and processed by the user", an especially important factor in LMS platforms. The development of this module was carried out in three stages:

1) Collection of Experimental Data: experimental data were obtained through the application of the test of Honey [24], to 34 students in the School of Marketing - UNSA, being the only group enrolled during that semester in that course, which had been implemented on the platform since before the beginning of the semester. Fig. 1 shows the results of the test, where it is possible to appreciate that, in a high percentage of cases it is not possible to determine a predominant LS.

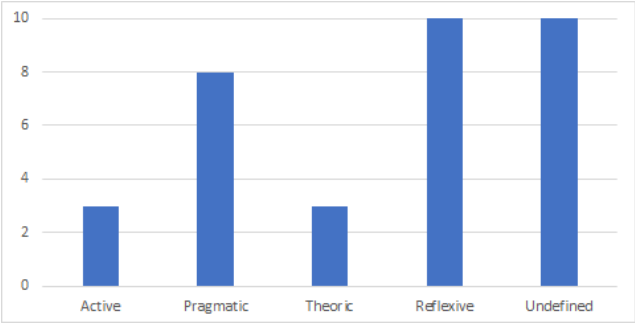


Fig. 1. Test data results obtained

The data collected through the traditional method (test) were utilized for the validation of the authors' research hypothesis, comparing it with the results obtained by the NN. For this purpose, the data were divided into two sets, procuring that both sets maintain similarity with those that were utilized as training and test sets in the development of the NN.

2) Pre-processing of inputs: in the first tests, some deficiencies were found in the results obtained by the NN. The analysis of these incidents led to determine two factors that directly influenced the results, these are:

- The "noise" within the NN input-data sets.
- The input data in which it is difficult to determine a single output.

In order to reduce the impact of the issues in the performance of the proposed model, the authors the authors decided to add a pre-processing phase of the input data to the NN by way of a fuzzy set (FS) which "represents a better categorization of the students preferences for a determined type of resources" [1], taking into consideration the percentage and relevance of the students interactions in each one of the resource categories. This change is justified on the premise that determining the student's specific LS can turn into a fuzzy-nature problem due to the fact that "important situations and characteristics must be considered with a certain level of imprecision" [1]. This means that students' categories are ultimately created with imprecise limits and cannot be well-defined, allowing one to treat the problem with only an approximation of reality, especially when cognitive and subjective aspects are considered.

For the analysis and classification of user interactions, the proposed model defines 20 resource categories, according to the resources available in the Moodle platform, in such a way that each and every student selection (click) could be could be taken into account in a log-file and, therefore, a register of his/her preferences stored. Each one of these categories was related to the Honey-Alonso model through the application of fuzzy logic techniques related to the "theoric and behavioral characteristics associated with each LS", as proposed by [1].

3) LS identification: based on an backpropagation NN, composed of an input layer, a hidden layer, and an output layer (see Fig. 2). The main reason for the utilization of this model was that it allowed for the "training of weights of a NN with an undetermined number of layers" [1], an feature which makes it possible to utilize differentiable transference functions in order to execute approximation functions, association, and classification tasks.

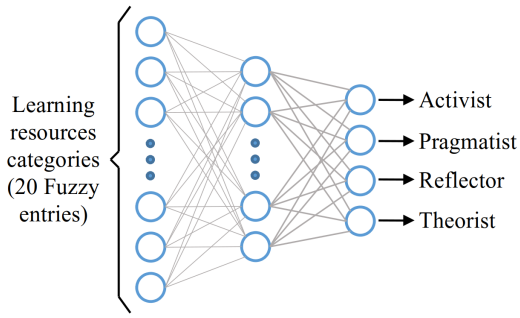


Fig. 2. Neural Network architecture. Source: [6]

The input neurons represents the resource categories defined in [6], which have as input value "the percentage of the interactions within a particular category as a value between 0 and 1" [1], obtained from the log-file. The value 0 indicates that the student did not choose any type of the resource category, and the value 1 indicates that the student only utilizes resources of that category.

During the initial runs with the NN, some problems were discovered in the efficiency of and the results obtained from the NN training session, which according to the analysis conducted correspond to the cases in which the user's responses or interactions do not allow for the clear identification of a single LS, either because of the closeness of all the values or because of multiple outputs with the same value, a situation which complicates the adjustment of weights in the NN and, therefore, the results obtained. This could also have something to do with the mixed and disperse preferences in some students. Two of these cases which are most representative of the phenomenon just described are shown in Table I.

To reduce the impact of these inconveniences, a prepro-

TABLE I. NOISE EXAMPLES

Activist	Reflector	Theorist	Pragmatist
14	16	15	17
15	15	13	15

cessing phase of the inputs based on fuzzy logic techniques was included, with the objective of achieving a more precise categorization of student preferences regarding the categories of learning resources, taking considering the percentage and importance of their interactions with the LMS.

B. CBR-based Project Selection and Assignment

The architecture shown in Fig. 3, corresponds to the model based on the application of CBR techniques. For [1], the projects of the courses are associated with "cases" from the perspective of the PCBL being able to define the problem in the following way: *What is the project that should be assigned to a student or group of students, who have a learning style according to an established learning objective?*.

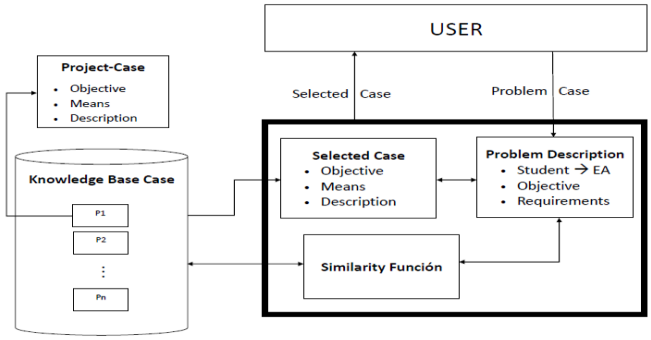


Fig. 3. CBR model architecture.

The starting point for the implementation of the techniques was the open source tool J-Colibri, which presents a layered architecture, and Colibri Studio. The methodology used to solve the problem consists of five steps or tasks:

1) Representation of cases (Fig. 4): Previous experiences are represented by a case, which could be projects that come from previous years, similar courses, etc., which must contain the knowledge that the reasoning model required for its operation. The set of previous cases is called chaos library or case base. A project or case must have at least the following attributes: (i) the problem: this represents the competencies that the student must develop, (ii) the solution: which is described in the project for the resolution of a problem or requirement described above, (iii) the LO and resources available for the execution of the project, (iv) characteristics required or suggested for the team members, which could include LS, (v) other limitations or restrictions.
2) Case recovery: Using some measure of similarity, the current problem should be compared or compared with the problems stored in the base of cases, with the objective of establishing the relevance of the same one, to finally recover the project more similar

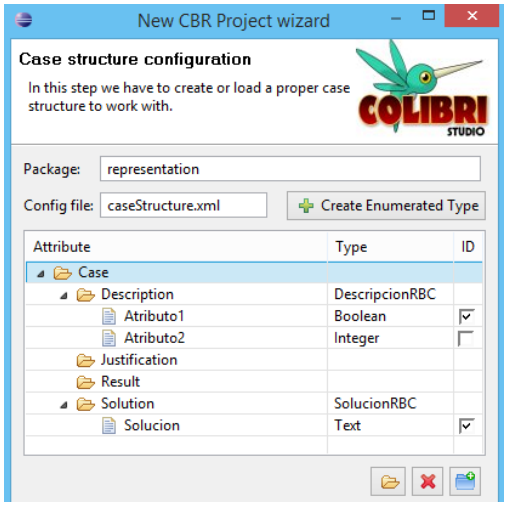


Fig. 4. Definition of Case structure

to the proposed problem. For comparison measures, averages, etc., a relative weight will be established for each of the attributes, according to their importance in the selection of course projects, with the adequate establishment of similarity measures for the selection of course projects constituting an element of great importance, as shown in Fig. 5.

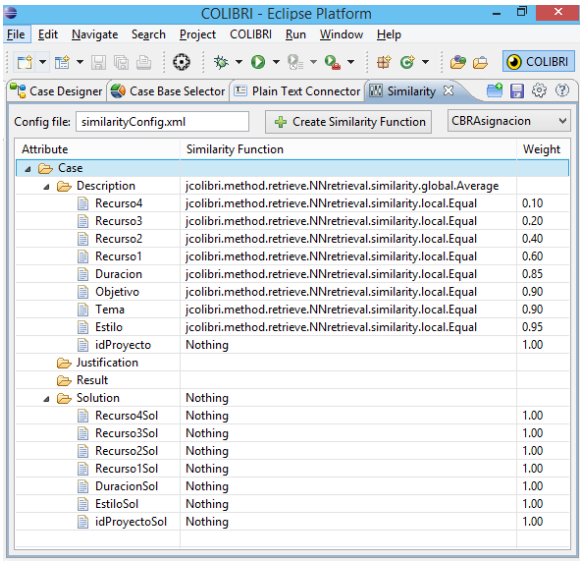


Fig. 5. Case attributes, similarity functions and weights.

3) Reuse of cases: The cases recovered for the resolution of the current problem must be copied and integrated, also formulating a new project in accordance with the objectives and requirements raised. This phase is also known as case adaptation. It can be developed in three ways: substitution, transformation and generative adaptation.
4) Case review: It consists of evaluating the solution originated in the reuse or adaptation phase in the resolution of the new case. It is generally carried out by the experts of the domain. If the solution requires

some adjustment, the solution is repaired iteratively until it is validated.

5) Retention of cases: When a new solution is validated by experts, the new case is stored in the case base with clearly defined criteria for future use, thus increasing the case base, which will allow a greater repertoire of new solutions. It is necessary to emphasize that a greater number of stored cases has a direct impact on the new solutions, which will also be increasingly complete and precise.

In the context of the proposed platform, a LO is defined as a digital material with different granularity, which can be used for educational purposes from an intentionality that is defined implicitly or explicitly by educational objectives, and contains metadata that allow its description and retrieval, facilitating its reuse and adaptation to different environments, among which the following can be used: the theme, pedagogical style, format, difficulty level, age range or copyright restrictions, and may also include keywords, descriptors, among others.

In this way, a typical query, as shown in Fig. 6, is based on the student's learning style, taken as input from the output of the automatic identification of learning styles module, the topic to be resolved, the learning objective, the duration of the project, the available resources or LO that are accessed in the online repository, among other attributes.

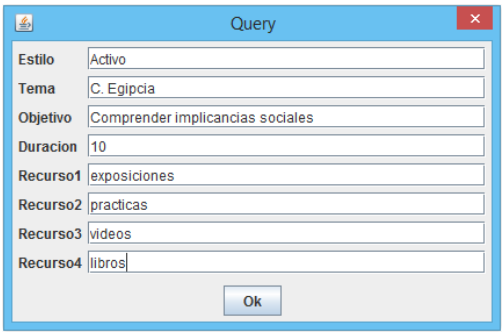


Fig. 6. Example of RBC query.

In the proposed model, the comparison of the attributes is performed using a linear similarity function, according to the procedure detailed in [1].

IV. Results

A. Results of the Learning-styles-identification Module

For the tests of this module, students perform a series of activities throughout a semester. According to [1]: "the identification must be made throughout a period of utilization time, given that the data analyzed in just one session might be seen as influenced by various factors". Fig. 7 presents the identification of the LS of four students throughout each week, where it can be appreciated that, for the 'case-1' student, the identification realized in weeks 3 and 20 might indicate that the student fits into the second category, however, it is clear that this student instead fits into the first category.

This result is understandable due to the proximity among some LS and the mixed preferences of some students. For

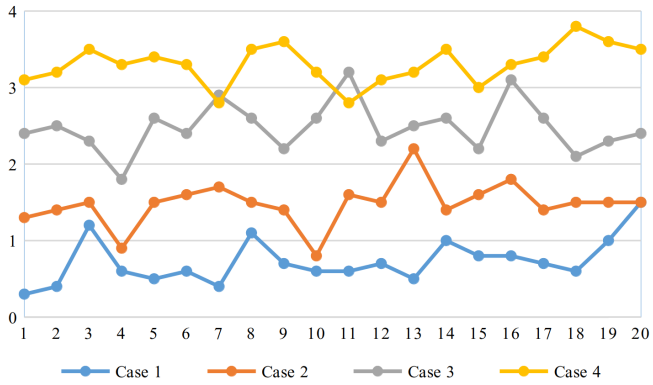


Fig. 7. Identification of LS of four students throughout time. Source: [6]

this reason, the best option would be to identify the LS during some introductory course or previous activities, before initiating the adaptation tasks and content personalization of the LMS, and refine iteratively in the following activities or courses throughout the student's training process.

For the tests of this model, the users' interactions, which were obtained through the LMS, and the test data, collected through the traditional method (test), were utilized, having been divided into two data sets:

- Training set: utilized for the training of the NN, to determine the parameters (weights) of the hidden layer's neurons. For this, user responses to the questionnaire are utilized, correlated with the resource categories proposed as input data. The network is trained in order to obtain the same result as the traditional method, which is to say the tabulation of responses to the questionnaire, which were known beforehand.
- Test set: utilized in order to compare the results obtained by the NN based on user interactions in the platform, as input, with the LS previously identified for the same group of users.

The results obtained by the NN had a 76.5% of coincidence with the results collected under the traditional method, for which the proposed model is considered to have reached a 76.5% of efficiency. Table II shows a comparison of the distinct approaches proposed for the automatic identification of LS, taking into consideration the models or learning-style schemes utilized in each approach. It must be pointed out that in the case of the approaches where the efficiency for each LS in isolation is calculated, the average of the same ones as a measurement of global efficiency has been considered with the objective of facilitating the comparison between the distinct approaches.

B. Results of the Selection-module Tests and Project Assignment

For this tests, 50 case-retrieval queries were conducted according to the procedure shown in Fig. 6, taking as principal search criteria the learning objective, the problem to solve, project duration, and the LS. These queries were conducted utilizing a case base defined in a general way by LS and LO specialists, obtaining a state of affairs in which for 80% of the queries the module could find at least 3 similar cases and

TABLE II. COMPARISON OF EVALUATED MODELS. SOURCE: [6]

Evaluated techniques	Learning Styles model	Efficiency
Bayesian networks	Felder & Silverman	66%
NBTree y CRB	Felder & Silverman	67.5%
Genetic algorithms and K-NN	Ad-hoc	96%
Monitoring of interactions	Felder & Silverman	79.6%
LOs and time estimation	Felder & Silverman	69.6%
NN and navigation maps	Vincent & Ross	90%
Stochastic models	Felder & Silverman	70%
NeuroFuzzy model (Proposal)	Honey & Mumford	77.1%

at least one of them with a similarity measurement superior to 80% (see Fig. 8), where the alternatives are ordered from greatest to least according to the degree of similitude.

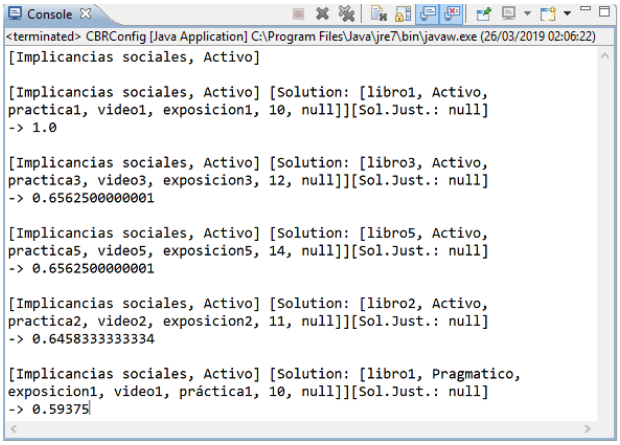


Fig. 8. Results of example query for project assignment. Source: [6]

Table III shows the summary of the tests conducted and the best similitude value obtained for each one of them, utilizing the comparison by linear similitude, in which it can be appreciated that the results are highly acceptable.

TABLE III. RESULTS OF RBC TEST QUERYS

Test	Result	Test	Result	Test	Result
1	85%	18	84%	35	93%
2	81%	19	74%	36	81%
3	100%	20	88%	37	61%
4	82%	21	94%	38	88%
5	91%	22	78%	39	84%
6	69%	23	81%	40	100%
7	100%	24	80%	41	80%
8	80%	25	68%	42	95%
9	87%	26	100%	43	85%
10	92%	27	83%	44	77%
11	84%	28	90%	45	82%
12	72%	29	86%	46	83%
13	82%	30	80%	47	91%
14	86%	31	57%	48	87%
15	91%	32	87%	49	80%
16	100%	33	100%	50	66%
17	66%	34	95%		%

The objective of this work is to integrate the LS identification module, based on Fuzy NN and user behavior analysis, with the RBC-based project selection module, so that, according to the student's profile, it is possible to assign projects that have the structure, learning objectives, and what best contributes to user interaction in the LMS, and thus enable more meaningful learning using the PBL approach. It would be convenient to subsequently add experimental data corresponding to several semesters or courses, in order to analyze the real impact of the proposal in the learning process.

V. CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

A subsystem was proposed online, of automatic recognition of LS that uses the information of the interaction of the student with the system, and is based on NN and Fuzzy Logic, whose tests and comparison, with works carried out by other groups of investigation is quite satisfactory. Likewise, a subsystem was proposed and developed based on CBR, which allowed an assignment of projects, with quite high relevance of percentages in relation to the LS and profiles of the students.

Likewise, the design and incremental development of the Hybrid adaptive e-Learning system model that uses the PBL approach has a high degree of customization for the assignment of projects and LOs that correspond to the personal characteristics of the learner, contributing to facilitate and simplify the teaching/learning process of the adaptive e-Learning platform.

Finally, the use of a neurofuzzy NN and the CBR techniques, allowed an original approach to the treatment of this complex problem, obtaining satisfactory results.

B. Recommendations

It is necessary to broaden the base of cases with information from future semester courses, for the realization of new tests, which will allow a better analysis of the results obtained.

It is necessary to explore the possibilities of incorporating the students thinking styles to the Adaptive e-Learning model.

Finally, it is necessary to explore the possibilities of incorporating the resources of Immersive Virtual Reality and Augmented Reality to the adaptive e-Learning model, so that the interaction with the LOs does not have linguistic, symbolic and/or computational mediation, which necessarily incorporates reflection, a fact that does not contribute to students building their knowledge through direct experiences.

ACKNOWLEDGMENT

The authors would like to thank the Universidad Nacional de San Agustín de Arequipa, for supporting this research.

REFERENCES

[1] L. Alfaro, C. Rivera, and J. Luna-Urquizo, "Using project-based learning in a hybrid elearning system model," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019.

[2] C.-I. Lee and F.-Y. Tsai, "Internet project-based learning environment: the effects of thinking styles on learning transfer," *Journal of Computer Assisted Learning*, vol. 20, pp. 31–39, feb 2004.

[3] F. Abdallah, C. Toffolon, and B. Warin, "Models transformation to implement a project-based collaborative learning (PBCL) scenario: Moodle case study," in *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, IEEE, 2008.

[4] M. Maldonado, "Aprendizaje basado en proyectos colaborativos. una experiencia en educacion superior," *Laurus - Revista de Educacion*, vol. 14, no. 28, pp. 158–180, 2008.

[5] D. M. Verstegen, N. de Jong, J. van Berlo, A. Camp, K. D. Könings, J. J. van Merriënboer, and J. Donkers, "How e-learning can support pbl groups: A literature review," in *Educational technologies in medical and health sciences education*, pp. 9–33, Springer, 2016.

[6] L. Alfaro, C. Rivera, J. Luna-Urquizo, E. Castaneda, and F. Fialho, "Utilization of a neuro fuzzy model for the online detection of learning styles in adaptive e-learning systems," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, 2018.

[7] S. Aldabbus, "Project-based learning: Implementation & challenges," *International Journal of Education, Learning and Development*, vol. 6, no. 3, pp. 71–79, 2018.

[8] B. Condliffe, "Project-based learning: A literature review. working paper.," *MDRC*, 2017.

[9] E. d. S. Zancul, T. T. Sousa-Zomer, and P. A. Cauchick-Miguel, "Project-based learning approach: improvements of an undergraduate course in new product development," *Production*, vol. 27, no. SPE, 2017.

[10] R. E. Gewurtz, L. Coman, S. Dhillon, B. Jung, and P. Solomon, "Problem-based learning and theories of teaching and learning in health professional education," *Journal of Perspectives in Applied Academic Practice*, vol. 4, no. 1, 2016.

[11] L. N. R. Agudelo, V. S. Urbina, and F. J. M. Gutiérrez, "Estilos de aprendizaje basados en el modelo de kolb en la educación virtual," *Apertura*, vol. 2, no. 1, 2010.

[12] L. Fewster-Thuente and T. J. Batteson, "Kolb's experiential learning theory as a theoretical underpinning for interprofessional education," *Journal of allied health*, vol. 47, no. 1, pp. 3–8, 2018.

[13] R. Rodríguez Cepeda, "The learning models of kolb, honey and mumford: implications for science education," *Sophia*, vol. 14, no. 1, pp. 51–64, 2018.

[14] A. F. Hoffman and M. F. Liporace, "Cuestionario honey-alonso de estilos de aprendizaje: Análisis de sus propiedades psicométricas en estudiantes universitarios," *Summa psicológica UST*, vol. 10, no. 1, pp. 103–117, 2013.

[15] F. Coffield, D. Moseley, E. Hall, and K. Ecclestone, *Should we be using learning styles? What research has to say to practice*. London: Learning and skills research centre, 2004.

[16] A. Klašnja-Milićević, B. Vesin, M. Ivanović, and Z. Budimac, "E-learning personalization based on hybrid recommendation strategy and learning style identification," *Computers & Education*, vol. 56, pp. 885–899, apr 2011.

[17] P. Honey and A. Mumford, *Learning styles questionnaire*. Organization Design and Development, Incorporated, 1989.

[18] F.-Z. Berriche, B. Zeddini, H. Kadima, and A. Rivière, "Cbr-mining approach to improve learning system engineering in a collaborative e-learning platform.," in *COMPLEXIS*, pp. 94–101, 2018.

[19] S. Shekapure and D. D. Patil, "Enhanced e-learning experience using case based reasoning methodology," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 10, no. 4, pp. 236–241, 2019.

[20] A. Soundoss, G. Mohamed, and F. Abdelhadi, "Case-based reasoning for e-learning systems: State of the art," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pp. 353–356, IEEE, 2018.

[21] B. López, "Case-based reasoning: a concise introduction," *Synthesis lectures on artificial intelligence and machine learning*, vol. 7, no. 1, pp. 1–103, 2013.

[22] D. B. Leake, B. Smyth, and R. Weber, "Guest editors' introduction: special issue on case-based reasoning," *Journal of Intelligent Information Systems*, vol. 46, no. 2, pp. 235–236, 2016.

[23] B. Dai, J. Wang, X. Shen, and A. Qu, "Smooth neighborhood recommender systems," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 589–612, 2019.

[24] C. Alonso, D. Gallego, and P. Honey, *Los estilos de aprendizaje. Procedimientos de diagnóstico y mejora*. Bilbao: Mensajero, 1994.

[25] D. A. Kolb, *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[26] A. Vincent and D. Ross, "Personalize training: determine learning styles, personality types and multiple intelligences online," *The Learning Organization*, vol. 8, pp. 36–43, mar 2001.